Control Theory for Physicists

JOHN BECHHOEFER

Contents

PART I

CORE MATERIAL

2 Dynamical Systems

Problems

2.1 Balancing a pendulum by moving a cart.

a. Using a Lagrangian, derive Eq. (2.11).
b. For the case of a uniform stick of mass m and length ℓ, show that

$$\ddot{x} + \tfrac{1}{2}\left(\ddot{\theta}\cos\theta - \dot{\theta}^2\sin\theta\right) = u\,, \qquad \ddot{\theta} + \sin\theta + \tfrac{3}{2}\ddot{x}\cos\theta = d\,,$$

where $u(t)$ acts directly on the stick and $d(t)$ is a torque on the pendulum.
c. Write these equations as an equivalent first-order equation of the form $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u})$.

Solution.

a. The Lagrangian $L = T - V$ is

$$\begin{aligned} L &= \tfrac{1}{2}M\dot{x}^2 + \tfrac{1}{2}m\left[(\dot{x} + \dot{x}_\ell)^2 + \dot{y}_\ell^2\right] - (1 - mg\ell\cos\theta) \\ &= \tfrac{1}{2}(M+m)\dot{x}^2 + \tfrac{1}{2}m\ell^2\dot{\theta}^2 + m\ell\cos\theta\,\dot{x}\dot{\theta} + mg\ell\cos\theta - 1\,, \end{aligned}$$

with $x_\ell = \ell\sin\theta$ and $y_\ell = \ell\cos\theta$. If we neglect friction, the Lagrangian gives nonlinear equations of motion for x and θ. We begin with the x equation:

$$\partial_{\dot{x}}L = (M+m)\dot{x} + m\ell\cos\theta\,\dot{\theta}\,, \qquad \partial_x L = 0\,.$$

The x equation is then $d_t\partial_{\dot{x}}L - \partial_x L = u$, where $u(t)$ is the external force:

$$(M+m)\ddot{x} - m\ell\sin\theta\,\dot{\theta}^2 + m\ell\cos\theta\,\ddot{\theta} = u\,.$$

The θ equation is $d_t\partial_{\dot{\theta}}L - \partial_\theta L = d$, where $d(t)$ is an external torque disturbance:

$$\partial_{\dot{\theta}}L = m\ell^2\dot{\theta} + m\ell\cos\theta\,\dot{x}\,, \qquad \partial_\theta L = -m\ell\sin\theta\,\dot{x}\dot{\theta} - mg\ell\sin\theta\,,$$

which leads to

$$m\ell^2\ddot{\theta} - \cancel{m\ell\sin\theta\,\dot{x}\dot{\theta}} + m\ell\cos\theta\,\ddot{x} + \cancel{m\ell\sin\theta\,\dot{x}\dot{\theta}} + mg\ell\sin\theta = d\,.$$

Note that we put in the external force "by hand." More formally, they can be included in the Lagrangian: $L \to L - u(t)x - d(t)\theta$. The forcing terms are then generated automatically by the Euler-Lagrange equations.

Collecting the two equations, we have

$$(M+m)\,\ddot{x} + m\ell\,(\ddot{\theta}\cos\theta - \dot{\theta}^2\sin\theta) = u\,,$$
$$m\ell\,(\ddot{x}\cos\theta + \ell\ddot{\theta} + g\sin\theta) = d\,.$$

Next, we scale the equations to make them dimensionless. We will use the same variables to represent the scaled quantities. We define $\omega^2 = g/\ell$ and let $t \to \omega t$, $x \to x/\ell$, $u \to u/[(M+m)g]$, and $d \to d/(mg\ell)$. We find

$$\ddot{x} + \left(\frac{m}{M+m}\right)\left(\ddot{\theta}\cos\theta - \dot{\theta}^2\sin\theta\right) = u\,, \qquad \ddot{\theta} + \sin\theta + \ddot{x}\cos\theta = d\,,$$

b. Now we consider the closely related case where we impose a uniform force on a stick of mass m and length ℓ. The easiest is to calculate the linear kinetic energy relative to the center of mass, as $\frac{1}{2}\ell$, and similarly for the potential energy. The kinetic energy about the center of mass is then $\frac{1}{2}I\dot{\theta}^2$, where $I = \frac{1}{12}m\ell^2$ for a stick *about its center of mass*. The Lagrangian becomes

$$\begin{aligned} L &= \tfrac{1}{2}m\left[\left(\dot{x} + \tfrac{1}{2}\ell\dot{\theta}\cos\theta\right)^2 + \left(\tfrac{1}{2}\ell\dot{\theta}\sin\theta\right)^2\right] + \tfrac{1}{24}m\ell^2\dot{\theta}^2 - \left(1 - \tfrac{1}{2}mg\ell\cos\theta\right) \\ &= \tfrac{1}{2}m\dot{x}^2 + \tfrac{1}{6}m\ell^2\dot{\theta}^2 + \tfrac{1}{2}m\ell\cos\theta\,\dot{x}\dot{\theta} + \tfrac{1}{2}mg\ell\cos\theta - 1 \end{aligned}$$

Note that there is no "cart" in this problem: the hand exerts a force directly on the stick. Then

$$\partial_{\dot{x}}L = m\dot{x} + \tfrac{1}{2}m\ell\cos\theta\,\dot{\theta}\,, \qquad \partial_x L = 0\,,$$

and the x-equation is

$$m\ddot{x} + \tfrac{1}{2}m\ell\left(\ddot{\theta}\cos\theta - \dot{\theta}^2\sin\theta\right) = u\,.$$

The θ-equation is obtained by first calculating

$$\partial_{\dot{\theta}}L = \tfrac{1}{3}m\ell^2\dot{\theta} + \tfrac{1}{2}m\ell\cos\theta\,\dot{x}\,, \qquad \partial_\theta L = -\tfrac{1}{2}\sin\theta\,\dot{x}\dot{\theta} - \tfrac{1}{2}mg\ell\sin\theta\,,$$

and, thus,

$$\tfrac{1}{3}m\ell^2\ddot{\theta} - \cancel{\tfrac{1}{2}m\ell\sin\theta\,\dot{x}\dot{\theta}} + \tfrac{1}{2}m\ell\cos\theta\ddot{x} + \cancel{\tfrac{1}{2}\sin\theta\,\dot{x}\dot{\theta}} + \tfrac{1}{2}mg\ell\sin\theta = d\,.$$

Together, the two dimensional equations are

$$m\ddot{x} + \tfrac{1}{2}m\ell\left(\ddot{\theta}\cos\theta - \dot{\theta}^2\sin\theta\right) = u\,, \qquad \tfrac{1}{3}m\ell^2\ddot{\theta} + \tfrac{1}{2}mg\ell\sin\theta + \tfrac{1}{2}m\ell\cos\theta\ddot{x} = d\,.$$

In scaling the equations, one difference is that we define $\omega^2 = \frac{3}{2}g/\ell$, so that ω is the frequency of small oscillations of the stick. Then it is straightforward to verify that

$$\ddot{x} + \tfrac{1}{2}\left(\ddot{\theta}\cos\theta - \dot{\theta}^2\sin\theta\right) = u\,, \qquad \ddot{\theta} + \sin\theta + \tfrac{3}{2}\ddot{x}\cos\theta = d\,,$$

where we scale u by $\frac{3}{2}mg$ and d by $mg(\frac{1}{2}\ell)$. The last quantity is again the torque to hold the stick horizontally. The factors of $\frac{3}{2}$ can be interpreted as defining an effective mass $\frac{3}{2}m$.

c. To write these equations as a system of four coupled, nonlinear equations in the form $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u})$, we must first decouple the $\ddot{x}$ and $\ddot{\theta}$ terms in the second-order equations. We write

$$\begin{pmatrix} 1 & \frac{1}{2}\cos\theta \\ \frac{3}{2}\cos\theta & 1 \end{pmatrix} \begin{pmatrix} \ddot{x} \\ \ddot{\theta} \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\dot{\theta}^2 \sin\theta + u \\ -\sin\theta + d \end{pmatrix}.$$

Inverting the matrix gives

$$\begin{aligned} \begin{pmatrix} \ddot{x} \\ \ddot{\theta} \end{pmatrix} &= \frac{1}{1 - \frac{3}{4}\cos^2\theta} \begin{pmatrix} 1 & -\frac{1}{2}\cos\theta \\ -\frac{3}{2}\cos\theta & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{2}\dot{\theta}^2 \sin\theta + u \\ -\sin\theta + d \end{pmatrix} \\ &= \frac{1}{1 - \frac{3}{4}\cos^2\theta} \begin{pmatrix} \frac{1}{2}\dot{\theta}^2 \sin\theta + u + \frac{1}{2}\sin\theta\cos\theta - \frac{1}{2}d\cos\theta \\ -\sin\theta + d - \frac{3}{4}\sin\theta\cos\theta - \frac{3}{2}u\cos\theta \end{pmatrix}. \end{aligned}$$

By defining

$$\boldsymbol{x} = \begin{pmatrix} x \\ \dot{x} \\ \theta \\ \dot{\theta} \end{pmatrix} \equiv \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \quad \text{and} \quad \boldsymbol{u} = \begin{pmatrix} u \\ d \end{pmatrix} \equiv \begin{pmatrix} u_1 \\ u_2 \end{pmatrix},$$

we can write first-order equations of motion:

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} x_2 \\ \frac{1}{1 - \frac{3}{4}\cos^2 x_3} \left(\frac{1}{2} x_4^2 \sin x_3 + u_1 + \frac{1}{2}\sin x_3 \cos x_3 - \frac{1}{2} u_2 \cos x_3 \right) \\ x_4 \\ \frac{1}{1 - \frac{3}{4}\cos^2 x_3} \left(-\sin x_3 + u_2 - \frac{3}{4}\sin x_3 \cos x_3 - \frac{3}{2} u_1 \cos x_3 \right) \end{pmatrix}.$$

Comments:

- The nonlinear state-space form $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u})$ in the last part is neither intuitive nor convenient for derivations done by hand. For humans, the coupled second-order equations are easier. The state-space form, however, eases automated symbolic computations and numerical solution.
- We choose a coordinate system where θ increases *counterclockwise* from 0, the down equilibrium orientation of the pendulum. Often, people choose coordinates where $\theta = 0$ corresponds to the unstable top position of the pendulum and increases for *clockwise* rotation. The transformation between them is $\theta = \pi - \phi$, which implies $\cos\theta = -\cos\phi$, $\sin\theta = +\sin\phi$, $\dot{\theta} = -\dot{\phi}$, and $\ddot{\theta} = -\ddot{\phi}$.

2.2 First-order systems, frequency domain. Derive analytically the curves in the graphs of the first-order dynamical system $G(s) = \frac{1}{1+s}$ (Bode plots, pole-zero, and Nyquist plots) shown in Figure 2.4. For the Nyquist plot, derive the geometric shape of the curve, not just the parametric form.

Solution.

a. *Bode plots*. The frequency-domain transfer function is $G(\mathrm{i}\omega) = \frac{1}{1+\mathrm{i}\omega} = \frac{1-\mathrm{i}\omega}{1+\omega^2}$. The magnitude and phase relations are then

$$|G| = \frac{1}{\sqrt{(1+\mathrm{i}\omega)(1-\mathrm{i}\omega)}} = \frac{1}{\sqrt{1+\omega^2}}$$
$$\varphi = \tan^{-1}\left(\frac{-\omega}{1}\right) = -\tan^{-1}\omega\,.$$

b. *Pole-zero plots*. From $G(s) = \frac{1}{1+s}$, we see that there is a pole at $s = -1$.

c. *Nyquist plots*. We plot Im $G(\mathrm{i}\omega)$ vs. Re $G(\mathrm{i}\omega)$:

$$G = \frac{1-\mathrm{i}\omega}{1+\omega^2} \equiv x + iy\,,$$

where $x = \frac{1}{1+\omega^2}$ and $y = \frac{-\omega}{1+\omega^2}$. Thus,

$$x^2 + y^2 = \frac{1}{1+\omega^2} = x\,,$$

and

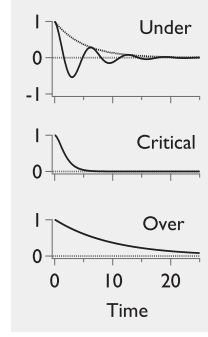
$$x^2 - x + \tfrac{1}{4} - \tfrac{1}{4} + y^2 = 0$$
$$(x - \tfrac{1}{2})^2 + y^2 = \tfrac{1}{4}\,,$$

which is a circle with center $(\frac{1}{2}, 0)$ and radius $\frac{1}{2}$. Because Nyquist plots are conventionally the locus of *positive* frequencies, we see, from the explicit form of $G(\mathrm{i}\omega)$, that the *negative* half-circle branch is covered.

2.3 Second-order systems.

a. Consider a generic second-order system $G(s) = (1 + 2\zeta s + s^2)^{-1}$. As in Problem 2.2, derive analytic expressions for the Bode, pole-zero, and Nyquist plots (Figure 2.5), for $\zeta < 1$, $\zeta = 1$, and $\zeta > 1$, (underdamped, critically damped, and overdamped), respectively, for some parts of this problem. Find analytic approximations for $\zeta \ll 1$ and $\zeta \gg 1$. Give the Nyquist plots in parametric form.

b. A common control goal is to have critically damped closed-loop dynamics ($\zeta = 1$). Show that the damping time of the decay of an oscillator is shortest for critical damping. Confirm the time-decay plots at right? ($\zeta = 0.2,\ 1,\ 5$) and the plot below.

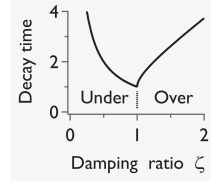


Solution.

a. *Graphical aids*

i. *Bode plots*. We write

$$G(\mathrm{i}\omega) = \frac{1}{1-\omega^2+2\mathrm{i}\zeta\omega} = \frac{1-\omega^2-2\mathrm{i}\zeta\omega}{(1-\omega^2)^2+4\zeta^2\omega^2}\,,$$



so that

$$|G(\mathrm{i}\omega)| = \frac{1}{\sqrt{(1-\omega^2)^2 + 4\zeta^2\omega^2}}\,, \qquad \tan\varphi = \frac{-2\zeta\omega}{1-\omega^2}\,.$$

ii. *Pole-zero plot*. The poles are the roots of

$$s^2 + 2\zeta\, s + 1 = 0\,,$$

which have three cases:

- $\zeta < 1$: $s = -\zeta \pm \mathrm{i}\sqrt{1-\zeta^2}$ (Complex-conjugate pair). For $\zeta \ll 1$, we have $s \approx -\zeta \pm \mathrm{i}$.
- $\zeta = 1$: $s = -1$ (two-fold degeneracy).
- $\zeta > 1$: $s = -\zeta \pm \sqrt{\zeta^2 - 1}$ (two real roots). For $\zeta \gg 1$, we have $s_+ \approx -\frac{1}{2\zeta}$ and $s_- \approx -2\zeta$. Notice how the second root "comes in from infinity" as ζ increases.

iii. *Nyquist plot*. This is complicated to express geometrically. In parametric from (for $0 < \omega < \infty$), we have

$$x = \frac{1-\omega^2}{(1-\omega^2)^2 + 4\zeta^2\omega^2}\,, \qquad y = \frac{-2\zeta\omega}{(1-\omega^2)^2 + 4\zeta^2\omega^2}\,.$$

b. The transfer function of the closed-loop system might typically be given by

$$T(s) = \tfrac{1}{1+2\zeta s+s^2}\,,$$

whose decay time is given by

$$\frac{1}{\mathrm{Re}\,\left(\zeta - \sqrt{\zeta^2-1}\right)}\,.$$

For $\zeta \le 1$, this is just $1/\zeta$. For $\zeta > 1$, the expression is real. Plotting this gives the graph in the main text.

2.4 **Transfer function for thermal conduction**. We showed that a semi-infinite, one-dimensional thermal conductor gives rise to a transfer function between heater and thermometer of $G(s) = \mathrm{e}^{-\sqrt{s}}/\sqrt{s}$. See Eq. (2.44). The probe is at $x = \ell$ (scaled to 1).

a. Derive explicit expressions for the magnitude and phase of the frequency response. Use the results to reproduce the Bode and Nyquist plots.
b. How are Bode and Nyquist plots altered for a different sensor point ($x = 2$)?
c. We can approximate non-rational transfer functions by rational polynomials (Padé approximants; Cf. Section 3.6.4). The second-order Padé approximation to $G(s)$ about $s = 1$ is $G_{2,2}(s) = (\mathrm{e}^{-1})\,\frac{44-76s+8s^2}{5+26s-55s^2}$. Compare its Bode plot with that of $G(s)$. Use a computer-algebra program to compute and compare $G_{3,3}(s)$.
d. Show that if we use two temperature probes, at distances x_1 and x_2, the transfer function between the two probes is $G(s) = \mathrm{e}^{-\sqrt{s}\ell}$, where $\ell = x_2 - x_1$.

Solution.

a. *Frequency response*. For $G(s) = \mathrm{e}^{-\sqrt{s}} / \sqrt{s}$, the frequency response is

$$G(\mathrm{i}\omega) = \frac{\mathrm{e}^{-\sqrt{\mathrm{i}\omega}}}{\sqrt{\mathrm{i}\omega}}$$

We recall that $\sqrt{\mathrm{i}} = \mathrm{e}^{\mathrm{i}\,\pi/4} = \frac{1}{\sqrt{2}}(1 + \mathrm{i})$. Substituting and collecting terms gives

$$G(\mathrm{i}\omega) = \frac{\mathrm{e}^{-\frac{(1+\mathrm{i})\sqrt{\omega}}{\sqrt{2}}}}{\sqrt{\omega}}\,\mathrm{e}^{-\mathrm{i}\,\pi/4} = \frac{\mathrm{e}^{-\sqrt{\omega/2}}}{\sqrt{\omega}}\,\mathrm{e}^{-\mathrm{i}(\pi/4+\sqrt{\omega/2})}\ .$$

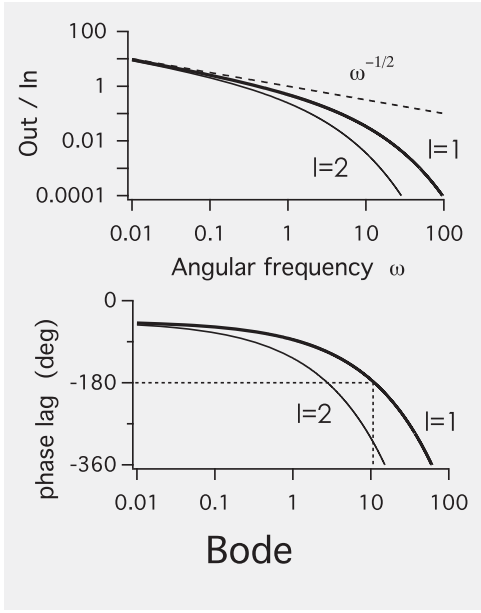
Thus,

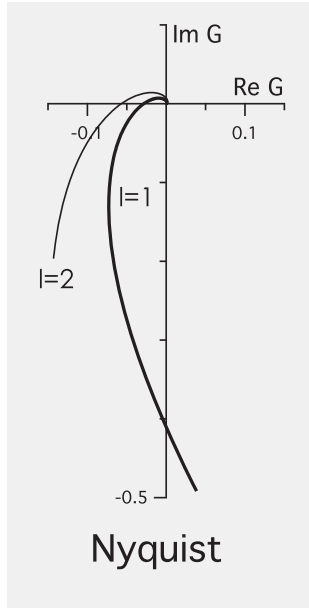
$$|G(\mathrm{i}\omega)| = \frac{\mathrm{e}^{-\sqrt{\omega/2}}}{\sqrt{\omega}}\,, \qquad \varphi = -\left(\pi/4 + \sqrt{\omega/2}\right).$$

b. *Change the sensor point*. Repeating the derivation for a sensor at position $x = \ell$, we have

$$G(s) = \frac{\mathrm{e}^{-\ell\sqrt{s}}}{\sqrt{s}}\,.$$

The Bode plots are shown below.



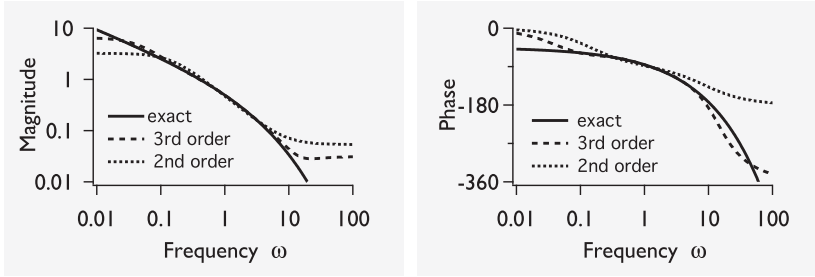


The $\ell = 1$ curves are shown as a heavy trace and correspond to $\ell = 1$. The $\ell = 2$ curves basically have a bit more delay, are closer to instability, and so forth.

c. The approximation $G_{3,3}(s)$ is

$$G_{3,3}(s) = \left(\mathrm{e}^{-1}\right)\frac{551s^3 - 10521s^2 + 115029s + 32701}{6508s^3 + 84468s^2 + 44940s + 1844}\,.$$

The Bode plots are given below:



We see that the approximations are better for the magnitude than the phase. One of the issues with the phase is that the exact transfer function has unbounded phase lag, while the nth-order Padé approximation asymptotes to a phase lag of $180(n-1)$.

For the magnitude, the second-order approximation is reasonable in the range $\omega \in (0.1, 3)$ and the third-order approximation is reasonable in the range $\omega \in (0.03, 10)$.

In practice, a range of 100 in frequency is reasonable, so that the (2,2) or (3,3) Padé approximation is useful.

d. The problem is the same as that solved in the text, except that

$$G(s) = \frac{T(x_2, s)}{T(x_1, s)} = \frac{B\,\mathrm{e}^{-\sqrt{s}x_2}}{B\,\mathrm{e}^{-\sqrt{s}x_1}} = \mathrm{e}^{-\sqrt{s}\ell},$$

where $\ell = x_2 - x_1$. Note that one might want to reconsider the scaling of distance. If $x_1 \approx 0$ (probe near heater), we can use the same scaling as before. The two-temperature configuration is nice because we can measure the transfer function without knowing many details about the heater. For example, although temperature probes can be pretty small, heaters tend to be big, and their size might need to be modeled to understand the heat flow near the heater.

2.5 Ångström's method for measuring the thermal diffusion coefficient. Ångström (1861) derived a "remarkably simple" result for the thermal diffusion constant that led to far more accurate measurements of D. Distributed heat losses to the environment modify the diffusion equation to be $\partial_t T(x,t) = D\partial_{xx}T - \mu T$. Unfortunately, μ is difficult to measure and reflects all the details of the geometry of the experiment. Ångström found an expression for D that was *independent* of μ.

a. Show that the transfer function between two points on the rod that are separated by a distance $\Delta\ell$ is given by $G(s) = \exp\left[-\sqrt{\mu + s}\,(\Delta\ell/D^{1/2})\right]$. The temperature is the "input" at a point ℓ_1 and the "output" at point $\ell_2 = \ell_1 + \Delta\ell$.

b. Derive Ångström's result: $[\ln|G(\mathrm{i}\omega)|]\,\theta(\omega) = \frac{(\Delta\ell)^2\omega}{2D}$, which is indeed independent of μ. Hints: $G = |G|\,\mathrm{e}^{\mathrm{i}\theta} \implies \ln G = \ln|G| + \mathrm{i}\theta$. Then look at $(\ln G)^2$.

Thus, you simply oscillate one end of your material at ω and measure the log of the ratio of temperature-oscillation amplitudes at two different points and their phase difference. Ångström varied the temperature by alternating cold water and steam using a valve. He used the method to measure the thermal diffusivity of copper (expressed as a conductivity). His result of 382 W/m/K is 5% below the modern value of 401 W/m/K (Haynes, 2014). The best previous measurement, 80 W/m/K, was far too low because it did not account correctly for heat losses. Ångström's method became the standard one for measuring thermal diffusivity and was the first use of thermal "diffusion waves" to probe material properties (Mandelis, 2000).

Solution.

a. *Transfer function between two points*. From

$$\partial_t T(x,t) = D\partial_{xx} T - \mu T\,,$$

we Laplace transform the time variable to find

$$\begin{aligned} sT(x,s) &= D\partial_{xx} T(x,s) - \mu T(x,s) \\ (s+\mu)T(x,s) &= D\partial_{xx} T(x,s) \\ T(x,s) &= \cancel{Ae^{\sqrt{\frac{s+\mu}{D}}x}} + Be^{-\sqrt{\frac{s+\mu}{D}}x}\,. \end{aligned}$$

Then

$$G(s) = \frac{\text{output}}{\text{input}} = \frac{T(\ell_2,s)}{T(\ell_1,s)} = \frac{B\,\mathrm{e}^{-\sqrt{\frac{s+\mu}{D}}\ell_2}}{B\,\mathrm{e}^{-\sqrt{\frac{s+\mu}{D}}\ell_1}} = \mathrm{e}^{-\sqrt{\frac{s+\mu}{D}}\Delta\ell}\,.$$

b. **Eliminating μ**. As suggested, the easy way to do this is to write

$$(\ln G)^2 = \frac{(\Delta\ell)^2}{D}(\mu + \mathrm{i}\omega)\,.$$

But

$$(\ln G)^2 = (\ln|G| + \mathrm{i}\theta)^2 = (\text{real terms}\ldots) + 2\,\mathrm{i}(\ln|G|)\theta\,.$$

Equating the imaginary parts then gives

$$2\ln|G|\theta = \frac{(\Delta\ell)^2\,\omega}{D}\,.$$

The naive way of proceeding, by calculating explicit expressions for $\ln|G|$ and θ, is much harder.

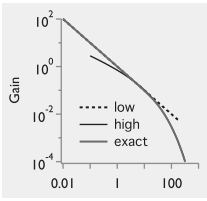
2.6 From lumped-element circuits to infinite objects. A physical object, such as the one-dimensional conductor considered in Problem 2.4, can show three qualitatively different types of behavior depending on ω, the signal frequency (Frick et al., 2018).

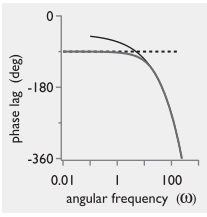
a. Derive the transfer function for thermal conduction of a one-dimensional material of finite length L, with insulating boundary conditions at $x = L$ and

a probe at $x = l_0$. Following Problem 2.4 and scaling L to one, show that $G(s,\ell) = \frac{\cosh\sqrt{s}(1-\ell)}{\sqrt{s}\sinh\sqrt{s}}$.

b. Simplify the transfer function at low frequencies ($s \to 0$) and high frequencies ($s \to \infty$). Be precise about what sets the scale that defines low and high frequencies. For each limit, express your result in dimensional as well as dimensionless units, and interpret. (Hint: for dimensional units, use $D = \lambda/(\rho C_p)$.)

c. Reproduce the Bode plot at left of the exact solution and its two limits, which is plotted for $\ell = l_0/L = 0.5$. Explore other values of ℓ.





Solution.

a. From Eq. (2.38), the equation of motion is

$$\partial_t T = D\,\partial_{xx} T\,,$$

with boundary conditions

$$\begin{aligned} -\lambda\,\partial_x T(x=L,t) &= 0\,,\\ -\lambda\,\partial_x T|_{(x=0,t)} &= P(t)/a\\ T(x,0) &= T_0\,, \end{aligned}$$

As before, we scale $x \to x/L$, $t \to t/\,(L^2/D)$, $l_0 \to \ell$, $T \to (T-T_0)/T_0$, and define $u = PL/(\lambda a T_0)$, with a the heater area, which gives

$$\begin{aligned} \partial_t T &= \partial_{xx} T\,, & y &= T(\ell,t)\,,\\ \partial_x T|_{(x=1,t)} &= 0\,, & -\partial_x T|_{(x=0,t)} &= u(t)\\ T(x,0) &= 0\,, \end{aligned}$$

We Laplace transform in time:

$$sT = \partial_{xx} T \quad\Longrightarrow\quad T(s,x) = A\,\mathrm{e}^{\sqrt{s}x} + B\,\mathrm{e}^{-\sqrt{s}x}\,.$$

Imposing the boundary condition at $x = 1$ gives

$$\partial_x T|_{(x=1,t)}\,\sqrt{s}\left(A\,\mathrm{e}^{\sqrt{s}} - B\,\mathrm{e}^{-\sqrt{s}}\right) = 0 \quad\Longrightarrow\quad B = A\,\mathrm{e}^{2\sqrt{s}}\,.$$

Imposing the boundary condition at $x = 0$ gives

$$-\partial_x T|_{(x=0,t)}\,\sqrt{s}(A-B) = u(s) \quad\Longrightarrow\quad A\,\sqrt{s}\left(\mathrm{e}^{2\sqrt{s}} - 1\right) = u(s)\,.$$

The transfer function from $x = 0$ to the point $x = \ell$ is then

$$\begin{aligned} G(s) = \frac{y(s)}{u(s)} = \frac{T(\ell,s)}{u(s)} &= \frac{\mathrm{e}^{\sqrt{s}\ell} + \mathrm{e}^{2\sqrt{s}}\,\mathrm{e}^{-\sqrt{s}\ell}}{\sqrt{s}\left(\mathrm{e}^{2\sqrt{s}} - 1\right)}\\ &= \frac{\cosh\sqrt{s}(1-\ell)}{\sqrt{s}\sinh\sqrt{s}}\,. \end{aligned}$$

Note that, in our scaling, $\ell < 1$.

b. Low- and high-frequency limits.
   i. In the low-frequency limit, $\cosh\sqrt{s}(1-\ell) \to 1$ and $\sinh\sqrt{s} \to \sqrt{s}$, so that

$$G(s) \to \frac{1}{s}\,,$$

   which is a pure integrator. Putting back in dimensional units and recalling that $D = \lambda/(\rho C_\mathrm{p})$ gives

$$G(s) \to \frac{D}{L^2}\,\frac{LT_0}{\lambda a T_0}\,\frac{1}{s} = \frac{1}{La\rho C_\mathrm{p}}\,\frac{1}{s} = \frac{1}{C_\mathrm{p} M}\,\frac{1}{s}\,.$$

   where $M = \rho(aL)$ is the mass of the object and $C_\mathrm{tot} = C_\mathrm{p} M$ its total heat capacity. Going back to the time domain, this is equivalent to

$$C_\mathrm{tot}\frac{\mathrm{d}T}{\mathrm{d}t} = P\,,$$

   which is one of the elementary equations describing thermal circuits where the heat capacity of the object plays the role of a capacitor.

   ii. In the high-frequency limit, $\cosh x \to \frac{1}{2}\,\mathrm{e}^x$ and $\sinh x \to \frac{1}{2}\,\mathrm{e}^x$, so that

$$G(s) \to \frac{\mathrm{e}^{\sqrt{s}(1-\ell)}}{\sqrt{s}\;\mathrm{e}^{\sqrt{s}}} = \frac{\mathrm{e}^{-\sqrt{s}\ell}}{\sqrt{s}}\,.$$

   The difference from our previous result comes because we chose our scaling there to make $\ell = 1$.

   To be in the high-frequency regime requires

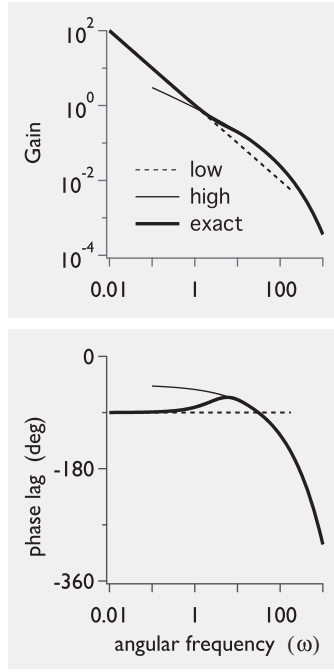
$$\mathrm{e}^{-\sqrt{s}(1-\ell)} \ll \mathrm{e}^{\sqrt{s}(1-\ell)} \quad\Longrightarrow\quad \mathrm{e}^{-2\sqrt{s}(1-\ell)} \ll 1 \quad\Longrightarrow\quad 2\sqrt{s}(1-\ell) \gg 1\,.$$

   In dimensional units, this implies

$$4\omega\frac{L^2}{D}\left(1-\frac{\ell}{L}\right)^2 \gg 1 \quad\Longrightarrow\quad \omega \gg \frac{D}{4L^2}\frac{1}{\left(1-\frac{\ell}{L}\right)^2} \to \frac{D}{4L^2}\,.$$

   The latter limit requires $\ell \ll L$, but it is interesting to see that the probe can be anywhere, as long as the first (more restrictive) frequency limit applies. An interesting numerical limit occurs for $\ell = L/2$ (probe in the middle of the rod), where we require $\omega \gg \frac{D}{L^2}$.
   In the low-frequency regime, simply change $\gg$ to $\ll$.

c. The numerics consist in simply plotting the magnitude and phase response. Note that many pre-packaged "Bode Plot" routines will not work with non-rational arguments such as $\sqrt{s}$. But it is easy to plot the magnitude and phase of an arbitrary complex function. The graph in the text is for $\ell/L = 0.5$. For lower values of ℓ, there is a pronounced peak in the phase response, as illustrated below for $\ell/L = 0.2$. Higher values of ℓ are similar to $\ell/L = 0.5$.



Qualitatively, we can interpret the three regimes as a comparison between the time it takes heat to diffuse across the object, $\frac{L^2}{D}$, to the period of heater oscillations, $\sim \omega^{-1}$. In the lumped-element regime, heat diffuses across the element so quickly that the temperature of the object rigidly follows the heater. In the infinite limit, temperature fluctuations produced at the boundary damp out before they reach the other side, so that the two boundaries do not influence each other. The mathematics is simple in the two extreme limits and messier in the intermediate, finite case. While this problem deals with thermal conduction, analogous results hold in all parts of physics. For example, the familiar resistors, capacitors, and inductors of elementary circuit theory behave as lumped elements with respect to Maxwell's equations. In this case, one compares the period $\sim \omega^{-1}$ to the time it takes light to cross the object, $\frac{L}{c}$.

As an aside, engineering texts[1] often define a dimensionless *Biot number*, Bi $\equiv hL/\lambda$. The coefficient h is known as the *heat transfer coefficient* and has units of W/(m^{2}K). In general, it can include contributions from thermal convection and radiation. In a simple situation, we can model it crudely as coming a conduction-like term $h = J/\Delta T = P/(a\Delta T)$, which is the heat flux divided by the temperature at the heater relative to the temperature of the surrounding environment. In this language, Bi $\ll 1$ implies that the lumped-element approximation is valid. In the

[1] See, for example, T. L. Bergman and A. S. Lavine, *Fundamentals of Heat and Mass Transfer*, Wiley, 7th ed., 2011, Sections 5.1 and 5.2.

language developed in this problem, the lumped-element approximation is valid for $|G(s)| \approx 1$. In our scaling, this means that

$$\underbrace{\frac{\Delta T}{T_0}}_{y}\,\underbrace{\frac{\lambda a T_0}{PL}}_{u^{-1}} = \frac{\Delta T a \lambda}{PL} = \frac{\lambda}{hL} = \mathrm{Bi}^{-1} \gg 1\,,$$

which is equivalent to the engineering criterion. For more discussion, see also Frick et al. (2018).

2.7 **Example 2.6**. Give the steps in the calculations.

Solution.

We first find the eigenvalues of $\boldsymbol{A} = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$. The condition $\det(\lambda\mathbb{I} - \boldsymbol{A}) = 0$ gives

$$\det(\lambda\mathbb{I} - \boldsymbol{A}) = \begin{vmatrix} \lambda & -1 \\ 1 & \lambda \end{vmatrix} = \lambda^2 + 1 = 0\,, \quad \Longrightarrow \quad \lambda = \pm \mathrm{i}\,.$$

The eigenvalue for $\lambda = +i$ is given by

$$\begin{pmatrix} \mathrm{i} & -1 \\ 1 & \mathrm{i} \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \Longrightarrow \quad \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ \mathrm{i} \end{pmatrix},$$

where the $1/\sqrt{2}$ normalizes the eigenvector to be a unit vector. A similar calculation for $\lambda = -\mathrm{i}$ gives $\frac{1}{\sqrt{2}}\left(\begin{smallmatrix} 1 \\ -\mathrm{i} \end{smallmatrix}\right)$ for the other eigenvector. Putting them together gives

$$\boldsymbol{R} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{pmatrix} \quad \Longrightarrow \quad \boldsymbol{R}^{-1} = \boldsymbol{R}^{\dagger} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -\mathrm{i} \\ 1 & \mathrm{i} \end{pmatrix}.$$

Thus,

$$\boldsymbol{A} = \boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^{\dagger} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{pmatrix}\begin{pmatrix} \mathrm{i} & 0 \\ 0 & -\mathrm{i} \end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -\mathrm{i} \\ 1 & \mathrm{i} \end{pmatrix}$$

and

$$\mathrm{e}^{\boldsymbol{A}t} = \boldsymbol{R}\left(\mathrm{e}^{\boldsymbol{D}t}\right)\boldsymbol{R}^{\dagger} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{pmatrix}\begin{pmatrix} \mathrm{e}^{\mathrm{i}t} & 0 \\ 0 & \mathrm{e}^{-\mathrm{i}t} \end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -\mathrm{i} \\ 1 & \mathrm{i} \end{pmatrix} = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}.$$

Once we understand how to do such calculations, it is usually easier to do them numerically or symbolically by computer.

2.8 **Example 2.7**. Check the matrix calculations.

Solution.

We first investigate

$$G(s) = \underbrace{\begin{pmatrix} 1 & 0 \end{pmatrix}}_{\boldsymbol{C}}\left[s\underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{\mathbb{I}} - \underbrace{\begin{pmatrix} 0 & 1 \\ -1 & -2\zeta \end{pmatrix}}_{\boldsymbol{A}}\right]^{-1}\underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{\boldsymbol{B}}.$$

The middle matrix is

$$\begin{pmatrix} s & -1 \\ 1 & s+2\zeta \end{pmatrix}^{-1} = \frac{1}{s^2+2\zeta s+1}\begin{pmatrix} s+2\zeta & 1 \\ -1 & s \end{pmatrix}$$

and, thus,

$$G(s) = \frac{1}{s^2+2\zeta s+1}\begin{pmatrix}1 & 0\end{pmatrix}\begin{pmatrix} s+2\zeta & 1 \\ -1 & s \end{pmatrix}\begin{pmatrix}0\\1\end{pmatrix} = \frac{1}{s^2+2\zeta s+1}\,.$$

Similarly, if we observe *both* position and velocity,

$$\begin{aligned} G(s) &= \frac{1}{s^2+2\zeta s+1}\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\begin{pmatrix} s+2\zeta & 1 \\ -1 & s \end{pmatrix}\begin{pmatrix}0\\1\end{pmatrix} \\ &= \frac{1}{s^2+2\zeta s+1}\begin{pmatrix} s+2\zeta & 1 \\ -1 & s \end{pmatrix}\begin{pmatrix}0\\1\end{pmatrix} \\ &= \frac{1}{s^2+2\zeta s+1}\begin{pmatrix}1\\s\end{pmatrix}. \end{aligned}$$

2.9 Critically damped harmonic oscillator. Confirm Eq. (2.68)b that the response of a critically damped harmonic oscillator to a unit-velocity "kick" at time $t = 0$ is $y(t) = t\,\mathrm{e}^{-t}$. Hint: Write the dynamical matrix $\boldsymbol{A} = \left(\begin{smallmatrix} 0 & 1 \\ -1 & -2 \end{smallmatrix}\right)$ as $(-1)[\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right) + \left(\begin{smallmatrix} -1 & -1 \\ 1 & 1 \end{smallmatrix}\right)]$ and exponentiate directly. See Problem 13.4a for a more sophisticated approach.

Solution.

Decompose $\boldsymbol{A} = (-1)\,(\mathbb{I} + \boldsymbol{J})$, where $\mathbb{I}$ is the 2×2 identity matrix and

$$\boldsymbol{J} = \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix} \quad \Longrightarrow \quad \boldsymbol{J}^2 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Thus,

$$\begin{aligned} \mathrm{e}^{\boldsymbol{A}t} = \mathrm{e}^{(-t)(\mathbb{I}+\boldsymbol{J})} &= \sum_{n=0}^{\infty}\frac{(-t)^n}{n!}(\mathbb{I}+\boldsymbol{J})^n = \sum_{n=0}^{\infty}\frac{(-t)^n}{n!}(\mathbb{I}+n\boldsymbol{J}) \\ &= \mathrm{e}^{-t}\,\mathbb{I} + \underbrace{\sum_{n=1}^{\infty}\frac{(-t)^n}{(n-1)!}}_{-t\,\mathrm{e}^{-t}}\boldsymbol{J} \\ &= \mathrm{e}^{-t}(\mathbb{I} - t\boldsymbol{J}) \\ &= \mathrm{e}^{-t}\begin{pmatrix} 1+t & t \\ -t & 1-t \end{pmatrix}. \end{aligned}$$

The position response to a unit kick is then given by

$$y(t) = \begin{pmatrix}1 & 0\end{pmatrix}\mathrm{e}^{-t}\begin{pmatrix} 1+t & t \\ -t & 1-t \end{pmatrix}\begin{pmatrix}0\\1\end{pmatrix} = t\,\mathrm{e}^{-t}\,.$$

2.10 Converting transfer functions to canonical state-space form. Show:

a. State-space equations in control-canonical form, Eq. (2.59) correspond to the strictly proper transfer function, Eq. (2.57).
b. A "merely proper" transfer function can still be written in control-canonical form.

$$G(s) = \frac{b_0 s^n + b_1 s^{n-1} + \cdots + b_n}{s^n + a_1 s^{n-1} + \cdots + a_n} \quad \longrightarrow \quad \left(\begin{array}{c:c} \boldsymbol{A} & \boldsymbol{B} \\ \hdashline \boldsymbol{C} & \boldsymbol{D} \end{array}\right),$$

Solution.

a. The claim is that

$$G(s) = \frac{b_0 s^k + b_1 s^{k-1} + \cdots + b_k}{s^n + a_1 s^{n-1} + \cdots + a_n} \equiv \frac{b(s)}{a(s)}\,.$$

is equivalent to

$$\dot{\boldsymbol{x}} = \underbrace{\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_2 & -a_1 \end{pmatrix}}_{\boldsymbol{A}} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} + \underbrace{\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}}_{\boldsymbol{B}} u$$

$$y = \underbrace{\begin{pmatrix} b_k & b_{k-1} & \cdots & b_1 & b_0 & 0 & \cdots & 0 \end{pmatrix}}_{\boldsymbol{C}} \boldsymbol{x}\,.$$

We have $\dot{x}_1 = x_2$, $\ddot{x}_1 = x_3$, $x_1^{(3)} = x_4$, and $x_1^{(n-1)} = x_n$ and

$$\frac{\mathrm{d}^n x}{\mathrm{d}t^n} = -a_n x_1 - a_{n-2}\frac{\mathrm{d}x_1}{\mathrm{d}t} - \cdots - a_1 \frac{\mathrm{d}^{n-1} x_1}{\mathrm{d}t^{n-1}} + u(t)\,.$$

Then, taking the Laplace transform gives,

$$\left(s^n + a_1 s^{n-1} + a_2 s^{n-2} + \cdots + a_{n-1} s + a_n\right) x_1(s) = u(s)\,.$$

Conversely, taking the Laplace transform of

$$\begin{aligned} y(t) &= b_k x_1(t) + b_{k-1} x_2 + \cdots b_0 x_{k+1} \quad (k < n) \\ &= b_k x + b_{k-1}\dot{x} + \cdots b_0 x^{(k)} \end{aligned}$$

leads to

$$y(s) = \left(b_k + b_z k - 1 s + \cdots + b_0 s^k\right) x(s)\,.$$

Solving for $x(s)$ and substituting gives the original expression for $G(s)$.

b. We show that the "merely proper" transfer function

$$G(s) = \frac{b_0 s^n + b_1 s^{n-1} + \cdots + b_n}{s^n + a_1 s^{n-1} + \cdots + a_n} \quad \longrightarrow \quad \left(\begin{array}{c:c} \boldsymbol{A} & \boldsymbol{B} \\ \hdashline \boldsymbol{C} & \boldsymbol{D} \end{array}\right)$$

can still be written in control-canonical ***ABCD*** state-space form,

$$\left(\begin{array}{cccccc:c} 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_2 & -a_1 & 1 \\ \hdashline b_n - a_n b_0 & b_{n-1} - a_{n-1} b_0 & b_{n-2} - a_{n-2} b_0 & \cdots & b_2 - a_2 b_0 & b_1 - a_1 b_0 & b_0 \end{array}\right) .$$

To see this, write the transfer function as the sum of a constant term (representing the feedthrough) and a strictly proper transfer function, so that the rules given in Section 2.4.1 apply. Thus,

$$\begin{aligned} G(s) &= \frac{b_0 s^n + b_1 s^{n-1} + \cdots + b_n}{s^n + a_1 s^{n-1} + \cdots + a_n} \\ &= \frac{b_1 s^{n-1} + \cdots + b_n}{s^n + a_1 s^{n-1} + \cdots + a_n} \\ &\quad + b_0 \left(\frac{s^n}{s^n + a_1 s^{n-1} + \cdots + a_n} + 1 - \frac{s^n + a_1 s^{n-1} + \cdots + a_n}{s^n + a_1 s^{n-1} + \cdots + a_n} \right) \\ &= \frac{b_1 s^{n-1} + \cdots + b_n}{s^n + a_1 s^{n-1} + \cdots + a_n} + b_0 \left(1 - \frac{a_1 s^{n-1} + \cdots + a_n}{s^n + a_1 s^{n-1} + \cdots + a_n} \right) \\ &= \frac{(b_1 - a_1 b_0) s^{n-1} + \cdots + (b_n - a_n b_0)}{s^n + a_1 s^{n-1} + \cdots + a_n} + b_0 . \end{aligned}$$

We can then convert the biproper transfer function to a state-space representation using the results from (a). The feedthrough "matrix" $\boldsymbol{D}$ is just the scalar b_0. (For a MIMO system, it could be a matrix.)

Notice that setting $b_0 = 0$ gives back the result in (a), which is also Eq. (2.59).

2.11 Feedthrough as unmodeled dynamics. Consider highly overdamped motion, with a small term multiplying the highest derivative (*singular perturbation*): $\varepsilon \ddot{x} + \dot{x} + x = u$ and $y = \dot{x}$, where $u(t)$ is the input, the output $y(t)$ is the *velocity*, and the mass $\varepsilon \ll 1$.

a. Transform to a two-dimensional state-space form with $\boldsymbol{x} = (x_1 \;\; x_2)^{\mathsf{T}}$. Find the two poles, to lowest order in ε. One mode should be slow, the other fast.

b. Solve for $x_2(t)$ in the quasistatic limit, $\varepsilon \dot{x}_2 \approx 0$, and show that the reduced, one-dimensional state-space equations have an output with feedthrough $D = 1$.

Solution.

a. In standard state-space form, the equations are

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_1 - x_2 + u , \qquad y = x_2 . \end{aligned}$$

From the characteristic equation, $\varepsilon s^2 + s + 1 = 0$, the two modes have poles at $s \approx -1$ and $s \approx \varepsilon^{-1}$, to lowest order in ε. The former is the slow mode, the latter the fast mode, since its decay rate $\varepsilon^{-1} \gg 1$.

b. In the quasistatic limit $\varepsilon x_2 \approx 0$, which implies

$$x_2 \approx -x_1 + u\,.$$

Substituting gives a reduced equation in the one-dimensional space of $x_1(t)$. Explicitly, we have

$$\dot{x}_1 \approx -x_1 + u\,, \qquad y \approx -x_1 + u\,.$$

The state-space "matrices" are all constants: $\{A, B, C, D\} = \{-1, +1, -1, +1\}$. We see that a feedthrough term, $D = 1$, has arisen as a result of transforming the output. To understand this more physically, we note that the output is the *fast* mode. In the quasistatic limit, we take it to be infinitely rapid. Then it makes sense that the input $u(t)$ appears directly in the output $y(t)$, as the signal propagates infinitely rapidly through the "dynamical" system. Note that had we added, as would be likely, a second output for x_1, it would not have been affected. It is only the fast modes that contribute to the feedthrough.

We thus see how a finite feedthrough matrix $\boldsymbol{D}$ can represent fast modes that have been adiabatically eliminated via the quasistatic approximation. Notice that if we observed only the position and not the velocity, we would have a state-space model that is fully proper.

2.12 Invariance of the transfer function.

a. Show that the transfer function is invariant under coordinate transformation:

$$G(s) = \boldsymbol{C}\,(s\mathbb{I} - \boldsymbol{A})^{-1}\,\boldsymbol{B} = \boldsymbol{C}'\,(s\mathbb{I} - \boldsymbol{A}')^{-1}\,\boldsymbol{B}'\,.$$

b. Consider the second-order system $\boldsymbol{A} = \left(\begin{smallmatrix} 0 & 1 \\ -1 & -2\zeta \end{smallmatrix}\right)$, $\boldsymbol{B} = \left(\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right)$, $\boldsymbol{C} = (1\ 0)$, and transformation $\boldsymbol{T} = \frac{1}{\sqrt{2}}\left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right)$. Physically, $\boldsymbol{T}$ corresponds to a 45° rotation in the x_1-x_2 plane. Find new matrices $\{\boldsymbol{A}', \boldsymbol{B}', \text{and } \boldsymbol{C}'\}$ and show that the corresponding transfer function calculated is the same as that calculated using $\{\boldsymbol{A}, \boldsymbol{B}, \text{and } \boldsymbol{C}\}$.

Solution.

a. *Proving invariance*:

$$\begin{aligned} G &= \boldsymbol{C}\,(s\mathbb{I} - \boldsymbol{A})^{-1}\,\boldsymbol{B} \\ &= \boldsymbol{C}'\boldsymbol{T}\left(s\boldsymbol{T}^{-1}\boldsymbol{T} - \boldsymbol{T}^{-1}\boldsymbol{A}'\boldsymbol{T}\right)^{-1}\,\boldsymbol{T}^{-1}\boldsymbol{B}' \\ &= \boldsymbol{C}'\boldsymbol{T}\left(\boldsymbol{T}^{-1}\,(s\mathbb{I} - \boldsymbol{A}')\,\boldsymbol{T}\right)^{-1}\,\boldsymbol{T}^{-1}\boldsymbol{B}' \\ &= \boldsymbol{C}'\boldsymbol{T}\boldsymbol{T}^{-1}\,(s\mathbb{I} - \boldsymbol{A}')^{-1}\,\boldsymbol{T}\boldsymbol{T}^{-1}\boldsymbol{B}' \\ &= \boldsymbol{C}'\,(s\mathbb{I} - \boldsymbol{A}')^{-1}\,\boldsymbol{B}' \end{aligned}$$

b. *2nd-order example*: This is a good problem to use a computer algebra program such as Mathematica. Using that program, I get

$$\boldsymbol{A}' = \begin{pmatrix} -\zeta & \zeta+1 \\ \zeta-1 & -\zeta \end{pmatrix}, \quad \boldsymbol{B}' = \frac{1}{\sqrt{2}}\begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \boldsymbol{C}' = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \end{pmatrix}.$$

Both sets of matrices give $G(s) = \boldsymbol{C}\,(s\mathbb{I} - \boldsymbol{A})^{-1}\,\boldsymbol{B} = \boldsymbol{C}'\,(s\mathbb{I} - \boldsymbol{A}')^{-1}\,\boldsymbol{B}' = \frac{1}{1+2\zeta s+s^2}$.

2.13 First-order systems, step and impulse response. Derive the step and impulse response from Section 2.4.4 for the first-order system $G(s) = (1 + s)^{-1}$. Do it directly in the time domain and also by Laplace transforms. For responses, the system is in equilibrium at $x = 0$ for $t < 0$. For the step response, see Eq. (2.66) or Footnote 22.

Solution.

a. *Step response, time domain*. We solve $\dot{y} + y = \theta(t)$, with initial condition $y(0) = 0$, to find

$$y(t) = 1 - \mathrm{e}^{-t}, \qquad t > 0\,.$$

b. *Impulse response, time domain*. We have to be careful in how we specify the initial conditions. The equation of motion is

$$\dot{y} + y = \delta\,(t)\,, \qquad y(0) = 0\,.$$

We integrate this equation from $t = -\epsilon$ to $t = +\epsilon$. The integral over $y(t)$ gives zero, but the integral over the derivative gives

$$\lim_{\epsilon\to 0}\int_{-\epsilon}^{\epsilon} \mathrm{d}t\,(\dot{y} + y) = \lim_{\epsilon\to 0}\int_{-\epsilon}^{\epsilon} \mathrm{d}t\,\delta\,(t) = 1$$
$$\lim_{\epsilon\to 0}\left[y(\epsilon) - \cancel{y(-\epsilon)}\right] = 1\,,$$

In other words, there is a discontinuity in $y(t)$ at $t = 0$, and, for $t \geq 0$, we solve $\dot{y} + y = 0$, with $y(0^+) = 1$ as our initial condition. The solution is then

$$y(t) = \mathrm{e}^{-t}\;\theta(t)\,.$$

Note that, in this calculation, we wrote

$$\lim_{\epsilon\to 0}\int_{-\epsilon}^{\epsilon} \mathrm{d}t\,y = 0\,.$$

The general justification is that, in general, $\int_{-\epsilon}^{\epsilon} \mathrm{d}t\,y \approx y(0)(2\epsilon) \to 0$ when $\epsilon \to 0$. You might worry whether the finite discontinuity in $y(t)$ at 0 changes this result. It does not: you can divide the integral into $(-\epsilon, 0)$ and $(0, \epsilon)$ portions and argue that on each of these sub-intervals, the integral vanished (when $\epsilon \to 0$), so that the argument holds for finite jump discontinuities.

c. *Impulse response, via Laplace*: We have

$$G(s) = C(s - A)^{-1}B = (1)\,\frac{1}{1+s}\,(1) = \frac{1}{1+s}\,,$$

which corresponds to $G(t) = \mathrm{e}^{-t}$.

d. *Step response, via Laplace*: We have

$$y(s) = G(s)\,\frac{1}{s} = \frac{1}{s(s+1)} = \frac{1}{s} - \frac{1}{s+1}\,,$$

which corresponds to $y(t) = 1 - \mathrm{e}^{-t}$.

2.14 Lyapunov function for linear dynamics, 1. Let $\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x}$, with all the eigenvalues of $\boldsymbol{A}$ having a negative real part, and let $\boldsymbol{Q}$ be an arbitrary $n \times n$ positive definite matrix.

a. Show $V(\boldsymbol{x}) = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{x}$ is a Lyapunov function, where $\boldsymbol{P} = \int_0^\infty \mathrm{d}t\, \mathrm{e}^{\boldsymbol{A}^{\mathsf{T}}t}\, \boldsymbol{Q}\, \mathrm{e}^{\boldsymbol{A}t}$ and $\boldsymbol{P}$ satisfies the *Lyapunov equation*, $\boldsymbol{A}^{\mathsf{T}}\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} = -\boldsymbol{Q}$.

b. If we can choose $\boldsymbol{Q}$, then Lyapunov functions are not unique. For one-dimensional dynamics $\dot{x} = -x$ and arbitrary $Q > 0$, construct P, show that it obeys the Lyapunov equation, and find V. Why does an arbitrary, positive Q work?

Solution.

a. As suggested, we first show that if $\boldsymbol{P} = \int_0^\infty \mathrm{d}t\, \mathrm{e}^{\boldsymbol{A}^{\mathsf{T}}t}\, \boldsymbol{Q}\, \mathrm{e}^{\boldsymbol{A}t}$ for stable dynamics, then it satisfies the Lyapunov Equation, $\boldsymbol{A}^{\mathsf{T}}\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} = -\boldsymbol{Q}$:

$$\begin{aligned}
\boldsymbol{A}^{\mathsf{T}}\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} &= \int_0^\infty \mathrm{d}t \left(\boldsymbol{A}^{\mathsf{T}}\, \mathrm{e}^{\boldsymbol{A}^{\mathsf{T}}t}\, \boldsymbol{Q}\, \mathrm{e}^{\boldsymbol{A}t} + e^{\boldsymbol{A}^{\mathsf{T}}t}\, \boldsymbol{Q}\, \mathrm{e}^{\boldsymbol{A}t}\, \boldsymbol{A}\right) \\
&= \int_0^\infty \mathrm{d}t\, \frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{e}^{\boldsymbol{A}^{\mathsf{T}}t}\, \boldsymbol{Q}\, \mathrm{e}^{\boldsymbol{A}t}\right) \\
&= \mathrm{e}^{\boldsymbol{A}^{\mathsf{T}}t}\, \boldsymbol{Q}\, \mathrm{e}^{\boldsymbol{A}t}\Big|_0^\infty \\
&= -\boldsymbol{Q}\,,
\end{aligned}$$

where the term at $t = \infty$ goes to zero since "stable dynamics" implies that all the eigenvalues of $\boldsymbol{A}$ have negative real part.

Then we show that V is indeed a Lyapunov function:

$$\frac{\mathrm{d}V}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{x}\right) = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{x} + \boldsymbol{x}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^{\mathsf{T}}(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A})\boldsymbol{x} = -\boldsymbol{x}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{x} \leq 0\,.$$

The last step follows because $\boldsymbol{Q}$ is positive definite. Since the derivative equals zero only when $\boldsymbol{x} = \boldsymbol{0}$, the solution $\boldsymbol{x} = \boldsymbol{0}$ is globally stable.

b. For the one-dimensional problem, we have

$$P = \int_0^\infty \mathrm{d}t\, \mathrm{e}^{-t}\, Q\, \mathrm{e}^{-t} = Q \int_0^\infty \mathrm{d}t\, \mathrm{e}^{-2t} = \frac{Q}{2}\,.$$

Then $V = Px^2 = \frac{1}{2}Qx^2$ and

$$\frac{\mathrm{d}V}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}Qx^2\right) = Qx\dot{x} = -Qx^2 \,.$$

Clearly, this works for any $Q > 0$.

2.15 Lyapunov function for linear dynamics, 2. Consider two-dimensional linear dynamics, with $\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x}$ and Lyapunov equation $\boldsymbol{A}^{\mathsf{T}}\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} = -\boldsymbol{Q}$, with $\boldsymbol{A} = \left(\begin{smallmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{smallmatrix}\right)$, $\boldsymbol{Q} = \left(\begin{smallmatrix} q_1 & 0 \\ 0 & q_2 \end{smallmatrix}\right)$, $\boldsymbol{P} = \left(\begin{smallmatrix} p_a & p_b \\ p_b & p_c \end{smallmatrix}\right)$. Solve the linear matrix equation with a trick:

a. Define the *vector* $\boldsymbol{p}^{\mathsf{T}} = (p_a \;\; p_b \;\; p_c)$. Then write down and solve the corresponding 3×3 matrix version of the Lyapunov equation for $\boldsymbol{p}$.
b. Write $V(\boldsymbol{x})$ explicitly in terms of the elements of $\boldsymbol{A}$ and $\boldsymbol{Q}$ for $a_{12} = a_{21} = 0$.

Solution.

a. By direct inspection (or via a computer-algebra program), we can find an equation of the form $\boldsymbol{M}\boldsymbol{p} = \boldsymbol{v}$, with

$$\underbrace{\begin{pmatrix} 2a_{11} & 2a_{21} & 0 \\ a_{12} & a_{11}+a_{22} & a_{21} \\ 0 & 2a_{12} & 2a_{22} \end{pmatrix}}_{\boldsymbol{M}} \underbrace{\begin{pmatrix} p_a \\ p_b \\ p_c \end{pmatrix}}_{\boldsymbol{p}} = \underbrace{\begin{pmatrix} -q_1 \\ 0 \\ -q_2 \end{pmatrix}}_{\boldsymbol{q}} .$$

The results in Problem 2.14 then guarantee that we can invert to find $\boldsymbol{p} = \boldsymbol{M}^{-1}\boldsymbol{v}$, as long as $q_1, q_2 > 0$ and the eigenvalues of $\boldsymbol{A}$ have negative real part. The solution is

$$\boldsymbol{p} = \begin{pmatrix} p_a \\ p_b \\ p_c \end{pmatrix} = \frac{1}{2\mathrm{Tr}\,\det} \begin{pmatrix} a_{12}\,a_{21}\,q_1 - a_{22}\,(\mathrm{Tr})\,q_1 - a_{21}^2\,q_2 \\ a_{12}\,a_{22}\,q_1 + a_{11}\,a_{21}\,q_2 \\ a_{12}\,a_{21}\,q_2 - a_{11}\,(\mathrm{Tr})\,q_2 - a_{12}^2\,q_1 \end{pmatrix},$$

where $\mathrm{Tr} = a_{11} + a_{22} = \lambda_1 + \lambda_2 < 0$ and $\det = a_{11}\,a_{22} - a_{12}\,a_{21} = \lambda_1\,\lambda_2 > 0$. For completeness, we rewrite the solution in terms of the original matrix $\boldsymbol{P}$:

$$\boldsymbol{P} = \frac{1}{2\mathrm{Tr}\,\det} \begin{pmatrix} a_{12}\,a_{21}\,q_1 - a_{22}\,(\mathrm{Tr})\,q_1 - a_{21}^2\,q_2 & a_{12}\,a_{22}\,q_1 + a_{11}\,a_{21}\,q_2 \\ a_{12}\,a_{22}\,q_1 + a_{11}\,a_{21}\,q_2 & a_{12}\,a_{21}\,q_2 - a_{11}\,(\mathrm{Tr})\,q_2 - a_{12}^2\,q_1 \end{pmatrix}.$$

b. The Lyapunov function is $V(\boldsymbol{x}) = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{x} = p_a x_1^2 + 2p_b x_1 x_2 + p_c x_2^2$. This is a bit messy in general but simplifies nicely for diagonal $\boldsymbol{A}$:

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix} \quad \Longrightarrow \quad V = \frac{1}{2}\left(q_1 \frac{x_1^2}{-a_{11}} + q_2 \frac{x_2^2}{-a_{22}}\right).$$

We see that we need stable dynamics ($a_{11} < 0$ and $a_{22} < 0$) for V to be a Lyapunov function – i.e., for $V(\boldsymbol{x})$ to be positive definite. This makes sense, since the Lyapunov function is used to prove stability! As in Problem 2.14, we see that any positive definite $\boldsymbol{Q}$ will work.

2.16 Lyapunov and the damped pendulum. Some Lyapunov functions are more useful than others. Consider a damped pendulum whose angle $\theta(t)$ obeys $\ddot{\theta} + \dot{\theta} + \sin\theta = 0$.

a. Show that the total energy V_1 is a negative semidefinite Lyapunov function.
b. Use the Krasovskii–LaSalle Invariance Principle to conclude from the analysis of V_1 that the down position is locally asymptotically stable.
c. Show that another Lyapunov function $V_2 = \frac{1}{2}\dot{\theta}^2 + \frac{1}{2}(\dot{\theta} + \theta)^2 + 2(1 - \cos\theta)$ is locally negative definite. Discuss the local stability of the down equilibrium.

Solution.

a. The energy $E \equiv V_1$ is given by

$$V_1 = \frac{1}{2}\dot{\theta}^2 + (1 - \cos\theta) \, .$$

Then

$$\dot{V}_1 = \dot{\theta}\,\ddot{\theta} + \sin\theta\,\dot{\theta} = \dot{\theta}\left(\ddot{\theta} + \sin\theta\right) = -\dot{\theta}^2 \leq 0 \, .$$

Note that $\dot{V}_1 = 0$ for $\boldsymbol{x} \neq \boldsymbol{0}$. Physically, $\dot{V}$ is the power dissipated by the moving pendulum.

b. At the turning points of each swing of a pendulum, $\dot{\theta} = 0$ and $\theta \neq 0$. Thus, at these moments, $\dot{V} = 0$. Notice that the linearized damped pendulum (second-order system) will have the same features. Thus, the same reasoning as given in Example 2.10 shows that the origin is asymptotically stable. It is not globally asymptotically stable, as there are multiple equilibria ($\theta = 0$ or π).

c. The alternate choice, which has no obvious physical motivation, is

$$V_2 = \frac{1}{2}\dot{\theta}^2 + \frac{1}{2}\left(\dot{\theta} + \theta\right)^2 + 2(1 - \cos\theta) \, .$$

It is clearly positive definite: $V \geq 0$, except for $\theta = \dot{\theta} = 0$, where $V = 0$. Differentiating gives

$$\begin{aligned}
\dot{V}_2 &= \dot{\theta}\,\ddot{\theta} + (\dot{\theta} + \theta)\,(\ddot{\theta} + \dot{\theta}) + 2(\sin\theta)\,\dot{\theta} \\
&= 2\dot{\theta}\,\ddot{\theta} + 2(\sin\theta)\,\dot{\theta} + \theta(\ddot{\theta} + \dot{\theta}) + \dot{\theta}^2 \\
&= 2\dot{\theta}\,(\ddot{\theta} + \sin\theta) + \theta(\ddot{\theta} + \dot{\theta}) + \dot{\theta}^2 \\
&= 2\dot{\theta}(-\dot{\theta}) + \theta(-\sin\theta) + \dot{\theta}^2 \\
&= -\dot{\theta}^2 - \theta\,\sin\theta \\
&= -(\dot{\theta}^2 + \theta\,\sin\theta) \leq 0 \, ,
\end{aligned}$$

where we have substituted the original equation of motion in two places. In contrast to our previous choice of Lyapunov function, we see here that the only way to get $\dot{V}_2 = 0$ is to have $\dot{\theta} = 0$ (motionless) and to have $\theta = 0$, π, etc. Thus, V_2 is a Lyapunov function for the down equilibrium of the damped pendulum.

We saw previously that the energy V_1 never increases. Consider, then, any perturbation to the solution $\theta = \dot{\theta} = 0$ with energy $E \leq 2$. Such a perturbation will kick the down equilibrium and make the pendulum swing, but not so much that the amplitude reaches π. In this region of state space, $\dot{V}_2 < 0$ unless $\theta = \dot{\theta} = 0$, and we have a true Lyapunov function and, hence, local stability of the origin.

This exercise shows that we can define different Lyapunov functions. V_1 is the total energy and thus has a physical meaning. Because $\dot{V} \leq 0$ is only negative semidefinite, we have to do some extra work and invoke Krasovskii-LaSalle to infer asymptotic stability of the origin. V_2 has no obvious physical motivation, but it satisfies $\dot{V}_2 < 0$, which allows us to immediately conclude that the origin is asymptotically stable, just using the basic Lyapunov theorem.

This exercise is partly based on material from Slotine and Li (1991).

2.17 Krasovskii's method for Lyapunov functions. Although there are no general methods for finding Lyapunov functions, there are tricks. Here is one: let $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$ be an n-dimensional nonlinear dynamical system with $\boldsymbol{A} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$ its $n \times n$ Jacobean matrix.

a. Show that if $\boldsymbol{A} + \boldsymbol{A}^{\mathsf{T}}$ is negative definite, then $V = \boldsymbol{f}^{\mathsf{T}}\boldsymbol{f}$ is a Lyapunov function.
b. Show that the origin is globally stable for $\dot{x}_1 = -2x_1 + x_2\,,\ \dot{x}_2 = x_1 - 2x_2 - x_2^3$.

Solution.

a. Consider $V = \boldsymbol{f}^{\mathsf{T}}\boldsymbol{f}$. Note that

$$\dot{\boldsymbol{f}} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\,\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{f}\,.$$

Then, differentiating V gives

$$\begin{aligned}\dot{V} &= \boldsymbol{f}^{\mathsf{T}}\dot{\boldsymbol{f}} + \dot{\boldsymbol{f}}^{\mathsf{T}}\boldsymbol{f} \\ &= \boldsymbol{f}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{f} + (\boldsymbol{A}\boldsymbol{f})^{\mathsf{T}}\boldsymbol{f} \\ &= \boldsymbol{f}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{f} + \boldsymbol{f}^{\mathsf{T}}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{f} \\ &= \boldsymbol{f}^{\mathsf{T}}\left(\boldsymbol{A} + \boldsymbol{A}^{\mathsf{T}}\right)\boldsymbol{f}\,.\end{aligned}$$

Since $\boldsymbol{F} \equiv \boldsymbol{A} + \boldsymbol{A}^{\mathsf{T}}$ is symmetric and negative definite, we can diagonalize it via $\boldsymbol{F} = \boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^{\mathsf{T}}$, where the diagonal matrix $\boldsymbol{D}$ has real eigenvalues $-\lambda_1^2$, $-\lambda_2^2$, etc. We then use the rotation matrix $\boldsymbol{R}$ to define new coordinates $\boldsymbol{h} = \boldsymbol{R}^{\mathsf{T}}\boldsymbol{f}$. Then,

$$\begin{aligned}\dot{V} &= \boldsymbol{f}^{\mathsf{T}}\boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^{\mathsf{T}}\boldsymbol{f} \\ &= \boldsymbol{h}^{\mathsf{T}}\boldsymbol{D}\boldsymbol{g} \\ &= -\lambda_1^2 g_1^2 - \lambda_2^2 g_2^2 - \ldots \leq 0\,.\end{aligned}$$

b. The dynamical system is

$$\begin{aligned}\dot{x}_1 &= -2x_1 + x_2 \\ \dot{x}_2 &= x_1 - 2x_2 - x_2^3\,.\end{aligned}$$

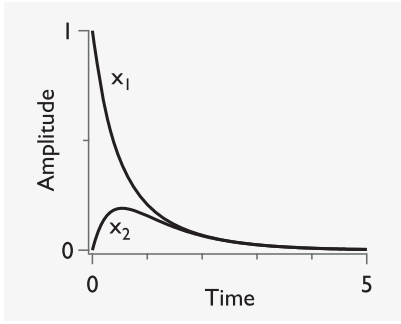
Thus,

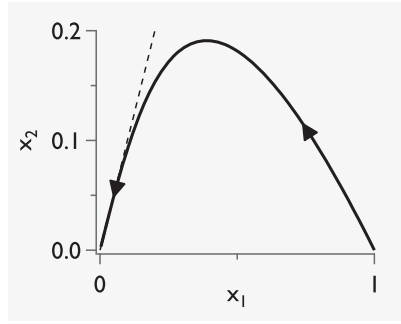
$$\boldsymbol{A} = \begin{pmatrix} -2 & 1 \\ 1 & -2 - 3x_2^2 \end{pmatrix} \quad \Longrightarrow \quad \boldsymbol{F} = \begin{pmatrix} -4 & 2 \\ 2 & -4 - 6x_2^2 \end{pmatrix}.$$

We have to show that $\boldsymbol{F}$ is negative definite (both eigenvalues have real part < 0) for all values of x_2. We could solve explicitly for the eigenvalues λ_1 and λ_2, but it is easier to make use of the fact that for a 2d matrix, the determinant $\det = \lambda_1 \lambda_2$ and the trace $\mathrm{Tr} = \lambda_1 + \lambda_2$. Thus, to be negative definite requires $\det > 0$ and $\mathrm{Tr} < 0$. Here, we have

$$\det = 16 + 24x_2^2 - 4 = 12(1 + 2x_2^2) > 0\,, \qquad \mathrm{Tr} = -8 - 6x_2^2 < 0\,.$$

Thus, the conditions are satisfied. Below are graphs and phase-plane plots for a typical initial condition, $x_1(0) = 1$, $x_2(0) = 0$, which confirm (in one instance) the stability of the origin. Dashed line in parametric plot is $x_1 = x_2$.





3 Frequency-Domain Control

Problems

3.1 **Bandwidth of integral control**. For the first-order system with integral control discussed in Section 3.3.1, derive the feedback bandwidth for arbitrary K_i. Find simpler expressions for the limits $K_\mathrm{i} \ll 1$ and $K_\mathrm{i} \gg 1$. Also, rescale the closed-loop transfer function of Eq. (3.36) to find the damping coefficient ζ as a function of K_i.

Solution.

a. *Bandwidth*. With $G(s) = \frac{1}{1+s}$ and $K(s) = \frac{K_\mathrm{i}}{s}$, we have the closed-loop transfer function is

$$T(s) = \frac{1}{1 + (KG)^{-1}} = \frac{K_\mathrm{i}}{s^2 + s + K_\mathrm{i}}\,,$$

with poles satisfying

$$s^2 + s + K_\mathrm{i} = 0\,, \qquad \Longrightarrow \qquad s = -\tfrac{1}{2} \pm \sqrt{\tfrac{1}{4} - K_\mathrm{i}}\,.$$

The closed-loop feedback bandwidth is defined as the lowest frequency for which $|T(\omega_\mathrm{c})| = \frac{1}{\sqrt{2}}\,|T(0)|$. Here, this implies

$$\frac{K_\mathrm{i}}{\sqrt{(K_\mathrm{i} - \omega^2)^2 + \omega^2}} = \frac{1}{\sqrt{2}}\,,$$

or

$$2K_\mathrm{i}^2 = K_\mathrm{i}^2 - 2K_\mathrm{i}\omega^2 + \omega^4 + \omega^2$$

so that

$$\omega^4 - 2(K_\mathrm{i} - \tfrac{1}{2})\omega^2 - K_\mathrm{i}^2 = 0\,.$$

The solution is

$$\omega_\mathrm{c}^2 = (K_\mathrm{i} - \tfrac{1}{2}) \pm \sqrt{(K_\mathrm{i} - \tfrac{1}{2})^2 + K_\mathrm{i}^2}\,,$$

where we take the positive solution, to make sure that ω is real.
In the limit $K_\mathrm{i} \ll 1$, we have

$$\omega_\mathrm{c}^2 \approx -\tfrac{1}{2} + \sqrt{\tfrac{1}{4} + K_\mathrm{i}^2} \approx \tfrac{1}{2}\left(-1 + \sqrt{1 + \tfrac{1}{4}K_\mathrm{i}^2}\right) \approx \tfrac{1}{2}\left(-1 + 1 + 2K_\mathrm{i}^2\right) \approx K_\mathrm{i}^2\,,$$

so that $\omega_\mathrm{c} \approx K_\mathrm{i}$.
In the limit $K_\mathrm{i} \gg 1$, we have

$$\omega_\mathrm{c}^2 \approx K_\mathrm{i} + \sqrt{K_\mathrm{i}^2 + K_\mathrm{i}^2}\,,$$

so that

$$\omega_\mathrm{c} \approx \left(\sqrt{1 + \sqrt{2}}\right)\sqrt{K_\mathrm{i}} \approx 1.55\sqrt{K_\mathrm{i}}\,.$$

b. *Damping*. Starting from

$$\begin{aligned} s^2 + s + K_\mathrm{i} &= 0 \\ \frac{s^2}{K_\mathrm{i}} + \frac{s}{K_\mathrm{i}} + 1 &= 0 \\ (s')^2 + \frac{1}{\sqrt{K_\mathrm{i}}}\,s' + 1 &= 0\,, \end{aligned}$$

where $s' = s \,/\, \sqrt{K_\mathrm{i}}$. Thus,

$$\zeta = \frac{1}{2\sqrt{K_\mathrm{i}}}\,.$$

Thus, for small gains, the system behaves as an overdamped oscillator. Critical damping occurs at $K_\mathrm{i} = \frac{1}{4}$, and it becomes more and more underdamped as K_i is further increased.

3.2 **Rejecting disturbances in undamped oscillators**. Let $G(s) = \frac{1}{1+s^2}$ and consider how an input disturbance d modifies the block diagram in Figure 3.1.

a. *Response functions*. Show that the disturbance response of the output $y(s)$ is given by $y(s) = \frac{G}{1+KG}\,d(s)$. Show that the controller output is $u(s) = \frac{-KG}{1+KG}\,d(s)$.
b. *Proportional (P) control*. Find and plot the $y(t)$ and $u(t)$ disturbance impulse responses for the P-control algorithm, $K(s) = K_\mathrm{p}$. What is wrong with P-control?
c. *Derivative (D) control*. Find and plot the $y(t)$ and $u(t)$ disturbance impulse responses for D control, $K(s) = K_\mathrm{d}s$. Find the critical value of K_d^* where the response changes from oscillatory to damped. Using the final-value theorem, show that the initial control input is $u(t \to 0^+) = -K_\mathrm{d}$. What is wrong with D control?
d. *Proportional-derivative (PD) control*. Consider the PD algorithm, $K = K_\mathrm{p} + K_\mathrm{d}s$. With K_p as a free parameter, find K_d^* for critical damping. Find and plot the y and u disturbance impulse responses for this choice of K_d. Find $u(0^+)$, and discuss the penalty for PD control relative to D control.

e. *Proportional-integral-derivative (PID) control.* For a step disturbance, $d(t) = \theta(t)$, show that PD control (d) results in a steady-state offset. Add integral control to make a full PID controller, $K = K_\mathrm{p} + K_\mathrm{i}/s + K_\mathrm{d}s$. By matching coefficients of $\{1, s, s^2, s^3\}$ of the denominator polynomial (pole placement), show that choosing $K_\mathrm{p} = 3a^2 - 1$, $K_\mathrm{i} = a^3$, and $K_\mathrm{d} = 3a$ leads to closed-loop dynamics with three degenerate poles at $s = -a$. Plot step and impulse responses for y and u, for $a = 2$.

Solution.

a. *Response functions.* From the block diagram, an input disturbance d implies that $y = G(u + d)$. For a zero reference, $u = Ke = K(r - y) = -Ky$. Thus,

$$y = G(u + d) = G(-Ky + d) \qquad \Longrightarrow \qquad y = \frac{G}{1 + KG}\, d = G\, S\, d\,,$$

and, thus,

$$u = -Ky = -\frac{KG}{1 + KG}\, d = -T\, d\,.$$

Here, the sensitivity function $S \equiv \frac{1}{1+KG}$ and its complement $T \equiv 1 - S = \frac{KG}{1+KG}$.

b. *Proportional control.* Using $G(s)^{-1} = 1 + s^2$, the load sensitivity function is

$$SG = \frac{G}{1 + KG} = \frac{1}{K + G^{-1}} = \frac{1}{(K_\mathrm{p} + 1) + s^2}\,,$$

which describes an undamped oscillator with natural frequency $\sqrt{K_\mathrm{p} + 1}$. To obtain explicit time-domain solutions, we should be careful and keep initial conditions in the Laplace transforms. Here, it is easiest just to work directly in the time domain. To handle the delta function disturbance $d(t) = \delta\,(t)$, we integrate the equation of motion from $t = 0^-$ to $t = 0^+$. After an input disturbance impulse, $y(t)$ is continuous but $\dot{y}(t)$ can have a jump discontinuity. Thus, we write,

$$\int_{0^-}^{0^+} \mathrm{d}t\, (\ddot{y} + y) \approx \dot{y}(0^+) - \dot{y}(0^-) = \int_{0^-}^{0^+} \mathrm{d}t\ \delta\,(t) = 1\,.$$

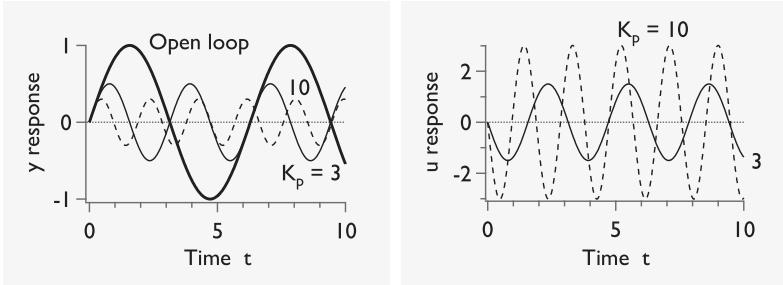
Since the initial conditions are $y = \dot{y} = 0$ for $t = 0$, we conclude that the impulse response is the same as free response for initial conditions $y(0^+) = 0$ and $\dot{y}(0^+) = 1$. The solution satisfying these initial conditions is, for $t > 0$,

$$y(t) = \frac{1}{\sqrt{K_\mathrm{p} + 1}}\, \sin\left(\sqrt{K_\mathrm{p} + 1}\, t\right),$$

$$u(t) = -K_\mathrm{p} y(t) = -\left(\frac{K_\mathrm{p}}{\sqrt{K_\mathrm{p} + 1}}\right) \sin\left(\sqrt{K_\mathrm{p} + 1}\, t\right).$$

At low gains $K_\mathrm{p} \ll 1$, and $SG \approx G$, the open-loop response: proportional feedback has little effect, and the response oscillates with $\omega \approx 1$ and amplitude ≈ 1.

At high gains, $K_p \gg 1$, the amplitude is reduced ($\approx 1/\sqrt{K_p}$), and the frequency increases to $\omega \approx \sqrt{K_p}$. The control signal is also sinusoidal, with amplitude $u \sim \sqrt{K_p}$.



Although reducing the amplitude of the response and speeding up its dynamics can be good features, this is nonetheless not a satisfactory response: because the dynamics are not damped, the system never "forgets" the past. Thus, the past controls the present, whereas the goal of a controller for disturbances is typically to make the behavior of a system independent of its long-time history. In addition, the control effort, defined as $u^2(t)$, never dies away. Thus, each disturbance requires the controller to act on the system forever after!

The intrinsic damping of a system modifies these conclusions only slightly. The past does decay, on a dimensionless time scale ζ^{-1} for the generic scaled oscillator, $G(s) = \frac{1}{1+2\zeta s+s^2}$. But in this case, proportional control does not alter the decay rate. If it is too slow for your purposes, you are out of luck.

c. *Derivative control.* Pure derivative control, $K(s) = K_d s$ leads to

$$S = \frac{1}{1+KG} = \frac{1}{1+K_d s+s^2}\,.$$

We see that derivative control adds damping but does not alter the natural frequency. The gain $K_d = 2$ corresponds to critical damping, which is a reasonable goal. Lower gain implies underdamped behavior; higher gain overdamped.

To calculate the maximum control signal required, we use the results in Part (a):

$$-T = -\frac{KG}{1+KG} = -\frac{1}{1+(KG)^{-1}} = -\frac{1}{1+\frac{1+s^2}{K_d s}} = \frac{-K_d s}{1+K_d s+s^2}\,.$$

The initial value theorem then says that

$$u(0^+) = \lim_{s\to\infty} s[-T(s)] = -K_d\,.$$

Note that a critically damped oscillator has the fastest decay envelope for fixed oscillator frequency, we conclude that we can use derivative control to damp on a time scale of the oscillator period, but not faster.

To summarize:

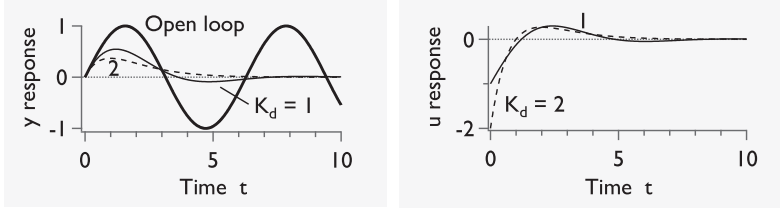
- *P control* decreases the amplitude of disturbances and makes the response faster but does not add damping.
- *D control* adds damping but does not alter the time scales or amplitudes.

We use Mathematica to solve for explicit time responses. As before, the impulse response is equivalent to free response with $y(0) = 0$ and $\dot{y} = 1$, for $t > 0$. Using $u = K_d\dot{y}$ and for $K_d = 1$ and $\omega = \sqrt{3}/2$, they are

$$K_d = 1: \qquad y(t) = \frac{2}{\sqrt{3}}\, e^{-t/2} \sin \omega t\,, \qquad u(t) = -\,e^{-t/2}\left(\cos \omega t - \frac{1}{\sqrt{3}} \sin \omega t\right)$$

$$K_d = 2: \qquad y(t) = t\, e^{-t}\,, \qquad u(t) = -2(1-t)\, e^{-t}\,.$$

The output y and input u responses to an impulse are shown below.



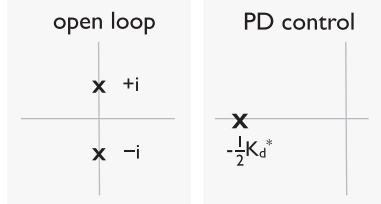
d. *Proportional-derivative control*. Here, $K(s) = K_p + K_d s$. Then,

$$\frac{y}{d} = \frac{G}{1+KG} = \frac{1}{K + G^{-1}} = \frac{1}{K_p + K_d s + (1+s^2)} = \frac{1}{(1+K_p) + K_d s + s^2}\,,$$

which corresponds to an oscillator with damping coefficient $K_d/2$ and frequency $\sqrt{1+K_p}$. For critical damping, $K_d^* = 2\sqrt{1+K_p}$, and

$$\frac{y}{d} = \frac{1}{\left(\sqrt{1+K_p} + s\right)^2} = \frac{1}{\left(\frac{1}{2}K_d^* + s\right)^2}\,.$$

With this choice of K_d^* in terms of K_p, we transform the undamped poles of the system at $p = \pm i$ to two stable, degenerate real poles at $p = -\frac{1}{2}K_d^*$, whose position can be chosen, as shown below.



For the control signal,

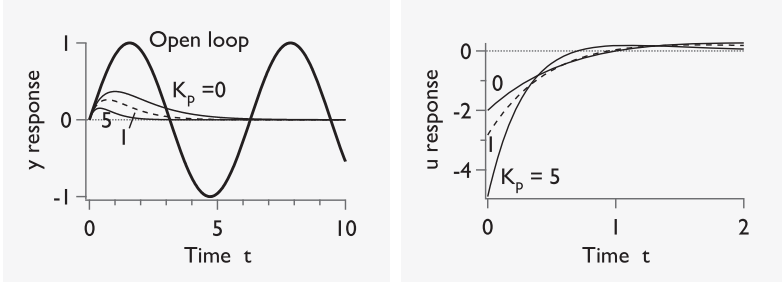
$$u = -\frac{1}{1+(KG)^{-1}} = -\frac{1}{1+\frac{1+s^2}{K_\mathrm{p}+K_\mathrm{d}s}} = -\frac{K_\mathrm{p}+K_\mathrm{d}s}{(1+K_\mathrm{p})+K_\mathrm{d}s+s^2}$$
$$= -\frac{K_\mathrm{p}+2\sqrt{1+K_\mathrm{p}}s}{\left(\sqrt{1+K_\mathrm{p}}+s\right)^2}\,.$$

where the last step chooses critical damping ($K_\mathrm{d} = 2\sqrt{K_\mathrm{p}^2+1}$). Just as with pure derivative control, the initial controller amplitude is $u(0^+) = -K_\mathrm{d}$. But if we choose $K_\mathrm{d} = 2\sqrt{K_\mathrm{p}^2+1}$ in order to maintain critical damping, we will use a larger derivative gain in the PD case. Thus, the improved control again comes at a cost of larger amplitude.

For critical damping, the explicit time-domain solutions are

$$y(t) = t\,\mathrm{e}^{-\sqrt{K_\mathrm{p}+1}t}\,, \qquad u(t) = -\left(2\sqrt{K_\mathrm{p}+1} - (K_\mathrm{p}+2)t\right)\mathrm{e}^{-\sqrt{K_\mathrm{p}+1}t}\,.$$

You might wonder about the total time-integrated control effort, defined, for example, as $\mathcal{E} \equiv \frac{1}{2}\int_0^\infty \mathrm{d}t\,u(t)^2$. It, too, does not greatly exceed that of D control, showing that the main limitation is the largest limiting value of u, here occurring at $t = 0^+$.
The time responses are given below.



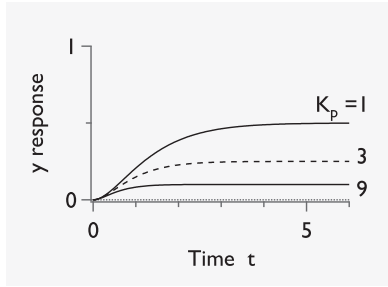
e. *Proportional-integral-derivative (PID) control.* There are several ways to see that a step disturbance, $d(t) = \theta(t)$, will lead to an offset for a PD controller with $K_\mathrm{d} = 2\sqrt{1+K_\mathrm{p}}$. We can use the transfer function and final-value theorem:

$$y(t=\infty) = \lim_{s\to 0}(s)(SG)(s^{-1}) = S(0)\,G(0) = \left.\frac{1}{(1+K_\mathrm{p})+K_\mathrm{d}s+s^2}\right|_{s\to 0} = \frac{1}{1+K_\mathrm{p}}\,.$$

Alternatively, using Mathematica, the explicit time-dependent step response is

$$y(t) = \frac{1}{1+K_\mathrm{p}}\left[1-\left(1+\sqrt{1+K_\mathrm{p}}\right)\mathrm{e}^{-\sqrt{1+K_\mathrm{p}}t}\right] \xrightarrow[t\to\infty]{} \frac{1}{1+K_\mathrm{p}}\,,$$

which is plotted below for $K_\mathrm{p} = 1$. The step disturbance results in a constant offset, whose value decreases as the proportional gain K_p is increased.



In order to get rid of the long-time offset to a step disturbance, we consider the full PID algorithm: For $K = K_\mathrm{p} + K_\mathrm{i}/s + K_\mathrm{d}s$, we match the coefficients of 1, s, and s^2 in the denominator of the desired response polynomial:

$$\begin{aligned}(s+a)^3 &= s^3 + 3as^2 + 3a^2 s + a^3 \\ &= s^3 + K_\mathrm{d}s^2 + (K_\mathrm{p}+1)s + K_\mathrm{i}\,.\end{aligned}$$

This is equivalent to *pole placement* (Section 3.7.2). By inspection, we choose

$$K_\mathrm{p} = 3a^2 - 1\,, \qquad K_\mathrm{i} = a^3\,, \qquad K_\mathrm{d} = 3a\,.$$

Having made this choice, the transfer function from d to y is (using Mathematica)

$$SG = \frac{G}{1+KG} = \frac{s}{(s+a)^3}\,,$$

which implies impulse and step responses

$$\begin{aligned}y_\mathrm{impulse}(t) &= t\left(1 - \frac{a}{2}t\right)\mathrm{e}^{-at}\;\theta(t) \\ y_\mathrm{step}(t) &= \frac{1}{2}t^2\,\mathrm{e}^{-at}\;\theta(t)\,.\end{aligned}$$

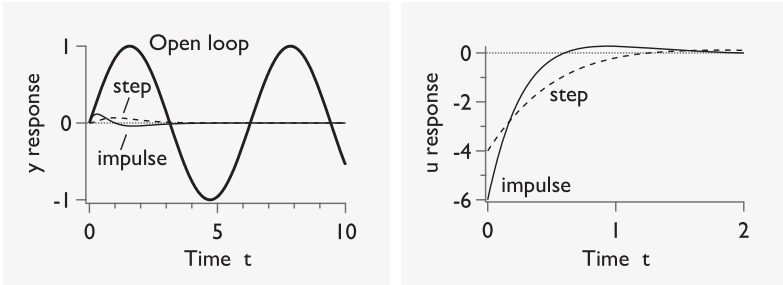
Similarly, the response from d to u is

$$-T = -\frac{KG}{1+KG} = -\frac{3as^2 + (3a^2-1)s + a^3}{(s+a)^3}\,,$$

which implies impulse and step responses

$$\begin{aligned}u_\mathrm{impulse}(t) &= -\frac{1}{2}\left((a^3+a)t^2 - 2(1+3a^2)t + 6a\right)\mathrm{e}^{-at}\;\theta(t) \\ u_\mathrm{step}(t) &= -\frac{1}{2}\left(-(a^2+1)t^2 + 4a\right)\mathrm{e}^{-at} - 1\;\theta(t)\,.\end{aligned}$$

Plots for $a = 2$ are shown below. By using the full flexibility of PID control, we have arrived at a controller that can compensate for step and impulse disturbances at the input to the system.



In further problems, we will consider the command response using feedforward and also strategies that limit $u(t)$ to some maximum value. We see that the control signal $u(t)$ typically "spikes" negative in response to a disturbance. Later, we will see that a smaller response, applied for a longer time, is another option. Finally, we will also consider what happens when the frequency of the physical oscillator differs from what we think it is.

3.3 Feedforward control for an undamped oscillator. In Problem 3.2, we discussed how a PID controller can damp impulse and step disturbances. Now let's make an undamped oscillator reject disturbances *and* track commands. Recall the system, $G(s) = \frac{1}{1+s^2}$, and the PID controller, $K(s) = K_\mathrm{p} + K_\mathrm{i}/s + K_\mathrm{d}s$, with $K_\mathrm{p} = 3a^2 - 1$, $K_\mathrm{i} = a^3$, and $K_\mathrm{d} = 3a$, chosen to make the closed-loop denominator $(1 + KG) \sim (1 + s/a)^3$.

a. Find $r \to \{y, u\}$ transfer functions and analytic time-domain formulas for the command step response, for $K_\mathrm{p} = 3a^2 - 1$, $K_\mathrm{i} = a^3$, and $K_\mathrm{d} = 3a$. Plot for $a = 2$.

b. Since the above PID controller worked well for disturbances, we would like to keep it and add feedforward instead. Modify the command signal $r(s)$ by an element $F(s)$. The simple strategy $F \sim T^{-1}$, for $T = \frac{KG}{1+KG}$, does not lead to proper response. Try $F(s) = \frac{T^{-1}(s)}{(1+s/a)^2}$, which is proper. Find $r \to \{y, u\}$ transfer functions and analytic formulas for the step response. Plot for $a = 2$. Compare the output and control time responses with and without feedforward.

Solution.

a. The PID controller from Problem 3.2 is

$$K(s) = (3a^2 - 1) + \frac{a^3}{s} + 3as = 11 + \frac{8}{s} + 6s\,,$$

with $a = 2$. It leads to a $r \to y$ transfer function $T(s)$ given by

$$T(s) = \frac{KG}{1 + KG} = \frac{8 + 11s + 6s^2}{(2 + s)^3}\,,$$

and an $r \to u$ transfer function

$$S\,K(s) = \frac{K}{1 + KG} = \frac{(1 + s^2)(8 + 11s + 6s^2)}{(2 + s)^3}\,.$$

From Mathematica calculations, the analytic solutions for the step responses, $r(t) = \theta(t)$, for $t > 0$,

$$y(t) = 1 + \tfrac{1}{2}\left[t\,(8 - 5t) - 2\right] \mathrm{e}^{-2t}$$
$$u(t) = 6\delta\,(t) + 1 - \left(1 + 2t(t+1)\ \mathrm{e}^{-2t}\right).$$

The delta function at $t = 0$ comes from the action of derivative feedback, $u \sim K_\mathrm{d}\dot{r}$ on the command signal and would be smoothed in a more realistic analysis. See below for plots of the response curves.

b. The $r \to y$ transfer function $T(s)\,F(s)$ for $a = 2$ is given by

$$T(s)\,F(s) = \frac{KGF}{1 + KG} = \frac{4}{(2+s)^2},$$

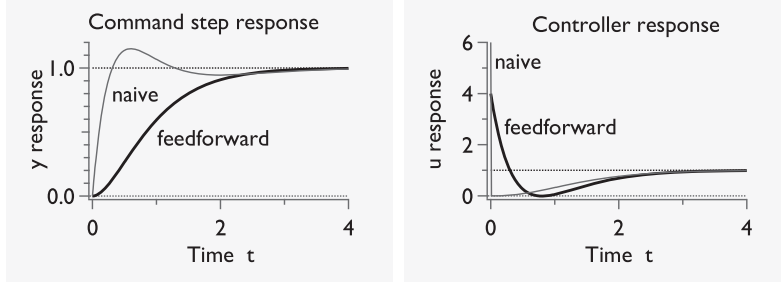
and the $r \to u$ transfer function

$$SKF(s) = \frac{KF}{1 + KG} = \frac{4(1+s^2)}{(2+s)^2}.$$

From Mathematica calculations, the analytic solutions for the step responses, $r(t) = \theta(t)$, for $t > 0$,

$$y(t) = 1 - (2t+1)\ \mathrm{e}^{-2t}$$
$$u(t) = 1 + (3 - 10t)\ \mathrm{e}^{-2t}\ .$$

The $y(t)$ and $u(t)$ response curves for $r(t) = \theta(t)$ are illustrated below for the naive (no feedforward) and feedforward cases. Notice how the feedforward has eliminated the overshoot in the control signal and reduced the control requirements, too. In particular, the delta function at $t = 0$ in (a) has been softened to a simple jump discontinuity at $t = 0$. You might think that it should to be possible to design a control where $u(t)$ does not exceed 1, and, in Section 9.1.1, we shall see how to do this. See also the direct digital design techniques discussed in Section 5.4.2.



3.4 Stabilizing an unstable oscillator. Consider an unstable oscillator $G(s) = \frac{1}{s^2-1}$, subject to an impulse input disturbance, $d(t) = \delta\,(t)$.

a. *Proportional (P) control.* For $K(s) = K_\mathrm{p}$, show that $K_\mathrm{p} > 1$ stabilizes the closed-loop disturbance response $\frac{G}{1+KG}$. Describe the response and its problems. What is the minimum proportional gain if the oscillator has natural frequency ω?

b. *Proportional-Derivative (PD) control.* For $K(s) = K_\mathrm{p} + K_\mathrm{d} s$, use pole placement (matching denominator polynomials) to show that choosing $K_\mathrm{d} = K_\mathrm{d}^* \equiv 2\sqrt{K_\mathrm{p} - 1}$ and $K_\mathrm{p} > 1$ gives the desirable critical-damping response. Choose a to give two poles at $s = -a$. For $a = 1$, plot the response for y and u.
c. *Filtered Proportional-Derivative (PD) control.* Limit the derivative control by filtering: For $K(s) = K_\mathrm{p} + \frac{K_\mathrm{d}^* s}{1 + s/\omega_\mathrm{d}}$, find the $d \to \{y, u\}$ transfer functions and also the explicit time-domain differential equations for $y(t)$ and $u(t)$.
d. Matching denominators, choose K_p, K_d, and ω_d to get a disturbance response with three poles at $s = -a$. For $a = 1$, plot the disturbance response $y(t)$ and $u(t)$.

Solution.

a. *Proportional (P) control.* For a PD controller with $K(s) = K_\mathrm{p}$, the closed-loop response to an input disturbance is

$$SG = \frac{G}{1 + KG} = \frac{1}{K + G^{-1}} = \frac{1}{K_\mathrm{p} + s^2 - 1} = \frac{1}{s^2 + (K_\mathrm{p} - 1)} .$$

For $K_\mathrm{p} > 1$, this gives undamped oscillation at frequency $\sqrt{K_\mathrm{p} - 1}$. It prevents instability but does not bring the system back to its equilibrium. (If there is system damping, it will bring it back at the normal damping rate, which can be slow.)
For $G = \frac{1}{s^2 - \omega^2}$, the same arguments leads to $K_\mathrm{p} > \omega^2$. Let's quickly show this in the time domain:

$$\ddot{y} + \omega^2 y(t) = u(t) + d(t) ,$$

where $y(t)$ is the oscillator state, $u(t)$ the control input, and $d(t)$ the disturbance. Choosing $u(t) = -K_\mathrm{p} y(t)$ gives

$$\ddot{y} + (K_\mathrm{p} - \omega^2) y(t) = d(t) ,$$

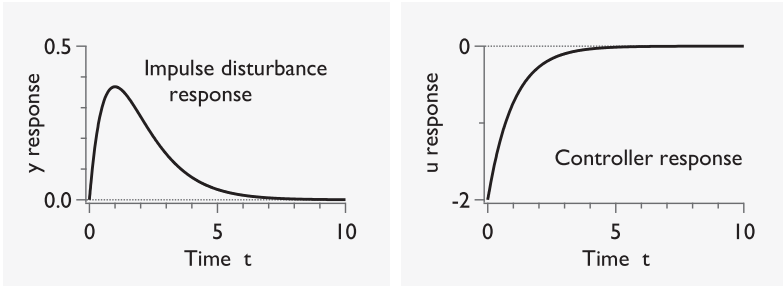
which has oscillatory response for $K_\mathrm{p} > \omega^2$.
b. *Proportional-Derivative (PD) control.* To have a critically damped response, the denominator $D(s) = s^2 + K_\mathrm{d} s + K_\mathrm{p} - 1$ should have degenerate roots. Matching coefficients to the desired denominator $(s + a)^2$, we have

$$\begin{aligned} & s^2 + K_\mathrm{d} s + (K_\mathrm{p} - 1) \\ = \; & s^2 + 2as + a^2 , \end{aligned}$$

so that,

$$a^2 = K_\mathrm{p} - 1 \text{ and } K_\mathrm{d} = 2a = 2\sqrt{K_\mathrm{p} - 1} .$$

Note that to achieve stabilization, we need to take the positive root for a, implying $K_\mathrm{d} > 0$. For $a = -1$, we have $K_\mathrm{p} = K_\mathrm{d} = 2$, which gives the response shown below.



c. *Proportional-Derivative (PD) control.* The controller is

$$K(s) = K_\mathrm{p} + \frac{K_\mathrm{d} s}{1 + s/\omega_\mathrm{d}} .$$

The $d \to y$ transfer function is

$$\frac{G}{1 + KG} = \frac{s + \omega_\mathrm{d}}{s^3 + \omega_\mathrm{d} s^2 + (K_\mathrm{p} - 1 + K_\mathrm{d}\omega_\mathrm{d})s + (K_\mathrm{p} - 1)\omega_\mathrm{d}} .$$

The $d \to u$ transfer function is

$$\frac{-KG}{1 + KG} = -\frac{K_\mathrm{p}\omega_\mathrm{d} + (K_\mathrm{p} + K_\mathrm{d}\omega_\mathrm{d})s}{s^3 + \omega_\mathrm{d} s^2 + (K_\mathrm{p} - 1 + K_\mathrm{d}\omega_\mathrm{d})s + (K_\mathrm{p} - 1)\omega_\mathrm{d}} .$$

These are probably more easily understood in the time domain. Multiplying through by $1 + s/\omega_\mathrm{d}$ and going back to the time domain gives (with all quantities functions of t), two coupled equations for $y(t)$ and $u(t)$:

$$\ddot{y} - y = u + d , \qquad \left(\frac{1}{\omega_\mathrm{d}}\right)\dot{u} + u = -K_\mathrm{p} y - \left(\frac{K_\mathrm{p}}{\omega_\mathrm{d}} + K_\mathrm{d}\right)\dot{y} ,$$
$$y(0) = 0 , \quad \dot{y}(0) = 0 , \qquad u(0) = 0 .$$

Notice that taking $\omega_\mathrm{d} \to \infty$ gives us back the simpler equations for straight PD control.

d. Matching

$$\begin{aligned}(s + a)^3 &= s^3 + 3as^2 + 3a^2 s + a^3 \\ &= s^3 + \omega_\mathrm{d} s^2 + (K_\mathrm{p} - 1 + K_\mathrm{d}\omega_\mathrm{d})s + (K_\mathrm{p} - 1)\omega_\mathrm{d}\end{aligned}$$

gives

$$K_\mathrm{p} = 1 + \frac{a^2}{3} , \quad K_\mathrm{d} = \frac{8}{9}a , \quad \omega_\mathrm{d} = 3a .$$

For numerics, we take $a = 1$, or $K_\mathrm{p} = 4/3$, $K_\mathrm{d} = 8/9$, and $\omega_\mathrm{d} = 3$. Notice that the derivative term is cut off a factor of 3/(8/9) = 3.375 times higher than the cutoff frequency.

For the above choice of controller coefficients, the transfer functions become as follows: The $d \to y$ transfer function is

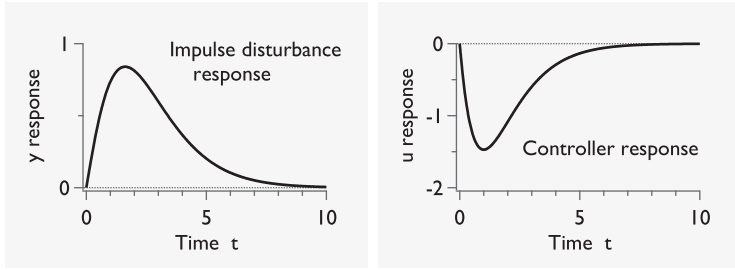
$$d \to y : \qquad \frac{G}{1 + KG} = \frac{3 + s}{(1 + s)^3}$$

$$d \to u : \qquad -\frac{KG}{1+KG} = -\frac{4}{(1+s)^2} .$$

Then, solving these (or the time-domain equations) gives, for $t > 0$,

$$y(t) = t(t+1)\,\mathrm{e}^{-t} , \qquad u(t) = -4t\,\mathrm{e}^{-t} .$$

These are plotted below. Comparing with the PD controller above, we see that the disturbance response is worse, increasing from $1/e \approx 0.37$ to ≈ 0.84. However, the maximum controller response has improved, from -2 to $-4/e \approx -1.47$. In addition, the noise sensitivity – the degradation that occurs when we allow for measurement noise – will be better in the filtered design. We will investigate such issues in Chapter 8. Notice that filtering the derivative term – adding ω_d – makes $u(t)$ change at a finite rate. By contrast, the controller signal had a jump discontinuity for pure PD control.



The above solution describes some of the issues involved in the linear control of an unstable pendulum. Problem 11.17 discusses how to stabilize the full nonlinear dynamics, assuming that the applied torque is strong enough to be able to overcome directly the gravitational torque.

3.5 **Integral-control instability**. Using Eq. (3.47), we studied an instability in integral control for $G = \frac{1}{(1+s)^2}$ and $K = \frac{K_\mathrm{i}}{s}$. Please verify that

a. At the onset of instability, $K_\mathrm{i} = 2$, the stable root is at -2.
b. For $K_\mathrm{i} = 0$, there is a double pole at -1 and a single pole at 0. Show that one of the -1 poles collides with the 0 pole at $K_\mathrm{i} = \frac{4}{27}$. Find the location of that collision (and that of the other pole while you're at it).
c. Verify that the gain margin is $2/K_\mathrm{i}$.
d. Plot the phase margin as a function of K_i over the parameter range $0 < K_\mathrm{i} < 2$. Evaluate the phase margin for $K_\mathrm{i} = 1$ numerically.

Solution.

a. *Stable root at onset*: In the text, we showed that the onset of instability occurs at $K_\mathrm{i}^* = 2$ and that it is a Hopf bifurcation, with $\omega^* = 1$. For $G = \frac{1}{(1+s)^2}$ and $K = \frac{K_\mathrm{i}}{s}$, the denominator of $T(s) = \frac{1}{1+(KG)^{-1}} = \frac{K_\mathrm{i}}{K_\mathrm{i}+(1+s)^2 s}$. Substituting $K_\mathrm{i} = 2$, we have

$$s(s+1)^2 + 2 = s^3 + 2s^2 + s + 2 ,$$

which factors into $(s + \mathrm{i})(s - \mathrm{i})(s + p)$, where p is the value of the 3rd root. To verify that the third root is at -2, we check that

$$(s^2 + 1)(s + 2) = s^3 + 2s^2 + s + 2\,.$$

b. *Double real roots.* We want to find the condition for a double real root. Thus, we want

$$s(s + 1)^2 + K_\mathrm{i} = (s - p)^2\,(s - q)$$

where p and q should be negative to give stable roots. Expanding out, we have

$$\cancel{s^3} + 2s^2 + s + K_\mathrm{i} = \cancel{s^3} - (2p + q)s^2 + (p^2 + 2pq)s - p^2 q$$

Matching powers of s (constant, linear, quadratic), gives two solutions. One is the trivial one mentioned in the problem ($K_\mathrm{i} = 0$, $p = -1$, $q = 0$). The other is the desired roots: $K_\mathrm{i} = \frac{4}{27}$, $p = -\frac{1}{3}$, and $q = -\frac{4}{3}$.
Alternatively, you can simply factor

$$s(s + 1)^2 + \tfrac{4}{27} = \left(s + \tfrac{1}{3}\right)^2 \left(s + \tfrac{4}{3}\right)\,.$$

c. *Gain margin.* The gain margin is the inverse of $|L(\mathrm{i}\omega)|$, evaluated at the frequency ω^* such that the phase of $L(\mathrm{i}\omega^*) = -180°$. In this case $\omega^* = 1$ and

$$|L(i)| = \left|\frac{1}{(1 + \mathrm{i})^2}\right| \left|\frac{K_\mathrm{i}}{\mathrm{i}}\right| = \frac{K_\mathrm{i}}{2}\,,$$

implying a gain margin of $2/K_\mathrm{i}$.

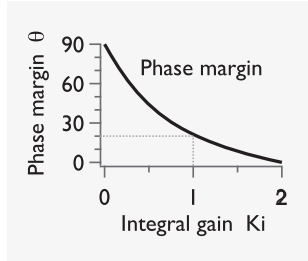
d. *Phase margin.* The phase margin is $\pi + \varphi$, where φ is the phase of $L(\mathrm{i}\omega)$, evaluated at a gain where $|L| = 1$. Here, the magnitude condition leads to

$$\frac{1}{\sqrt{(1 - \omega^2)^2 + 4\omega^2}} \left(\frac{K_\mathrm{i}}{\omega}\right) = 1\,,$$

which is a cubic equation that one can solve for $\omega = \omega_0$. Then, the phase angle φ is given by

$$\tan\varphi = -\frac{1 - \omega_0^2}{2\omega_0}\,,$$

which gives 21.4° for $K_\mathrm{i} = 1$. See plot below, too.

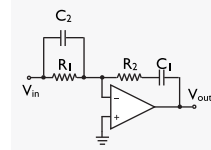


3.6 Analog circuits for PID control. Although less flexible than digital controllers, analog circuits can nonetheless be the easiest, cheapest solution. Simple algorithms such as PID control can be implemented with a single operational amplifier costing just a few cents, with bandwidths easily reaching 100 kHz.

a. Show how the circuit in Section 3.2.2 could be used as a proportional controller. If the controller output $u(t)$ is related to the error $e(t) = r(t) - y(t)$ by $u(t) = K_p e(t)$, what is K_p? It is important to get the signs correct. How is $V_{in}(t)$ related to $e(t)$?

b. The op-amp circuit at right implements a PID controller of the form $u(t) = [K_p + \frac{K_i}{s} + K_d s]\, e(t)$. Find K_p, K_i, and K_d.

c. Draw circuits for PI, PD, and I control.



Solution.

a. *Analog P control*. $K_p = \frac{R_2}{R_1}$. $e = -V_{in}$.

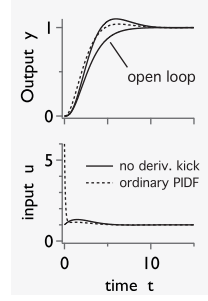
b. *Analog PID control*. $K_p = \frac{R_2}{R_1} + \frac{C_2}{C_1}$, $K_i = R_1 C_1$, $K_d = R_2 C_2$.

c. Circuit diagrams for PI, PD, and I control.

3.7 Eliminating derivative kick. If the basic PID algorithm is tuned to regulate against disturbances, it will tend to perform poorly when tracking a step command. A simple improvement is to apply the derivative term not to the error $e = (r - y)$ but to $-y$ directly. (The error still enters other terms, as usual.) This modification eliminates the output "kick" when changing the reference signal.

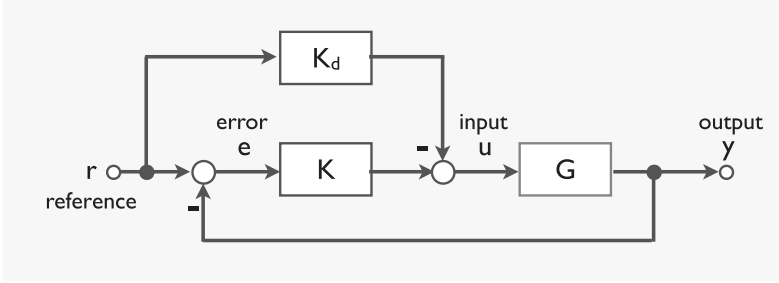
a. Derive this algorithm in the Laplace domain, and explain in simple terms why it works. Show that the corresponding block diagram has a new feedforward term.

b. The system $G(s) = \frac{1}{(1+s)^3}$ has a "sluggish" response. Explore PIDF control (Eq. 3.41), with and without derivative kick. For $K_p = 1$, $K_i = 0.5$, $K_d = 0.5$, and $\omega_d = 10$, reproduce the plots at right. The dashed line (ordinary PIDF) shows a large spike in the signal $u(t)$ sent to the system. The solid line shows that applying the DF part of the algorithm to y eliminates the spike, with only minor deterioration of the step response.



Solution.

a. The block diagram is as follows:



The controller is

$$K(s) = K_\mathrm{p} + \frac{K_\mathrm{i}}{s} + \frac{K_\mathrm{d} s}{1 + s/\omega_\mathrm{d}} \equiv K_\mathrm{pi}(s) + K_\mathrm{d}(s)\,.$$

In the block diagram, we subtract off the $K_\mathrm{d}(s)$ contribution so that the net direct path signal is $K - K_\mathrm{d} = K_\mathrm{pi}(s)$. Thus, the block diagram is equivalent to the transfer function

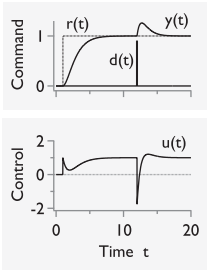
$$\frac{y}{r} = \frac{K_\mathrm{pi} G}{1 + KG}\,, \qquad \frac{u}{r} = \frac{K_\mathrm{pi}}{1 + KG}\,.$$

Notice that the disturbance response remains, for a disturbance d that enters at the input to G,

$$\frac{y}{d} = \frac{G}{1 + KG}\,.$$

b. See book website for Mathematica file.

3.8 **Decoupling feedforward from feedback**. Naive implementations of feedforward (e.g., Problem 3.3) lead to schemes where the feedforward filter F depends on the controller K. Figure 3.5 presented a scheme for combining feedforward and feedback that decouples the two transfer functions F and K.

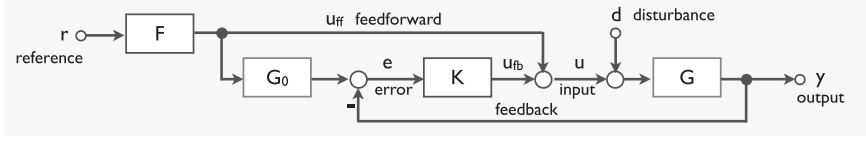


a. Solve the block diagram in Figure 3.5 for y. Show that when the model is perfect ($G_0 = G$) that $y = FG\,r + (\frac{G}{1+KG})\,d$. Thus, F acts only on r, and K only on d.

b. Plot command step and disturbance impulse response for the undamped oscillator $G = \frac{1}{s^2+1}$ (see left). Use a PD controller with $K_\mathrm{p} = 1$ and $K_\mathrm{d} = 2\sqrt{2}$ and a feedforward controller $F = G^{-1}/(1+s)^2$.

Solution.

a. For convenience, the Figure 3.5 block diagram is reproduced here:



From the block diagram,

$$y = G(u + d) \qquad u_\mathrm{ff} = Fr$$
$$u = u_\mathrm{ff} + u_\mathrm{fb} \qquad u_\mathrm{fb} = Ke$$
$$= Fr + K(FG_0 r - y) \qquad e = FG_0 r - y\,.$$

Then

$$\begin{aligned} y &= G[Fr + K(FG_0 r - y) + d] \\ &= FGr + FGKG_0 r - KGy + Gd \\ &= FG\left(\frac{1 + KG_0}{1 + KG}\right) r + \left(\frac{G}{1 + KG}\right) d\,. \end{aligned}$$

If the model is perfect, $G_0 = G$ and

$$y = FG\,r + \left(\frac{G}{1+KG}\right) d\,.$$

What if the model is not perfect? At frequencies where KG and $KG_0 \gg 1$,

$$\begin{aligned} y &\approx FG\left(\frac{KG_0}{KG}\right) r + \left(\frac{G}{1+KG}\right) d \\ &= FG_0\, r \quad + \quad \left(\frac{G}{1+KG}\right) d\,. \end{aligned}$$

Thus, even when the model is imperfect, F and K decouple at frequencies for which the feedback controller gain is large. You design F using the model G_0 and use high feedback gains to compensate for model inaccuracies.
Finally, it is useful to have a explicit expression for $u(s)$ as a function of the command r and disturbance d inputs. For a perfect model ($G_0 = G$),

$$u = Fr + K(FGr - y) = Fr - \left(\frac{KG}{1+KG}\right) d\,.$$

b. Substitute $K(s) = 1 + 2\sqrt{2}s$ and $F(s) = \frac{1+s^2}{(1+s)^2}$ into

$$\begin{aligned} y &= FG\,r + \left(\frac{G}{1+KG}\right) d \\ u &= Fr + K(FGr - y) = Fr - \left(\frac{KG}{1+KG}\right) d\,, \end{aligned}$$

for r a step at $t = 1$ and d an impulse at $t = 12$ and plot the result.

3.9 **Proportional temperature control of an extended rod**. In Section 2.3.2, we showed that the transfer function between a heater at $x = 0$ and a temperature probe at $x = \ell$ was $G(s) = \frac{1}{\sqrt{s}}\,\mathrm{e}^{-\sqrt{s}}$, with lengths scaled by ℓ and times by $\frac{\ell^2}{D}$. Here, D is the thermal diffusion constant. Find the maximum proportional gain K_p^* that can be applied without oscillation. Derive the frequency of oscillations at instability onset, ω^*.

Solution.

a. *Onset oscillation frequency* ω. To calculate the frequency of oscillations at onset, we look at the phase of $G(s)$. (Multiplying by a constant gain K_p does not alter the phase lag.) We have

$$G(\mathrm{i}\omega) = \frac{\mathrm{e}^{-\sqrt{\omega/2}}}{\sqrt{\omega/2}}\left(\cos\sqrt{\frac{\omega}{2}} - \mathrm{i}\sin\sqrt{\frac{\omega}{2}}\right)(1-\mathrm{i})$$

To shortcut the calculation, we note that a phase lag of π implies $\tan\phi = 0$ implies that $\mathrm{Im}\,G(\mathrm{i}\omega) = 0$. Thus, we extract from our expression for $G(\mathrm{i}\omega)$ the condition

$$\cos\sqrt{\frac{\omega}{2}} + \sin\sqrt{\frac{\omega}{2}} = 0\,,$$

which implies $\tan\sqrt{\frac{\omega}{2}} = -1$, so that the phase delay is $3\pi/4$. Thus,

$$\omega^* = \frac{9}{8}\pi^2 \approx 11.1 \to 11.1\frac{D}{\ell^2}\,.$$

b. *Critical gain K_p^*.* The loop gain is

$$L = KG = K_\mathrm{p}\frac{\mathrm{e}^{-\sqrt{s}}}{\sqrt{s}} = K_\mathrm{p}\frac{\mathrm{e}^{-\sqrt{\mathrm{i}\omega}}}{\sqrt{\mathrm{i}\omega}} = K_\mathrm{p}\frac{\mathrm{e}^{-\sqrt{\omega/2}}(1+\mathrm{i})}{\sqrt{\mathrm{i}\omega}}\,.$$

The magnitude $|L|$ is then

$$|L|^2 = \frac{K_\mathrm{p}^2}{\omega}\,\mathrm{e}^{-\sqrt{\frac{\omega}{2}}}(1+\mathrm{i})\,\mathrm{e}^{-\sqrt{\frac{\omega}{2}}}(1-\mathrm{i}) = \frac{K_\mathrm{p}^2}{\omega}\,\mathrm{e}^{-\sqrt{2\omega}} = 1\,,$$

with solution

$$K_\mathrm{p}^* = \sqrt{\omega^*}\,\exp\left\{\sqrt{\frac{\omega^*}{2}}\right\} = \sqrt{2}\left(\frac{3}{4}\pi\right)\exp\left\{\left(\tfrac{3}{4}\pi\right)\right\} \approx 35.2 \to 35.2\frac{\lambda}{\ell}\,.$$

3.10 Instability with a long delay. Consider the unstable system $\dot{x} = ax + u$, with $u(t) = -K_\mathrm{p}x(t-\tau)$. Show that $a < K_\mathrm{p} < \tau^{-1}$ for stability. Thus, when $\tau > a^{-1}$ no value of K_p can make all roots of s have a negative real part. Hint: When s is real, expand the exponential. Consider a possible Hopf bifurcation ($s = \pm\mathrm{i}\omega$), too.

Solution. For

$$\dot{x}(t) = ax(t) + u(t)\,, \qquad u(t) = -K_\mathrm{p}x(t-\tau)\,,$$

with $a > 0$, the Laplace transform is

$$\left(s - a + K_\mathrm{p}\,\mathrm{e}^{-s\tau}\right)x(s) = 0\,.$$

The basic strategy will be to look at the roots of the transcendental equation for s in the vicinity of the instability bifurcation point, where the solution with largest real part is zero. There are two cases: s is real and $s = 0$ at the instability, or s is complex and equals $\pm\mathrm{i}\omega$ at the instability. The latter signals a Hopf bifurcation with an oscillatory instability that is oscillating at an angular frequency ω at onset.

- s real. Since $s = 0$ at the bifurcation, we expand $\mathrm{e}^{-s\tau}$ to second order near the bifurcation. This implies

$$s - a + K_\mathrm{p}[1 - (s\tau) + \tfrac{1}{2}(s\tau)^2 + \cdots] = 0\,.$$

Rearranging gives a quadratic equation for s:

$$\tfrac{1}{2}K_\mathrm{p}\tau^2 s^2 + (1 - K_\mathrm{p}\tau)s + (K_\mathrm{p} - a) = 0$$
$$s^2 - \left(\frac{K_\mathrm{p}\tau - 1}{\frac{1}{2}K_\mathrm{p}\tau^2}\right)s + \left(\frac{K_\mathrm{p} - a}{\frac{1}{2}K_\mathrm{p}\tau^2}\right) = 0\,.$$

The product of the two (real) roots of s is $\sim (K_\mathrm{p} - a) > 0$, since the delay-free equation already tells us $K_\mathrm{p} > 0$. Thus, both roots must be negative (and their product positive). We also know that the sum of the roots must be $\sim (K_\mathrm{p}\tau - 1) < 0$. This implies that $a < K_\mathrm{p} < \tau^{-1}$, or $\tau < K_\mathrm{p}^{-1} < a^{-1}$.

- s complex. Now $s = \pm \mathrm{i}\omega$ at the bifurcation, and we substitute into the transcendental equation: $\mathrm{i}\omega - a + K_\mathrm{p}\,\mathrm{e}^{-\mathrm{i}\omega\tau} = 0$. Separating real and imaginary parts gives

$$-a + K_\mathrm{p}\cos\omega\tau = 0\,, \qquad \omega - K_\mathrm{p}\sin\omega\tau = 0\,.$$

Solving for $\cos\omega\tau$ and $\sin\omega\tau$ and using $\cos^2 + \sin^2 = 1$ gives

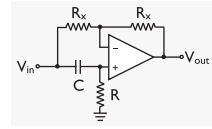
$$\left(\frac{a}{K_\mathrm{p}^2}\right)^2 + \left(\frac{\omega}{K_\mathrm{p}^2}\right)^2 = 1\,, \quad \Longrightarrow \quad a^2 + \omega^2 = K_\mathrm{p}^2\,.$$

Similarly, dividing the two equations gives

$$\tan\omega\tau = \omega/a\,,$$

which has a solution $\omega \neq 0$ only when $\tau < a^{-1}$. Otherwise, $\omega = 0$ is the only solution, and we are back in the first case. In conclusion, we need $\tau < a^{-1}$ to be able to find a gain $K_\mathrm{p} > a$ that ensures stability (all roots s have Re $s < 0$).

3.11 Op amp allpass. Consider the op-amp circuit shown at right. Derive the transfer function between V_in and V_out and show that it is all pass, with a zero in the RHP. You should find that the transfer function is independent of R_x.



Solution.

From the current path through the R_x resistors, we see that

$$\frac{V_\mathrm{in} - V_-}{R_x} = \frac{V_- V_\mathrm{out}}{R_x}\,,$$

which implies

$$V_- = \frac{1}{2}(V_\mathrm{in} + V_\mathrm{out})\,.$$

For the lower path through C and R, we use the op-amp rule $V_+ = V_-$ to write

$$\frac{V_\mathrm{in} - \frac{1}{2}(V_\mathrm{in} + V_\mathrm{out})}{1/(\mathrm{i}\omega C)} = \frac{\frac{1}{2}(V_\mathrm{in} + V_\mathrm{out})}{R}\,,$$

which leads to

$$V_\mathrm{out} = -\left(\frac{1 - \mathrm{i}\omega RC}{1 + \mathrm{i}\omega RC}\right) V_\mathrm{in} = -\left(\frac{1 - sRC}{1 + sRC}\right) V_\mathrm{in}\,.$$

Thus, the transfer function $G(s) = V_\mathrm{out}/V_\mathrm{in}$ is all pass, with a zero at $s = +(RC)^{-1}$, in the RHP.

3.12 Inverse response. NMP zeros can lead to *inverse* response: the initial response to a step input is in the opposite direction to the step. We saw such behavior in Example 3.3. Here, we explore this behavior analytically for that example, as well as for a slightly more general transfer function, $G_0(s) = \frac{1-s/z}{(1+s/p)(1+s/p^*)}$.

a. Show that the initial value of the derivative of the response $y(t)$ to a step function $\theta(t)$ is, for a general transfer function $G(s)$, given by $\dot{y}(t \to 0) = \lim_{s\to\infty} s\,G(s)$. Hint: see the discussion of the initial-value theorem in Appendix A.4.3.
b. Then show that the transfer function $G_{\rm NMP}$ in Ex. 3.3 has an inverse response.
c. Show that G_0 has an inverse response when there is an NMP zero.

Solution.

a. *Initial value of derivative*. Recall the Laplace transform,

$$\ddot{y}(s) = s^2\,y(s) - s\,\cancelto{0}{y(t=0)} - \dot{y}(t=0)\,.$$

Directly evaluating

$$\lim_{s\to\infty} \ddot{y}(s) = \lim_{s\to\infty} \int_0^\infty \mathrm{d}t\,\ddot{y}(t)\,\mathrm{e}^{-st} = 0\,,$$

we solve for $\dot{y}(t=0)$:

$$\left.\frac{\mathrm{d}y}{\mathrm{d}t}\right|_{t=0} = \lim_{s\to\infty} s^2 y(s) = \lim_{s\to\infty} s^2\,G(s)\,\frac{1}{s} = \lim_{s\to\infty} s\,G(s)\,.$$

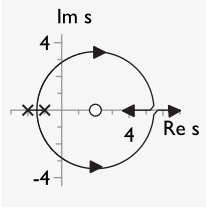
b. *Inverse response for zero at $z = 1$*. For G_{NMP},

$$\left.\frac{\mathrm{d}y}{\mathrm{d}t}\right|_{t=0} = \lim_{s\to\infty} s\,\frac{1}{(1+s)}\left(\frac{1-s/2}{1+s/2}\right) = \boxed{-1}\,,$$

c. *Inverse response for RHP zero*. For G_0,

$$\left.\frac{\mathrm{d}y}{\mathrm{d}t}\right|_{0} = \lim_{s\to\infty} \frac{-s^2/z}{(s/p)(s/p^*)} = \boxed{-\frac{pp^*}{z}}\,,$$

showing that the sign of $\dot{y}(t=0)$ is *opposite* that of z: if z is an RHP zero, the system initially goes the "wrong way." Notice that the initial response does not depend on whether the pole is stable or not (only $|p|^2$ enters). Thus, an unstable system with $z < 0$ would still show normal response.



3.13 Response of a non-minimum phase (NMP) system. In Section 3.6.2, we analyzed the system $G_{\rm NMP}(s) = \frac{1}{1+s}\left(\frac{1-s/2}{1+s/2}\right)$, which has a zero in the right-hand plane (RHP). Reproduce the pole-zero plot at left and describe its main features analytically.

Solution.

The closed-loop transfer function for proportional gain K is

$$T = \frac{K(s-2)}{s^2 + (3-K)s + 2(1+K)}\,.$$

The structure of the root-locus plot can be understood by solving for the pole positions of $T(s)$ as a function of gain K:

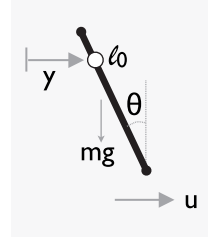
$$s = \tfrac{1}{2}\left(K - 3 \pm \sqrt{K^2 - 14K + 1}\right)\,.$$

There are two degenerate roots when $K^2 - 14K + 1 = 0$, which happens at $K = 7 \pm 4\sqrt{3}$, which corresponds to $s = 2(1 \pm \sqrt{3})$. The instability is calculated for Re $s = 0$, which occurs when $K = 3$.

To show that the shape of the locus is a circle centered on $s = 2$, we note that the equation of such a circle is of the form (Re $s - 2)^2 +$ Im $s^2 = R^2$, where R is the unknown radius of the circle. Using Re $s = \frac{1}{2}(K - 3)$ and Im $s^2 = -(K^2 - 14K + 1)$ gives $R = \sqrt{12}$.

Thus, the following "narrative" describes the root-locus plot, as K is increased from 0: The two poles first move towards each other, colliding at $s = -2(\sqrt{3} - 1) \approx -1.46$ for $K = 7 - 4\sqrt{3} \approx 0.072$. The roots then turn into a complex-conjugate pair that travel in a circle in the s-plane of radius $2\sqrt{3} \approx 3.46$. The system becomes unstable for $K^* = 3$, where $\omega^* = \sqrt{8} \approx 2.82$. If one further increases the gain, the circle closes back and the two complex-conjugate pairs meet at $s = 2(1 + \sqrt{3}) \approx 5.46$, where $K = 7 + 4\sqrt{3} \approx 13.93$. Thereafter, the poles become real. One of them approaches the RHP zero at $s = 2$, and the other goes off to infinity.

3.14 Balancing a stick by moving your hand. In Problem 2.1, we showed, neglecting any mass associated with the hand or arm, that $\ddot{x} + \frac{1}{2}\left(\ddot{\theta}\cos\theta - \dot{\theta}^2 \sin\theta\right) = u$ and $\ddot{\theta} + \sin\theta + \frac{3}{2}\ddot{x}\cos\theta = 0$, where θ increases counterclockwise, from the bottom, and x is the horizontal displacement of the stick bottom, relative to a reference position.



a. Linearize the equations of motion about the vertical equilibrium $\theta = \pi$.
b. Show that the transfer function from u to the *fixation point* $y = x + \ell_0 \sin\theta$ is $G(s) = [(1 - \frac{3}{2}\ell_0)s^2 - 1]/[s^2(\frac{1}{4}s^2 - 1)]$.
c. From Problems 3.16 and 3.17, show that one cannot balance a stick if $\ell_0 \leq \frac{1}{2}$.

Solution.

a. Linearizing about the unstable vertical equilibrium $\theta = \pi$ gives

$$\begin{aligned} \ddot{x} - \tfrac{1}{2}\ddot{\theta} &= u\,, \\ \ddot{\theta} - \theta - \tfrac{3}{2}\ddot{x} &= 0 \\ y &= x - \ell_0\theta\,. \end{aligned}$$

b. Take the Laplace transform and write the first two equations in matrix form:

$$\begin{pmatrix} s^2 & -\frac{1}{2}s^2 \\ -\frac{3}{2}s^2 & s^2 - 1 \end{pmatrix}\begin{pmatrix} x \\ \theta \end{pmatrix} = \begin{pmatrix} u \\ 0 \end{pmatrix}$$

Solve for $x(s)$ and $\theta(s)$ in terms of $u(s)$:

$$\begin{pmatrix} x \\ \theta \end{pmatrix} = \frac{1}{s^2\left(\frac{1}{4}s^2 - 1\right)}\begin{pmatrix} s^2 - 1 \\ \frac{3}{2}s^2 \end{pmatrix} u\,.$$

Writing $y = x - \ell_0\theta$ then gives the transfer function

$$G(s) = \frac{y(s)}{u(s)} = \frac{\left(1 - \frac{3}{2}\ell_0\right)s^2 - 1}{s^2\left(\frac{1}{4}s^2 - 1\right)}.$$

c. From the transfer function, the poles are at $p = 0, \pm 2$. The $+2$ pole is associated with the unstable motion. The relevant zero is at $z = +1/\sqrt{1 - (3/2)\ell_0}$. From Problem 3.17, the sensitivity function has a maximum of magnitude

$$\max S = \frac{p + z}{|p - z|},$$

which diverges for $p \to z$. For the relevant pole-zero combination, this condition occurs at $\ell_0 = 1/2$. More carefully, Problem 3.16 shows that the feedback bandwidth $\omega_c > p$. But $\omega_c < z$ for an unstable zero. We can satisfy both constraints when $z > p$ but not when $z < p$. Combining all these arguments, we conclude that one must look at a point $\ell_0 > 1/2$ (more than halfway up the stick from the hand).

3.15 Balancing a stick, with delay in applying feedback.

a. For small deviations of a stick from the vertical, show that the equation of motion with delayed PD feedback is $\ddot{\theta}(t) - (6g/\ell)\theta(t) = -K_p\theta(t-\tau) - K_d\dot{\theta}(t-\tau)$. As in Problem 3.14, neglect the mass of the hand in your calculation.

b. Inserting $\theta(t) \sim e^{st}$, show that no choice of K_p and K_d stabilizes the stick for $\tau > \sqrt{\ell/(3g)}$. Thus, there is instability for $\ell < 3g\tau^2$. Hint: Write $s = i\omega$; separate into real and complex equations; expand $\cos\omega\tau$; and look for ω roots in $0 < \omega\tau < \frac{\pi}{2}$. A careful argument about the roots is subtle (Stepan, 2009).

c. Whatever the controller, disturbances grow uncorrected over the delay time τ. Use this idea to argue that $\ell \lesssim g\tau^2$ implies instability. Experiments by Milton et al. (2016) suggest that, in humans, $\tau \approx 0.23$ s, while $\ell_{\min} \approx 0.32$ m. Show that these observations imply that uncontrolled disturbances grow by a factor ≈ 20. Humans thus seem to use memory and an internal model to predict motion.

Solution.

a. In terms of the scaling from Problem 3.14, the transfer function between u and θ is

$$\theta(s) = \frac{3/2}{s^2/4 - 1}u(s), \qquad \Longrightarrow \qquad \ddot{\theta} - 4\theta = 6u.$$

That problem scaled time by $\omega_0 = \frac{3}{2}g/\ell$, which implies

$$\ddot{\theta} - 4\left(\frac{3}{2}\frac{g}{\ell}\right) = 6\left(\frac{3}{2}\frac{g}{\ell}\right)u$$

$$\ddot{\theta}(t) - \left(\frac{6g}{\ell}\right)\theta(t) = -K_p\theta(t-\tau) - K_d\dot{\theta}(t-\tau),$$

where the PD gains K_p and K_d absorb all needed constants.

b. It is convenient to scale time by the slightly different scaling $t \to \sqrt{6g/\ell}\, t$. In this scaling, $\ddot{\theta} - \theta = -K_\mathrm{p}\theta(t-\tau) - K_\mathrm{d}\dot{\theta}(t-\tau)$, with K_d redefined, too. Substituting $\theta(t) = \theta_0\ \mathrm{e}^{st}$ into the equation of motion leads to the characteristic equation

$$s^2 - 1 + K_\mathrm{p}\,\mathrm{e}^{-s\tau} + K_\mathrm{d}\, s\,\mathrm{e}^{-s\tau} = 0\,.$$

At the threshold of instability $s = 0$ (real root) or $s = \pm \mathrm{i}\omega$ (Hopf bifurcation). For the former case $K_\mathrm{p} = 1$. For the latter, we set $s = \mathrm{i}\omega$:

$$-\omega^2 - 1 + K_\mathrm{p}\,\mathrm{e}^{-\mathrm{i}\omega\tau} + K_\mathrm{d}\, s\,\mathrm{e}^{-\mathrm{i}\omega\tau} = 0\,.$$

Separating the imaginary and real parts of the ω equation then gives

$$\begin{aligned} -K_\mathrm{p}\sin\omega\tau + \omega K_\mathrm{d}\cos\omega\tau &= 0 \\ -(1+\omega^2) + \omega K_\mathrm{d}\sin\omega\tau + K_\mathrm{p}\cos\omega\tau &= 0\,. \end{aligned}$$

In matrix form, this is

$$\begin{pmatrix} -K_\mathrm{p} & \omega K_\mathrm{d} \\ \omega K_\mathrm{d} & K_\mathrm{p} \end{pmatrix} \begin{pmatrix} \sin\omega\tau \\ \cos\omega\tau \end{pmatrix} = \begin{pmatrix} 0 \\ 1+\omega^2 \end{pmatrix}$$

$$\implies \quad \begin{pmatrix} \sin\omega\tau \\ \cos\omega\tau \end{pmatrix} = \frac{1}{K_\mathrm{p}^2 + \omega^2 K_\mathrm{d}^2} \begin{pmatrix} -K_\mathrm{p} & \omega K_\mathrm{d} \\ \omega K_\mathrm{d} & K_\mathrm{p} \end{pmatrix} \begin{pmatrix} 0 \\ 1+\omega^2 \end{pmatrix}$$

$$= \frac{1}{K_\mathrm{p}^2 + \omega^2 K_\mathrm{d}^2} \begin{pmatrix} \omega K_\mathrm{d} \\ K_\mathrm{p} \end{pmatrix} (1+\omega^2) = \begin{pmatrix} \omega K_\mathrm{d} \\ K_\mathrm{p} \end{pmatrix} (1+\omega^2)^{-1}\,,$$

using the relation established by the identity $\sin^2\omega\tau + \cos^2\omega\tau = 1$,

$$K_\mathrm{p}^2 + \omega^2 K_\mathrm{d}^2 = (1+\omega^2)^2\,.$$

We are interested in the roots of the cosine equation,

$$K_\mathrm{p} = (1+\omega^2)\cos\omega\tau\,,$$

for $K_\mathrm{p} = 1$, its smaller possible value. Numerically, there is a positive root ω that approaches 0 for $\tau \approx 1.4$. Exploring the characteristic equation for s numerically, it is easy to see that this root corresponds to a Hopf bifurcation and that all other roots are stable. Thus, there is critical delay τ_c and near that delay $\omega\tau \ll 1$. This suggesting expanding the cosine:

$$\begin{aligned} K_\mathrm{p} &\approx (1+\omega^2)\left(1 - \tfrac{1}{2}\omega^2\tau^2 + \tfrac{1}{24}\omega^4\tau^4 + \cdots\right) \\ &= 1 + \left(1 - \tfrac{1}{2}\tau^2\right)\omega^2 + \tfrac{1}{24}\left(\tau^4 - 12\tau^2\right)\omega^4 + \cdots\,. \end{aligned}$$

Recall for $\tau = 0$ that $K_\mathrm{p} \geq 1$. Then, for $\tau > \sqrt{2}$, there is no solution in the range $0 < \omega\tau < \frac{1}{2}\pi$. In physical units,

$$\tau_\mathrm{c} = \sqrt{2}\sqrt{\frac{\ell}{6g}} = \sqrt{\frac{\ell}{3g}}\,.$$

Filling in the steps we justified by numerical exploration turns out to be complicated. See Stepan (2009) and the references cited therein.

c. A disturbance at time $t-\tau$ will not become known until time t. According to the linear calculation, it will grow as

$$\theta(t) = \theta(t-\tau)\ \mathrm{e}^{\tau\sqrt{6g/\ell}}\,,$$

where $\theta(t-\tau)$ results from a disturbance a time τ in the past. What counts here is the growth factor, which magnifies any initial disturbance by $\exp\left\{[\tau\sqrt{6g/\ell}]\right\}$ before the feedback can kick in. Milton et al. (2016) find that the sensory delay $\tau \approx 0.23$ s and that the shortest stick (for experts, who had to remain seated) was $\ell \approx 0.32$ m. Inserting these numbers gives

$$\frac{\theta(t)}{\theta(t-\tau)} \approx \mathrm{e}^{0.23\sqrt{6\times 9.8/0.32}} \approx 20\,.$$

That is, for the shortest sticks that could be balanced, disturbances are amplified by a factor of about 20. Milton et al. (2016) take 20° as an indicator of instability, suggesting that the relevant perturbations are roughly 1°, which seems plausible. They also emphasize that the observed lower stick limit is well below what would be expected from the PD argument given in the previous section, ruling out that kind of control model.

If the mass M of the hand (or arm) plays a role, as arguably it might, we add $M\ddot{x}$ to the x-equation of motion and find a revised critical stick length of

$$\ell_{\mathrm{c}} = \left(\frac{M+m}{4M+m}\right) 3g\tau^2\,,$$

which, for $M \gg m$, approaches $\ell_{\mathrm{c}} = \frac{3}{4}g\tau^2$. The factor of four reduction brings the numbers closer to the experimental results based on PD control (Milton et al., 2016).

However, we should also note that Milton et al. (2016) consider a model that in addition assumes a *dead zone*, which can range from $\approx 0.8°$ for a trained expert stick balancer to 2–3° for a novice. In either case, we do not sense small angular deviations below some threshold. Including a dead zone will increase the minimum stick length. It is safest to conclude that the simple model presented in this problem is a start and captures at least some of the important physics, but "real stick balancing" likely requires a more detailed model.

3.16 Control of NMP systems, 1. Non-minimum phase systems restrict the possible feedback bandwidths. Consider an RHP zero and then an RHP pole (both real). Apply the Bode gain-phase relation to the minimum-phase part of the system. Decompose the loop transfer function $L(s) = G(s)K(s) = G_{\mathrm{mp}}(s)G_{\mathrm{ap}}(s)K(s)$, where G_{mp} is minimum phase, G_{ap} is all pass, and $K(s)$ is the controller. Let n be the slope of the gain curve at the crossover frequency ω^*, defined by $|L(\mathrm{i}\omega^*)| = 1$.

a. *Simple zero*. For $G_{\mathrm{ap}} = \frac{z-s}{z+s}$, show that we must choose $\omega^* < z\tan\frac{\varphi}{2}$, where z is the position of the zero and where the phase lag $\varphi \equiv \pi - \varphi_{\mathrm{m}} + n\frac{\pi}{2}$, with φ_{m} being the desired phase margin. For $\varphi_{\mathrm{m}} = 90° = \pi/2$, the bandwidth $\omega^* < z$.

b. *Simple pole*. For $G_{\rm ap} = \frac{s+p}{s-p}$, show that the *minimum* bandwidth to stabilize is $\omega^* > p/\tan\frac{\varphi}{2}$. For $\varphi_{\rm m} = 90°$, the bandwidth $\omega^* > p$. Intuitively, to stabilize an unstable system, the feedback must correct a perturbation faster than it grows.

Solution.

a. *Simple zero*. The loop transfer function is $L(s) = G_{\rm mp}(s)\,G_{\rm ap}(s)\,K(s)$, and $L(\mathrm{i}\omega^*) = -1$ at instability. Since we are considering fundamental limits, we assume that the controller has zero phase shift at ω^*. Thus, we can write for the phases

$$n\tfrac{\pi}{2} - \varphi - \varphi_{\rm m} = -\pi\,,$$

where $n\frac{\pi}{2}$ represents the contribution of $G_{\rm mp}$ at the crossover (with n typically negative), $-\varphi$ represents the contribution of $G_{\rm ap}$, $-\varphi_{\rm m}$ the desired phase margin, and $-\pi$ the instability condition. Thus,

$$\varphi = \pi + n\tfrac{\pi}{2} - \varphi_{\rm m}\,.$$

The next step is to find the frequency delay φ due to the all-phase component $G_{\rm ap}$.

$$G_{\rm ap}(\mathrm{i}\omega) = \frac{z - \mathrm{i}\omega}{z + \mathrm{i}\omega} = -\frac{\left(z^2 - \omega^2\right) - 2\,\mathrm{i}z\omega}{z^2 + \omega^2}\,,$$

so that

$$\tan\varphi = +\frac{2z\omega}{z^2 - \omega^2} \quad \rightarrow \quad \frac{2\omega}{1 - \omega^2}\,,$$

where in the last step we scale $\omega \to \omega/z$ and we switch the sign because the problem implies that we have defined the phase delay $\varphi > 0$. Then

$$\omega^2 + 2\left(\frac{1}{\tan\varphi}\right)\omega - 1 = 0\,,$$

whose solution is

$$\omega^* = -\frac{1}{\tan\varphi} \pm \sqrt{\left(\frac{1}{\tan\varphi}\right)^2 + 1} = -\frac{1}{\tan\varphi} + \frac{1}{\sin\varphi} = \frac{-\cos\varphi + 1}{\sin\varphi} = \tan\frac{\varphi}{2}\,,$$

where we take the positive root because $\varphi \geq 0$ and $\omega^* \geq 0$. Unscaling and remembering that φ is the limiting frequency gives

$$\omega^* < z\tan\tfrac{\varphi}{2}\,.$$

b. *Simple pole.* The story is similar.

$$G = \frac{s+p}{s-p} \quad \rightarrow \quad \frac{s+1}{s-1} \quad \rightarrow \quad \frac{\mathrm{i}\omega + 1}{\mathrm{i}\omega - 1} = -\frac{1 - \omega^2 + 2\,\mathrm{i}\omega}{\omega^2 + 1}\,,$$

where we scale $s \to s/p$. The phase is then

$$\tan\varphi = -\frac{2\omega}{1-\omega^2} = \frac{2\omega}{\omega^2-1},$$

which implies

$$\omega^2 - 2\left(\frac{1}{\tan\varphi}\right)\omega - 1 = 0,$$

with solution

$$\omega^* = \frac{1}{\tan\varphi} \pm \sqrt{\left(\frac{1}{\tan\varphi}\right)^2 + 1} = \frac{1}{\tan\varphi} + \frac{1}{\sin\varphi} = \frac{\cos\varphi + 1}{\sin\varphi} = \frac{1}{\tan\frac{\varphi}{2}},$$

which implies

$$\omega^* > p\left(\frac{1}{\tan\frac{\varphi}{2}}\right),$$

where we unscale ω and where the inequality comes because the right-hand side is decreasing with increasing frequency.

3.17 Control of NMP systems, 2. If pole and zero are both unstable, you are caught between a rock and a hard place: the pole imposes minimum bandwidth requirements while the zero imposes maximum requirements. The sensitivity function $S \equiv \frac{1}{1+L}$ then cannot be small at all frequencies. For $L(s) = K(\frac{z-s}{s-p})$ and $p, z > 0$, show that

a. the system is stable if $\frac{p}{z} < K < 1$ or $1 < K < \frac{p}{z}$;
b. the largest stability margin is $s_{\mathrm{m}} = \frac{|p-z|}{p+z}$, for $K = \frac{2p}{p+z}$ (Hint: Look at the Nyquist plot of the loop transfer function as a function of K.);
c. the maximum magnitude of S equals or exceeds $|S| = \frac{p+z}{|p-z|}$.

Thus, poles near zeros make control difficult. Recall that S gives the sensitivity to disturbances, with $|S| = 1$ being open loop. If $|S(\mathrm{i}\omega)| > 1$, then disturbances at that frequency are amplified. An RHP zero and pole guarantees such a frequency. And if they are close, you will do much worse (Åström and Murray, 2008).

Solution.

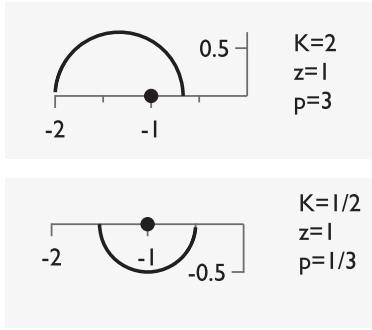
a. *Stability conditions*. The instability threshold is at $L(s) = -1$, which gives the equation

$$K\left(\frac{z-s}{s-p}\right) = -1 \qquad \Longrightarrow \qquad s = \frac{Kz-p}{K-1}.$$

The root is real and stable ($s < 0$) for $1 < K < p/z$ or $p/z < K < 1$.

b. *Largest stability margin*. Let's first consider $z < p$. Then, a quick plot of the Nyquist diagram (or a proof. . . .) shows that the Nyquist plot is a circle

whose center lies on the real axis of the s-domain. See below. The zero- and infinite-frequency loop gains are



$$L(0) = -\frac{Kz}{p} \qquad L(\infty) = -K\,.$$

The stability margins are just the closest distance between the critical point (-1) and the transfer function curve, in the s-plane. This distance is then the minimum of $(-K, -1)$ and $(-1, -Kz/p)$ for $z < p$ and similarly for the other case. The largest stability margin will occur when the left and right margins are equal (the bottom case in the plots at left). Thus,

$$-1 + K = -K\frac{z}{p} + 1\,, \qquad \Longrightarrow \qquad K = \frac{2p}{p+z}\,.$$

The margin itself is given by $s_\mathrm{m} = -1 + K = \frac{p-z}{p+z}$. Doing the $z > p$ case gives $s_\mathrm{m} = -\frac{p-z}{p+z}$, giving the absolute value.

c. *Sensitivity function*. $S = \frac{1}{1+L}$ is maximized by minimizing the denominator. But the stability margin is the point where L is closest to -1, meaning that max S occurs at the frequency that gives s_m, which means here that

$$\max S = \frac{1}{s_\mathrm{m}} = \frac{p+z}{|p-z|}\,.$$

3.18 Flexible string transfer function. A string supporting transverse waves $\psi(x,t)$ of unit velocity is driven at one end, $\psi(0,t) = u(t)$ and free at $x = 1$. Show that $G(s) = \frac{\psi(x,s)}{u(s)} = \frac{\cosh s(1-x)}{\cosh s}$ for an observation point $0 \le x \le 1$. Plot the magnitude of frequency response for $x = \{0, \frac{1}{2}, \frac{1}{4}, 1\}$, and discuss its structure.

Solution. The one-dimensional wave equation for transverse displacements of a string is

$$\partial_{xx}\psi = \partial_{tt}\psi\,, \qquad \psi(0,t) = u(t)\,, \qquad \partial_x\psi(1,t) = 0\,.$$

Because we are interested in steady-state response, we do not need to worry about the initial condition. Let us first Laplace transform all the equations in time:

$$\partial_{xx}\psi = s^2\psi\,, \qquad \psi(0,s) = u(s)\,, \qquad \partial_x\psi(1,s) = 0\,.$$

Solving for $\psi(x,s)$ gives

$$\psi(x,s) = a(s)\ \mathrm{e}^{sx} + b(s)\ \mathrm{e}^{-sx}\ .$$

The boundary condition at $x = 0$ implies

$$\psi(0,s) = a(s) + b(s) = u(s)\,.$$

The boundary condition at $x = 1$ implies

$$\partial_x\psi(1,s) = s\,a(s)\ \mathrm{e}^{s} - s\,b(s)\ \mathrm{e}^{-s} = 0\,, \qquad \Longrightarrow \qquad b(s) = a(s)\ \mathrm{e}^{2s}$$
$$\Longrightarrow \qquad a(s) = \frac{u(s)}{1+\mathrm{e}^{2s}}\,.$$

Then

$$\begin{aligned} G(s) \equiv \frac{\psi(x,s)}{u(s)} &= \frac{\mathrm{e}^{-s}}{\mathrm{e}^{-s}+\mathrm{e}^{s}}\,\mathrm{e}^{sx} + \frac{\mathrm{e}^{s}}{\mathrm{e}^{-s}+\mathrm{e}^{s}}\,\mathrm{e}^{-sx} \\ &= \frac{\mathrm{e}^{-s(1-x)} + \mathrm{e}^{s(1-x)}}{\mathrm{e}^{-s}+\mathrm{e}^{s}} \\ &= \frac{\cosh s(1-x)}{\cosh s}\,. \end{aligned}$$

The complex frequency response is

$$G(\mathrm{i}\omega) = \frac{\cos\omega(1-x)}{\cos\omega}\,.$$

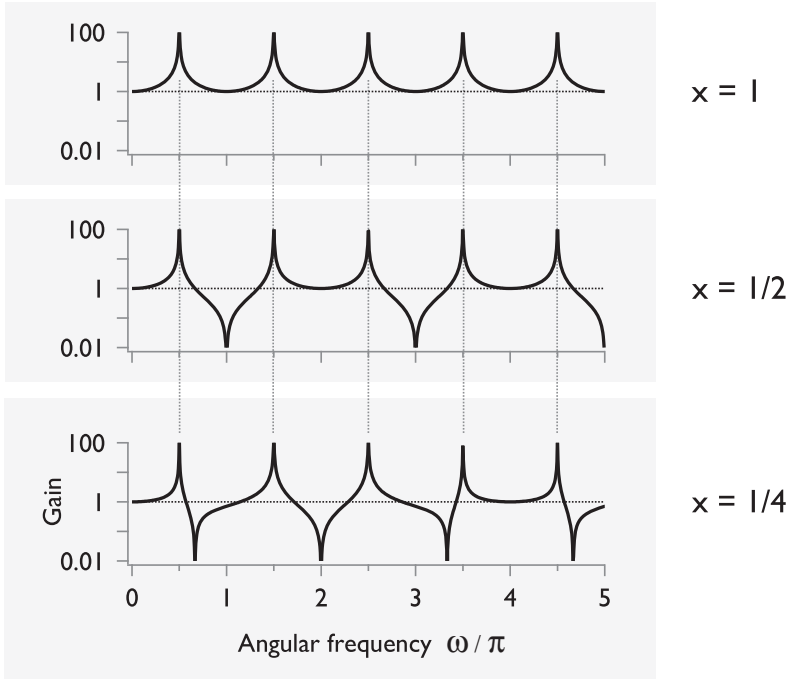
Note that in physical units $\omega \to \omega L/c$, where L is the domain size and c the wave propagation speed.

Now we consider this response for several observation points x. A first point to note, in general, is that there is no x dependence in the denominator, only the numerator. This means that the poles, which satisfy $\cos\omega = 0$, or $\omega = j\pi/2$ ($j = 1, 3, 5, \ldots$), are the same for all cases. But the position of zeros will depend on the observation point x.

- $x = 0$. Then $G(\mathrm{i}\omega) = 1$: the observation point is locked to the excitation and follows it exactly, at all frequencies.
- $x = 1$. Then $G(\mathrm{i}\omega) = 1/\cos\omega$. There are no zeros. The derivative $\mathrm{d}_\omega G = \sin\omega/\cos^2\omega$ vanishes at $\omega = n\pi$, with n an integer. At those values $G(in\pi) = 1$.
- $x = \frac{1}{2}$. Then $G(\mathrm{i}\omega) = (\cos\omega/2)/\cos\omega$ and $\mathrm{d}_\omega G \sim (3\sin\frac{\pi}{2}\omega + \sin\frac{3\pi}{2}\omega)$, implying that the numerator vanishes at $\omega = j\pi$, with $j = 1, 2, \ldots$. Unlike the $x = 1$, though, there is a zero only when $\omega = \pi, 3\pi, 5\pi, \ldots$. At the even values of n, the denominator also vanishes. L'Hôpital's Rule then implies $G = 1$ at those values.

- $x = \frac{1}{4}$. Then $G(\mathrm{i}\omega) = (\cos 3\omega/4)/\cos\omega$ and $\mathrm{d}_\omega G \sim (7\sin\frac{\pi}{4}\omega + \sin\frac{7\pi}{4}\omega)$. Whenever $\omega = 0,\ 4\pi,\ 8\pi, \ldots$, the derivative equation vanishes, leading to a minimum with $G = 1$. On the other hand, the numerator vanishes when $\omega/\pi = \frac{2}{3}, 2, \frac{8}{3}$, etc.

Below are magnitude response plots illustrating the above results, plotted using linear frequency to emphasize the periodic nature of the poles at $\omega = j\frac{\pi}{2}$, for j an odd integer.



Finally, we note that if x is a rational fraction, then so is $1 - x$. If $1 - x = k/\ell$, with k and ℓ both integers, then there is a zero at $\omega/\pi = \ell/(2k)$. Again, the actual response can be rather complex, for x not close to a "nice" fraction. Physically, the zeros result from interference between direct and reflected waves. Because they require a delicate cancellation, they are extremely sensitive to the observation point x (and frequency ω).

3.19 Zero cancellation. To see why cancelling a zero is dangerous, consider $G(s) = \frac{s+z}{(s+p_1)(s+p_2)}$. Find a controller $K(s)$ such that $T(s) = \frac{KG}{1+KG} = \frac{1}{(s+p_1)(s+p_2)}$ is the closed loop transfer function. What goes wrong?

Solution.

$$K(s) = G^{-1}\left(\frac{1}{T^{-1}-1}\right) = \frac{(s+p_1)(s+p_2)}{s+z}\,\frac{1}{(s+p_1)(s+p_2)-1}\,.$$

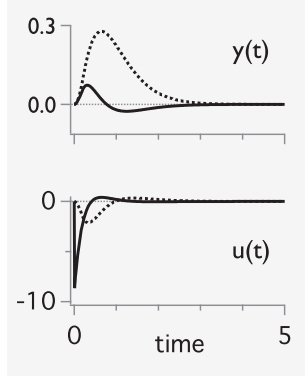
If you have an RHP zero, you would be adding an unstable pole. If the cancellation is not perfect – and it never is – then you will have unstable dynamics. You may also have unbounded internal signals. See the discussion in Section 3.5.3.

3.20 Synthesizing a PID controller. Example 3.5 used pole placement to synthesize a controller $K(s) = \frac{18+26s}{9+s}$ for the undamped oscillator $G(s) = \frac{1}{s^2+1}$. Repeat the controller synthesis using a PID form, $K(s) = \frac{K_i+K_p s+K_d s^2}{s}$.

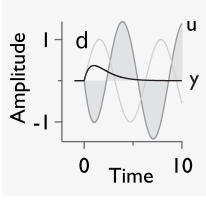
a. Find K_p, K_i, and K_d.
b. Plot the input disturbance response and controller input for the two controllers.

Solution.

a. The three coefficients are $K_p = 26$, $K_i = 27$, $K_d = 9$.
b. *Controller synthesis*. Below, we plot the response $y(t)$ to a delta-function input disturbance and the corresponding controller signal for the original PD controller (dotted lines) and the PID controller synthesized here (solid lines). We see that the response is faster but requires more control effort.



3.21 Rejecting an output disturbance. For $G = \frac{1}{1+s}$ and $d(t)$ a sinusoid of frequency ω:



a. Design a controller $K(s)$ to reject $d(t)$. Try choosing the controller to make the output $y(s) = \frac{N_d(s)}{(s+\omega)^2}$, where $d(s) = \frac{N_d(s)}{D_d(s)}$. Why is this a "nice" form for $y(s)$?
b. Calculate the time response of the output, $y(t)$, to a sinusoidal input, $d(t)$, of the form described above. Plot $y(t)$, $d(t)$, and $u(t)$, as shown at left for $\omega = 1$.
c. Investigate the output when $d(t) = \sin \omega_d t$ has the "wrong" frequency. For $\omega_d = \omega(1 + \epsilon)$, show that $y(t)$ converges to $\epsilon \cos \omega t + O(\epsilon^2)$.

Solution.

a. For an output disturbance,

$$y = \frac{d}{1 + KG}.$$

From the discussion in the book, we choose a controller proportional to the denominator of $d(s)$. We include a factor of G^{-1} to "cancel" the dynamics, too. This leads to

$$K(s) = K_0(s)\frac{G^{-1}(s)}{s^2+\omega^2} = \frac{K_0(s)(1+s)}{s^2+\omega^2}\,.$$

The output-disturbance sensitivity function S is

$$S(s) = \frac{1}{1+KG} = \frac{1}{1+\frac{K_0(s)}{s^2+\omega^2}} = \frac{s^2+\omega^2}{s^2+\omega^2+K_0(s)}\,.$$

The Laplace transform of a sinusoidal disturbance $d(t)$ of unknown amplitude and phase is

$$d(s) = \frac{N_\mathrm{d}(s)}{D_\mathrm{d}(s)} = \frac{sd(0)+\dot{d}(0)}{s^2+\omega^2}\,.$$

The closed-loop response to the disturbance is then

$$y(s) = S(s)d(s) = \frac{sd(0)+\dot{d}(0)}{s^2+\omega^2+K_0(s)} = \frac{sd(0)+\dot{d}(0)}{(s+\omega)^2}\,,$$

where the last form arises if we choose $K_0(s) = 2\omega s$. Notice that

$$y(t\to\infty) = \lim_{s\to 0}\frac{s[sd(0)+\dot{d}(0)]}{(s+\omega)^2} = 0$$

and

$$y(t\to 0) = \lim_{s\to\infty}\frac{s[sd(0)+\dot{d}(0)]}{(s+\omega)^2} = d_0\,.$$

The complete controller is then

$$K = \frac{(1+s)(2\omega s)}{s^2+\omega^2}\,,$$

which is biproper ($K\to 2\omega$ as $s\to\infty$).

b. The expected time response for $d(t) = \sin\omega t$ is $y(t) = \omega t\,\mathrm{e}^{-\omega t}$, which is quite satisfactory. Solving for the control $u(s) = -K(s)y(s)$ analytically (but numerically would be ok, too), we find

$$u(t) = \omega[1-(\omega-1)t]\,\mathrm{e}^{-\omega t} - \omega\cos\omega t - \sin\omega t\,,$$

which, after a transient, oscillates in order to cancel the disturbance, as plotted in the text.

c. It is straightforward, using symbolic manipulation, to find the solution when the controller is designed for a disturbance $\sin\omega t$ but the actual disturbance has a different frequency, $\sin\omega_\mathrm{d} t$. Ignoring the transient terms, we find a steady-state periodic response

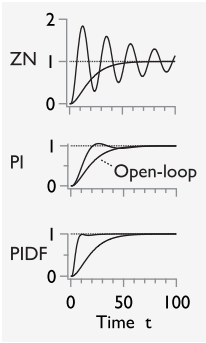
$$y_\mathrm{ss}(t) = \left(\frac{\omega^2-\omega_\mathrm{d}^2}{(\omega^2+\omega_\mathrm{d}^2)^2}\right)\left[\left(\omega^2-\omega_\mathrm{d}^2\right)\sin(\omega_\mathrm{d}t) - 2\omega\omega_\mathrm{d}\cos(\omega_\mathrm{d}t)\right]\,.$$

For a disturbance of frequency $\omega_d = \omega(1+\epsilon)$, this reduces to

$$y_{ss}(t) = \epsilon \cos \omega t + O(\epsilon^2) .$$

This $O(\epsilon)$ sensitivity is typical of naively designed feedback (or feedforward) compensation. In Chapter 9, we will see how to design controllers that increase the response order to $O(\epsilon^n)$, making it more robust to any mismatch between the designed and actual frequencies. (See, e.g., Figure Figure 9.2.) Robustness will require more control effort – bigger $u(t)$. Nothing comes for free!

3.22 Autotuning a PI controller.



a. Argue that relay feedback, Eq. (3.80), leads to output oscillations of period T_c, with critical gain $K_c = 4u_r/(\pi a)$. Here, u_r is the amplitude of the relay feedback and a the amplitude of the output oscillations. Hints: At instability, the loop gain $L(i\omega_c) = -1$. Why does only the first harmonic of the square-wave matter?

b. Consider a second-order system with delay, $G(s) = \frac{e^{-s}}{(10s+1)^2}$. Simulate the output responses $y(t)$ for both proportional control near the instability threshold and relay feedback. Show that both lead to $K_c \approx 20$ and $T_c \approx 14$.

c. Implement the PI Ziegler-Nichols rule and evaluate the closed-loop response for step command and for step input disturbance. Find PI parameters that are better than ZN. Then explore the full PIDF architecture. Reproduce the step commands at left and their associated $u(t)$. Plot the response to input step disturbances, too.

Solution.

a. In steady-state oscillation, the input $u(t)$ is a square wave of amplitude u_r and frequency ω_c, whose Fourier series is

$$u(t) = u_r \left(\frac{4}{\pi}\right) \sum_{n=1,3,\ldots}^{\infty} \frac{\sin n\omega_c t}{n} .$$

The response is essentially just a sine wave of frequency ω_c, because the amplitudes of higher harmonics are smaller at input (going as $\frac{1}{3}, \frac{1}{5}, \ldots$) and smaller again at output (because the transfer function of physical systems becomes small at high frequencies). We can thus think of our relay feedback as being equivalent to a sine wave of frequency ω_c and amplitude $\frac{4}{\pi}u_r$. This kind of approximation is sometimes called *describing function analysis*.

The second point is to argue that since the loop gain $L(s) = KG(s)$ equals -1 at instability, the frequency ω_c is the same as in proportional feedback. They are both ways of getting an instability with $L(i\omega) = -1$.

The last point is to realize that since the input amplitude is $\frac{4}{\pi}u_\mathrm{r}$, we must have $\frac{4}{\pi}u_\mathrm{r} = K_\mathrm{c}a$ in order to have $|L(\mathrm{i}\omega)| = 1$. Thus,

$$K_\mathrm{c} = \frac{4u_\mathrm{r}}{\pi a}\,.$$

b. For the "critical-gain" method using pure proportional feedback, I find $K_\mathrm{c} \approx 20.67$ and $T_\mathrm{c} \approx 14.2$. The method is to compute solutions for different gains and to look for the case where the oscillations neither grow nor decay. For relay feedback, I find $K_\mathrm{c} \approx 19.2$ and $T_\mathrm{c} \approx 14.9$. Remember the factor of $4/\pi$ to convert the amplitude of the relay output to the amplitude of the first harmonic.

You can also easily find the threshold for proportional feedback by solving the threshold condition $L(\mathrm{i}\omega) = -1$. Here, this implies

$$\frac{K\ \mathrm{e}^{-\mathrm{i}\omega}}{1 - 100\omega^2 + 20\,\mathrm{i}\omega} = -1\,,$$

or

$$K\ \mathrm{e}^{-\mathrm{i}\omega} = 100\omega^2 - 1 - 20\,\mathrm{i}\omega\,.$$

Isolating real and imaginary parts then gives two equations,

$$K\cos\omega = 100\omega^2 - 1\,, \qquad K\sin\omega = 20\omega\,.$$

Solving these transcendental equations simultaneously (using a nonlinear rootfinder routine) gives $K_\mathrm{c} \approx 20.671$ and $T_\mathrm{c} \approx 14.17$, consistent with the values found by direct numerical investigation of the solutions for different gains.

Arguably, though, finding the analytic threshold is not in the spirit of this problem. The point is that you have an unknown system and want to estimate K_c and T_c. We are comparing two different feedback algorithms (proportional control and relay control) that both give estimates of the quantity. The proportional feedback algorithm is accurate but slow, whereas relay feedback is quick but approximate. Usually, approximate values are fine. (The estimate of K_c was off by 7% in this case.)

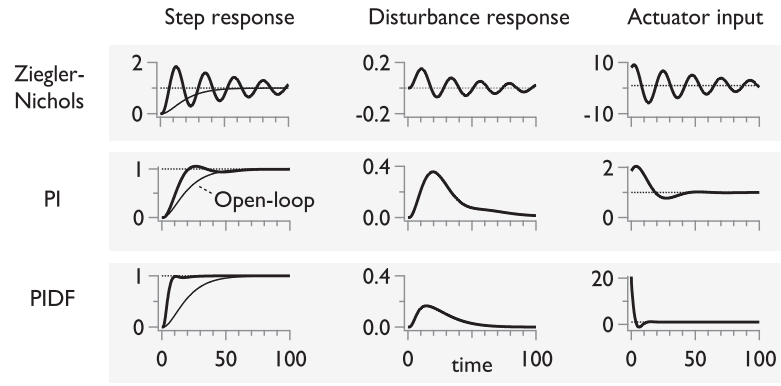
c. Using $K_\mathrm{c} = 20$ and $T_\mathrm{c} = 15$, the Ziegler-Nichols parameters are $K_\mathrm{p} = 8$ and $T_\mathrm{i} = 11.2$ (equivalently, $K_\mathrm{i} = 0.71$).

Playing around, I found $K_\mathrm{p} = 1.86$ and $K_\mathrm{i} = 0.08$, much lower gains! Notice the two time scales in the relaxation of the disturbance to equilibrium. One is from the proportional response, the other from the integral. We would want to increase the integral term to make disturbances recover more quickly, but we cannot because the controller would have too much lag, destabilizing the proportional part of the response. To do better, we need a controller with more parameters and a more complex frequency response.

For PIDF, I found $K_p = 4$, $K_i = 0.2$, $K_d = 20$, and $t_f = 1.2$, where the form is

$$K(s) = K_p + \frac{K_i}{s} + \frac{K_d s}{t_f s + 1}\,.$$

We see how adding more parameters improves the results markedly, albeit at the cost of increased control input magnitude. Note that the filtering term is needed to make the input requirements more reasonable. If the command is a step function, as here, then differentiating the response creates a delta function spike in the required input. Filtering softens this. See the collected plots for all the cases below.



3.23 Analysis of a two-sensor system. For the split-PI example of Section 3.8.1:

a. Derive the closed-loop transfer functions of Eqs. (3.81), (3.82), and (3.83).
b. Reproduce the step responses for all three cases.
c. For $K_p = 5$, why is there no instability in Case 1 for any K_i? For Cases 2 and 3, find the instability threshold K_i^* and oscillation frequency ω^* at onset, again fixing $K_p = 5$. Do this part analytically or numerically.

Solution.

a. *Transfer functions*. The first two follow the simpler rule,

$$\text{Transfer function} = \frac{\text{Direct}}{1 + \text{Loop}}\,,$$

discussed previously. We need to derive the third relation, where

$$u = K_{pi} r - K_p y_1 - \frac{K_i}{s} y_2\,,$$

with $K_{pi} = K_p + K_i/s$. We write $y_1 = G_1 u$ and $y_2 = G_2 y_1$, so that

$$\begin{aligned} y_2 &= G_1 G_2 K_{pi} r - K_p G_1 G_2 y_1 - \frac{K_i}{s} G_1 G_2 y_2 + G_1 G_2 d \\ \Longrightarrow \; &= \left(\frac{K_{pi} G_1 G_2}{K_p G_1 + \frac{K_i}{s} G_1 G_2} \right) r + \left(\frac{G_1 G_2}{K_p G_1 + \frac{K_i}{s} G_1 G_2} \right) d\,, \end{aligned}$$

where we substitute $G_2 y_1 = y_2$ to put the first line entirely in terms of y_2.

b. *Step responses*. With proper software, you just need to input the transfer functions and ask for the step response directly.

c. *Instability thresholds*. For Case 1, the system and controller are both first order, so that the closed-loop system is second order. Since a second-order system achieves a phase lag of $-180°$ only at infinite frequency, it can never be unstable.

Cases 2 and 3 are straightforward to solve in a symbolic-manipulation program. Requiring the real and imaginary parts of the denominator to vanish, we find

i. Case 2: $K_i^* = 18$ and $\omega^* = \frac{3}{2} = 1.5$.

ii. Case 3: $K_i^* = 63$ and $\omega^* = \sqrt{\frac{7}{2}} \approx 1.87$.

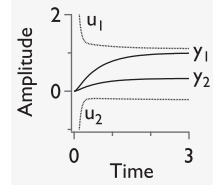
We can apply significantly higher gains in Case 3 before instability sets in.

3.24 Fixing the shower. Slightly altering the "shower" transfer function $\boldsymbol{G}_{sh}$ given in Eq. (3.101) turns an impossible control problem into a straightforward one. Consider $\boldsymbol{G}'_{sh} = \left(\begin{smallmatrix} 1/(1+s) & 1/(2+s) \\ 1/(2+s) & 1/(1+s) \end{smallmatrix}\right)$, which differs from the expression in Eq. (3.101) in that the DC cross gains equal $\frac{1}{2}$ rather than 1. The control goal is a good step response (e.g., to step rapidly the outputs from 0 to $y_1 = 1$ and $y_2 = \frac{1}{3}$, using inputs $(u_1, u_2) \in (-10, 10)$).

a. Make a singular value plot. Show that the condition number still diverges at $s \to \infty$ but not at $s \to 0$. Why is a high-frequency divergence allowable but not a low-frequency one? Can your computer program compute a step response in the time domain? Find a fix that leads asymptotically to the correct DC outputs.

b. Try to improve the controller by canceling a pole or zero in the inverse. Remember "tweaks" such as lag and lead.

At right are step responses for the modified-shower problem. Reference signals r_1 and r_2 are given steps of $(1, \frac{1}{3})$. Asymptotically, the outputs $y_1 \to r_1$ and $y_2 \to r_2$.

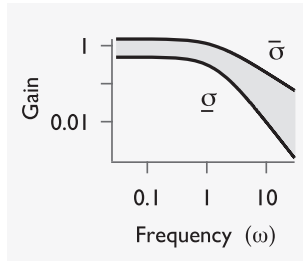


Solution.

a. Singular value plot. The intention was to generate a graph, as shown at right. It turns out that the singular values have a nice analytical form (as given by Mathematica):

$$\overline{\sigma}(\omega) = \sqrt{\frac{9 + 4\omega^2}{4 + 5\omega^2 + \omega^4}}, \qquad \underline{\sigma}(\omega) = \sqrt{14 + 5\omega^2 + \omega^4}$$

From the graph below and from the explicit algebraic expressions, we see that the condition number $\gamma = \overline{\sigma}/\underline{\sigma}$ diverges $\sim \omega$ at high frequencies. A divergence at high frequencies is acceptable, as any feedback system will have finite closed-loop bandwidth. By contrast, a divergence of the condition number at low frequencies can be disastrous, as we saw.



b. *Naive try*. The inverse to G_{sh} is

$$G_{\text{sh}}^{-1} = \begin{pmatrix} \frac{(s+1)(s+2)^2}{2s+3} & -\frac{(s+1)^2(s+2)}{2s+3} \\ -\frac{(s+1)^2(s+2)}{2s+3} & \frac{(s+1)(s+2)^2}{2s+3} \end{pmatrix},$$

whose elements $\sim \omega^2$ at high frequencies. To be realizable, a controller must not diverge. Thus, our controller must have a denominator of order at least 2 to be realizable. In other words,

$$\boldsymbol{K}(s) = \frac{K_{\text{i}}}{s} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \boldsymbol{G}'^{-1}_{\text{sh}}(s)$$

does not work because K_{i}/s reduces the order by only one. Thus, we need to add another power, for example K_{i}/s^2 or we might try to cancel the common factor of $2 + s$. For example, we can try $K_{\text{i}}/[s(s + 2)]$.

Remember that the above statements hold even though the whole loop $\boldsymbol{K}(s)\boldsymbol{G}(s)$ would be ok (since $\boldsymbol{G}$ cancels its inverse). But the controller needs to generate the signal before "cancellation" can occur.

c. *Best effort*. My best result was using a lag compensator to tweak. The transfer function

$$\boldsymbol{K}(s) = \left(\frac{1 + s/2}{1 + s/20}\right)\left(\frac{3}{s(s + 2)}\right)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \boldsymbol{G}_{\text{sh}}^{-1}$$

gave the step response in the problem, above.

Time-Domain Control

4

Problems

4.1 **Controllability of nearly identical systems**. Consider two first-order systems with relaxation rates $\lambda_1 = 1$ and $\lambda_2 = 2$ that are driven by identical inputs (Eq. 4.7). Find an input $u(t)$ that takes the system from an initial state $\boldsymbol{x}_0 = \binom{0}{0}$ to a final state $\boldsymbol{x}_\tau = \binom{1}{1}$, for $\tau = 1$. Plot your solution. Hint: Try a step function with two parameters.

Solution.

We need to find a solution $x_1(1) = x_2(1)$ for

$$\begin{pmatrix}\dot{x_1}\\ \dot{x_2}\end{pmatrix} = \begin{pmatrix}-\lambda_1 & 0\\ 0 & -\lambda_2\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix} + \begin{pmatrix}1\\ 1\end{pmatrix}u(t) \quad \Longrightarrow \quad \begin{aligned}\dot{x_1} &= -\lambda_1 x_1 + u\\ \dot{x_2} &= -\lambda_2 x_2 + u\end{aligned},$$

with initial state ($t = 0$) and final states ($\tau = 1$) given by

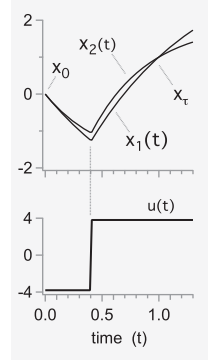
$$x_0 = \begin{pmatrix}x_1(0)\\ x_2(0)\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix}, \qquad x_\tau = \begin{pmatrix}x_1(1)\\ x_2(1)\end{pmatrix} = \begin{pmatrix}1\\ 1\end{pmatrix}.$$

There are an infinite number of ways to do this. A basic requirement is that there be two free parameters, since we are trying to fix two conditions ($x_1 = x_2$ at $\tau = 1$). One simple route is to use piecewise constant $u(t)$ functions. Let us try the simple form

$$u(t) = \begin{cases}-u_0 & 0 < t < \tau_0\\ +u_0 & \tau_0 < t < 1\end{cases}$$

We can explicitly integrate the x_1 and x_2 equations (they are the same, substituting 2 for 1, etc.) Denoting $x_{1,2}$ by $x(t)$ and $\lambda_{1,2}$ by λ, we get

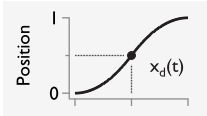
$$x(t) = \begin{cases}\frac{-u_0}{\lambda}\left(1 - e^{-\lambda t}\right) & 0 < t < \tau_0\\ \frac{u_0}{\lambda}\left(1 - e^{-\lambda\tau_0}\right)\mathrm{e}^{-\lambda(t-\tau_0)} - (1 - e^{-\lambda(t-\tau_0)}) & \tau_0 < t < 1\,.\end{cases}$$

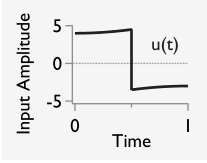


After a certain amount of playing around, I found the solution illustrated at right, where $u_0 = 3.8$ and $\tau_0 = 0.405$ works for $\lambda_1 = 1$ and $\lambda_2 = 2$. Again, we emphasize that we use the same $u(t)$ for both $x_1(t)$ and $x_2(t)$. It is possible to find explicit algebraic solutions for u_0 and τ_0 in terms of $\lambda_{1,2}$, etc., but the main point here is to understand intuitively how a solution works and why it

is possible. Similarly, although I found my solution by an iteration of varying coefficients and looking at the results, you could easily make such a procedure more systematic by formulating it as a nonlinear system of equations and using a general numerical solver to find its roots.

Finally, remember that you may not be able to generate the required values of $u(t)$. Any real input has saturation limits (cannot go outside a given range). This is a kind of nonlinearity and puts us outside the framework of linear equations.





4.2 Prescribing a path in state space. A system may be controllable, but that does not mean we can make it follow a desired trajectory $\boldsymbol{x}(t)$ in state space.

a. Show that you cannot prescribe a path for the system defined in Example 4.4.
b. Consider the undamped oscillator with torque control, $\ddot{x} + x = u$. Following Example 4.1, find and plot $u(t)$ that leads to the desired trajectory $x_\mathrm{d}(t) = 2(t/\tau)^2 - [1 - 2(t/\tau)^2]\theta[2(t/\tau) - 1]$, which is sketched at left. Verify by integrating the differential equation numerically that your $u(t)$ produces the desired $x_\mathrm{d}(t)$.
c. Comment on the required control effort for $\tau \to 0$, with fixed x_τ and ω_0.
d. Can you give any intuition about why the second case works but not the first?

Solution.

a. The dynamical system from Example 4.4 is

$$\begin{pmatrix}\dot{x}_1\\ \dot{x}_2\end{pmatrix} = \begin{pmatrix}-\lambda_1 & 0\\ 0 & -\lambda_2\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix} + \begin{pmatrix}1\\ 1\end{pmatrix}u(t) \quad \Longrightarrow \quad \begin{aligned}\dot{x}_1 &= -\lambda_1 x_1 + u\\ \dot{x}_2 &= -\lambda_2 x_2 + u\end{aligned},$$

If did not care about $x_2(t)$, we could specify a desired $x_1(t)$ and invert to find $u(t) = \dot{x}_1 + \lambda_1 x_1$. We could also do the same for a desired $x_2(t)$ if we did not care about $x_1(t)$. Clearly, we cannot specify the two functions independently! Again, keep in mind that the system *is* controllable and that Problem 4.1 showed that we could specify that the system reach an arbitrary point $(x_1, x_2)^\mathsf{T}$ at an arbitrary time τ. That is much less demanding than asking it

b. In other cases, we can indeed invert the dynamics. Here, the first derivative of the prescribed path is

$$\dot{x}_\mathrm{d} = \begin{cases}\frac{4t}{\tau^2} & 0 < t < \frac{\tau}{2}\\ \frac{4}{\tau}\left(1 - \frac{t}{\tau}\right) & \frac{\tau}{2} < t < \tau\end{cases}.$$

The second derivative is

$$\ddot{x}_\mathrm{d} = \begin{cases}\frac{4}{\tau^2} & 0 < t < \frac{\tau}{2}\\ -\frac{4}{\tau^2} & \frac{\tau}{2} < t < \tau\end{cases}.$$

Then, $u = \ddot{x}_\mathrm{d} + x_\mathrm{d} =$

$$u(t) = \begin{cases}\frac{4}{\tau^2} + 2\left(\frac{t}{\tau}\right)^2 & 0 < t < \frac{\tau}{2}\\ -\frac{4}{\tau^2} + \left[1 - 2\left(1 - \frac{t}{\tau}\right)^2\right] & \frac{\tau}{2} < t < \tau\end{cases},$$

For $x_\tau = \tau = 1$, we plot

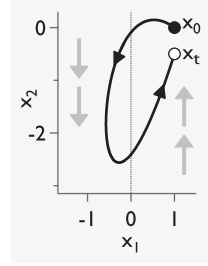
$$u(t) = \begin{cases} 4 + 2t^2 & 0 < t < \frac{1}{2} \\ -3 - 2(1-t)^2 & \frac{1}{2} < t < 1 \end{cases},$$

Numerically integrating the second-order equations with the above $u(t)$ as a forcing term reproduces the solution exactly.

c. The amplitude of control required goes as $1/\tau^2$, which diverges as $\tau \to 0$. Because the range of u is always limited, we cannot ask for trajectories that vary too rapidly.

d. An intuitive explanation is that in the first case, we have two first-order equations of the form $\dot{x} = -\lambda x + u(t)$ that have a single control. The two elements of the state vector can be specified independently as $x_1(t)$ and $x_2(t)$, but we have only one control function, $u(t)$. In the second case, $\ddot{x} + x = u$, we specify $x_1(t) = x(t)$ but then we have no choice about the other component of the state vector, $x_2(t) = \dot{x}(t)$ is determined once $x_1(t)$ is given. Thus, we really have only one independent function $x(t)$, which maps to a single function $u(t)$.
There is a general theory, *differential flatness*, that examines conditions for when the nonlinear system $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u})$ and $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{u})$ may be "inverted" to give $\boldsymbol{u}$ as a set of simple derivatives. (Notice that we do not have to solve any integrals in the harmonic-oscillator case.) For a very brief description, see Åström and Murray (2008) and for a more mathematical explanation, drawing on the differential-geometry formulation of Section 11.1.5, see Lévine (2009).

4.3 **Nonlocal control**. If there are fewer control nodes than state variables – and there usually are – then moving the system from an initial state $\boldsymbol{x}_0$ to a final state $\boldsymbol{x}_\tau$ may require a finite-length trajectory, even when $|\boldsymbol{x}_\tau - \boldsymbol{x}_0| = \varepsilon$, and $\varepsilon \to 0$. To illustrate this *nonlocality* of control trajectories, see the dynamics at right, which depict a kind of shear "flow" that is directed *down* for $x_1 < 0$ and *up* for $x_1 > 0$.



a. At right, $\dot{x}_1 = u$, $\dot{x}_2 = x_1 + u$. Write these equations in the form $\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}u$. Calculate analytically $e^{\boldsymbol{A}t}$, $e^{\boldsymbol{A}t}\,\boldsymbol{B}$, and the Gramian $\boldsymbol{P}(\tau) \equiv \int_0^\tau dt\, e^{\boldsymbol{A}t}\,\boldsymbol{B}\,\boldsymbol{B}^\mathsf{T} e^{\boldsymbol{A}^\mathsf{T}t}$.

b. Show, by substitution into the general solution, $\boldsymbol{x}(t) = e^{\boldsymbol{A}t}\boldsymbol{x}_0 + \int_0^t dt'\, e^{\boldsymbol{A}(\tau-t')}\,\boldsymbol{B}u(t')$, that the control $u(t) = \boldsymbol{B}^\mathsf{T} e^{\boldsymbol{A}^\mathsf{T}(\tau-t)}\,\boldsymbol{P}^{-1}(\tau)\Delta\boldsymbol{x}$ brings the initial state at $t = 0$, $\boldsymbol{x}_0$, to the final state at τ of $\boldsymbol{x}_\tau$. Here, $\Delta\boldsymbol{x} \equiv \boldsymbol{x}_\tau - e^{\boldsymbol{A}\tau}\boldsymbol{x}_0$. See also Problem 7.9.

c. Show the above formula gives $u(t) = 0.126(t - 5)$ and moves an initial state $\binom{1}{0} \to \binom{1}{-\varepsilon}$, with $\varepsilon = 0.5$ in a time $\tau = 10$. Reproduce the plot at right.

This problem is adapted from Sun and Motter (2013).

Solution.

a. The vector form of the equations is

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} u(t)\,.$$

Then

$$\boldsymbol{A}^2 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \boldsymbol{0} \quad \Longrightarrow \quad \mathrm{e}^{\boldsymbol{A}t} = \mathbb{I} + \boldsymbol{A}t = \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix}$$

Thus,

$$\mathrm{e}^{\boldsymbol{A}t}\,\boldsymbol{B} = \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1+t \end{pmatrix}$$

The outer product $\mathrm{e}^{\boldsymbol{A}t}\,\boldsymbol{B}\,\boldsymbol{B}^{\mathsf{T}}\,\mathrm{e}^{\boldsymbol{A}^{\mathsf{T}}t}$ is then

$$\begin{pmatrix} 1 \\ 1+t \end{pmatrix}\begin{pmatrix} 1 & 1+t \end{pmatrix} = \begin{pmatrix} 1 & 1+t \\ 1+t & (1+t)^2 \end{pmatrix}$$

Integrating gives the Gramian:

$$\boldsymbol{P}(\tau) = \int_0^\tau \mathrm{d}t\,\mathrm{e}^{\boldsymbol{A}t}\,\boldsymbol{B}\,\boldsymbol{B}^{\mathsf{T}}\,\mathrm{e}^{\boldsymbol{A}^{\mathsf{T}}t} = \begin{pmatrix} \tau & \tau + \frac{1}{2}\tau^2 \\ \tau + \frac{1}{2}\tau^2 & \tau + \tau^2 + \frac{1}{3}\tau^3 \end{pmatrix}.$$

b. The solution $\boldsymbol{x}(t)$ is given by

$$\boldsymbol{x}(t) = \mathrm{e}^{\boldsymbol{A}t}\,\boldsymbol{x}_0 + \int_0^t \mathrm{d}t'\,\mathrm{e}^{\boldsymbol{A}(\tau-t')}\,\boldsymbol{B}u(t').$$

Defining $\Delta\boldsymbol{x} \equiv \boldsymbol{x}_\tau - \mathrm{e}^{\boldsymbol{A}\tau}\,\boldsymbol{x}_0$, we have

$$\Delta\boldsymbol{x} = \int_0^\tau \mathrm{d}t'\,\mathrm{e}^{\boldsymbol{A}(\tau-t')}\,\boldsymbol{B}u(t').$$

Then, with $u(t) = \boldsymbol{B}^{\mathsf{T}}\,\mathrm{e}^{\boldsymbol{A}^{\mathsf{T}}(\tau-t)}\,\boldsymbol{P}^{-1}(\tau)\Delta\boldsymbol{x}$, we have

$$\begin{aligned}
\boldsymbol{x}(\tau) &= \mathrm{e}^{\boldsymbol{A}\tau}\,\boldsymbol{x}_0 + \int_0^\tau \mathrm{d}t'\,\mathrm{e}^{\boldsymbol{A}(\tau-t')}\,\boldsymbol{B}\,\underbrace{\boldsymbol{B}^{\mathsf{T}}\,\mathrm{e}^{\boldsymbol{A}^{\mathsf{T}}(\tau-t')}\,\boldsymbol{P}^{-1}(\tau)\Delta\boldsymbol{x}}_{u(t')} \\
&= \mathrm{e}^{\boldsymbol{A}\tau}\,\boldsymbol{x}_0 + \int_0^\tau \mathrm{d}t'\,\mathrm{e}^{\boldsymbol{A}(\tau-t')}\,\boldsymbol{B}\,\boldsymbol{B}^{\mathsf{T}}\,\mathrm{e}^{\boldsymbol{A}^{\mathsf{T}}(\tau-t')}\left(\boldsymbol{P}^{-1}(\tau)\Delta\boldsymbol{x}\right) \\
&= \mathrm{e}^{\boldsymbol{A}\tau}\,\boldsymbol{x}_0 + \boldsymbol{P}(\tau)\boldsymbol{P}^{-1}(\tau)\Delta\boldsymbol{x} \\
&= \mathrm{e}^{\boldsymbol{A}\tau}\,\boldsymbol{x}_0 + \boldsymbol{x}_\tau - \mathrm{e}^{\boldsymbol{A}\tau}\,\boldsymbol{x}_0 \\
&= \boldsymbol{x}_\tau\,.
\end{aligned}$$

Later, we will see that this choice of $u(t)$ minimizes the *control effort*,

$$\mathcal{E} \equiv \int_0^\tau \mathrm{d}t\,u^2(t).$$

c. Evaluate numerically. Note that since $\dot{x}_1 = u$, any trajectory with $x_1(\tau) = x_1(0)$ satisfies $\int_0^\tau \mathrm{d}t\,u(t) = 0$.

4.4 **Pole-zero cancellation**. In Example 4.9, we explored how the different input-output connections between two transfer functions can lead to different issues (loss of controllability vs. loss of observability). Verify the following:

a. Check the state-space forms for G_1 and G_2.
b. Show that the 12 and 21 series connections lead to different 3d systems.

c. Show that if you start from either the 12 or the 21 system, you derive the same transfer function (= G_1G_2 or G_2G_1).
d. Write down the observability and controllability matrices for the 12 and 21 systems, and verify that 12 is uncontrollable and 21 unobservable.

Solution.

You should use a symbolic-manipulation program to do this problem!

a. Use the controller-canonical form, Eq. (2.59).
b. Straightforward.
c. Using $G(s) = \boldsymbol{C}(s\mathbb{I} - \boldsymbol{A})^{-1}\boldsymbol{B}$ gives

$$G(s) = \frac{1}{(p-s)^2},$$

in both cases.
d. *Controllability and Observability*.
For the 12 connection:

$$\boldsymbol{W}_{\mathrm{c}} = \begin{pmatrix} 0 & 1 & 2p \\ 1 & 2p & 3p^2 \\ 0 & 1 & 2p \end{pmatrix}, \qquad \boldsymbol{W}_{\mathrm{o}} = \begin{pmatrix} 0 & 0 & 1 \\ -\alpha & 1 & \alpha \\ -p^2-\alpha^2 & 2p & \alpha^2 \end{pmatrix}$$

Det $\boldsymbol{W}_{\mathrm{c}} = 0$ and det $\boldsymbol{W}_{\mathrm{o}} = (p-\alpha)^2$.
For the 21 connection:

$$\boldsymbol{W}_{\mathrm{c}} = \begin{pmatrix} 1 & \alpha & \alpha^2 \\ 0 & 0 & 1 \\ 0 & 1 & 2p+\alpha \end{pmatrix}, \qquad \boldsymbol{W}_{\mathrm{o}} = \begin{pmatrix} 0 & -\alpha & 1 \\ 1 & -p^2 & 2p-\alpha \\ 2p & -p^2(2p-\alpha) & -p^2+2p(2p-\alpha) \end{pmatrix}$$

Det $\boldsymbol{W}_{\mathrm{c}} = -1$ and det $\boldsymbol{W}_{\mathrm{o}} = 0$.

4.5 **Zeros with more actuators and sensors**. In Section 4.1.3, we saw that if the state vector is n dimensional and there are either n independent inputs or outputs, the transfer function of the enlarged system cannot have a zero. Here, we verify this in a simple example. Consider the transfer function G with a single RHP zero,

$$G(s) = \frac{s-2}{(s-1)^2} \quad \Longleftrightarrow \quad \boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ -1 & 2 \end{pmatrix} \quad \boldsymbol{B} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \boldsymbol{C} = \begin{pmatrix} -2 & 1 \end{pmatrix}.$$

Now consider a second input or output, by taking $\boldsymbol{B}' = \left(\begin{smallmatrix} b & 0 \\ 0 & 1 \end{smallmatrix}\right)$ or $\boldsymbol{C}' = \left(\begin{smallmatrix} -2 & 1 \\ 0 & c \end{smallmatrix}\right)$.

a. Keeping the original $\boldsymbol{A}$ and $\boldsymbol{C}$, consider the new inputs $\boldsymbol{B}'$ and show that the 1×2 transfer function matrix has no zeros. Recall that in a MIMO transfer matrix, a zero is a value of s for which the transfer function matrix loses rank.
b. Repeat the calculation for the case where you keep $\boldsymbol{A}$ and $\boldsymbol{B}$ and use $\boldsymbol{C}'$.
c. Why do the above conclusions become invalid if b or $c = 0$?

Solution.

a. Adding an input changes the input coupling matrix and leads to a 2×1 transfer function matrix:

$$\boldsymbol{B}' = \begin{pmatrix} b & 0 \\ 0 & 1 \end{pmatrix}, \quad \Longrightarrow \quad G(s) = \begin{pmatrix} b\frac{3-2s}{(s-1)^2} & \frac{s-2}{(s-1)^2} \end{pmatrix},$$

We see that, for $b \neq 0$, no value of s makes G lose rank (go from 1 to 0). This contrasts with the situation with one input, where the zero at $s = 2$ makes the single transfer function vanish (also rank 1 to 0).

b. Adding an output changes the output coupling matrix and leads to a 1×2 transfer function matrix:

$$\boldsymbol{C}' = \begin{pmatrix} -2 & 1 \\ 0 & c \end{pmatrix}, \quad \Longrightarrow \quad G(s) = \begin{pmatrix} \frac{s-2}{(s-1)^2} \\ c\frac{s}{(s-1)^2} \end{pmatrix},$$

We see that, for $c \neq 0$, no value of s makes G lose rank (go from 1 to 0). This contrasts with the situation with one input, where there is a zero at $s = 2$ makes the single transfer function vanish (also rank 1 to 0).

c. When b is zero, one component of the transfer-function matrix is always zero, which causes a MIMO zero when the other component vanishes at $s = 2$. The argument is the same for $c = 0$.

4.6 **Feedforward gain for constant output**. For the SISO system $\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}u$, $y = \boldsymbol{C}\boldsymbol{x}$, show that choosing $u = -\boldsymbol{K}\boldsymbol{x} + k_{\mathrm{r}}r$ leads to $y = r$ if $k_{\mathrm{r}} = -[\boldsymbol{C}(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K})^{-1}\boldsymbol{B}]^{-1}$.

Solution.

$$\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}u\,, \qquad u = -\boldsymbol{K}\boldsymbol{x} + k_{\mathrm{r}}r\,, \qquad y = \boldsymbol{C}\boldsymbol{x}$$

Substituting and looking at the steady-state solution, we have

$$\boldsymbol{0} = (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K})\boldsymbol{x} + k_{\mathrm{r}}r \quad \Longrightarrow \quad \boldsymbol{x} = -(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K})^{-1}k_{\mathrm{r}}r \quad \Longrightarrow \quad y = -\boldsymbol{C}(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K})^{-1}k_{\mathrm{r}}r$$

Then, to make $y = r$ we set k_{r} to

$$k_{\mathrm{r}} = \frac{-1}{\boldsymbol{C}(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K})^{-1}\boldsymbol{B}}\,.$$

4.7 **Noise-tracking tradeoffs for observers**. If observer gains are too low, the observer states will not track the state vector well. If the gains are too high, too much measurement noise will be injected into the system. Here, we show this tradeoff explicitly.

a. Add measurement noise $\xi(t)$ to the observer equation for the dynamical system:

$$\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}u, \quad y = \boldsymbol{C}\boldsymbol{x}$$
$$\dot{\hat{\boldsymbol{x}}} = \boldsymbol{A}\hat{\boldsymbol{x}} + \boldsymbol{B}u + \boldsymbol{L}[y(t) + \xi(t) - \hat{y}(t)]\,.$$

Take the Laplace transform of the error dynamics ($\boldsymbol{e} = \boldsymbol{x} - \hat{\boldsymbol{x}}$), keeping the initial value term to give $\boldsymbol{e}(s)$ as the sum of two terms, one proportional to

$\boldsymbol{e}(t = 0) \equiv e_0$ and one proportional to $\xi(s)$. Argue that large values of $\boldsymbol{L}$ make the initial-value term decay quickly but will simultaneously keep the noise term large.

b. Specialize to the first-order system $\dot{x} = -x$ and $y = x + \xi \equiv x + \xi_0 \sin t$. The sinusoidal "noise" is a simple stand-in for stochastic noise, which would be the sum of sines of all frequencies and with random phases. Solve for the Laplace transform of the observer error, $e(s)$, in terms of the initial error e_0 and ξ_0.

The "best" value of observer gain ℓ balances the convergence rate of estimator errors against noise injection. One missing ingredient is the notion that disturbances continually inject new state-estimation errors, which the observer must try to remove. Here, an initial error e_0 will decay away for all values of ℓ so that an observer would not be necessary for long-time estimation. Continuously injecting new disturbances into the dynamics highlights the role of ℓ in balancing the rate that the observer removes disturbances against noise injection. See the *Kalman filter* in Chapter 8.

Solution.

a. *Transfer function*. The error dynamics for $\boldsymbol{e} = \boldsymbol{x} - \hat{\boldsymbol{x}}$ are given by

$$\dot{\boldsymbol{e}} = (\boldsymbol{A} - \boldsymbol{LC})\boldsymbol{e} - \boldsymbol{L}\xi(t)\,.$$

Now, we Laplace transform this equation, keeping the initial-value term:

$$\boldsymbol{e}(s) = (s\mathbb{I} - \boldsymbol{A} + \boldsymbol{LC})^{-1}\boldsymbol{e}(t = 0) - (s\mathbb{I} - \boldsymbol{A} + \boldsymbol{LC})^{-1}\boldsymbol{L}\xi(s)\,.$$

Thus, if we choose $\boldsymbol{L}$ so that the eigenvalues of $\boldsymbol{A} - \boldsymbol{LC}$ have large decay constants, we will have large values for the matrix $\boldsymbol{L}$, which will enhance the right-hand term. (In one dimension, the factors of L would cancel out, but in higher dimensions, the different terms are likely to generate a larger factor in front of $\xi(s)$.)

b. *First-order example*. In the example, $A = -1$, $C = 1$, and the observer equation is

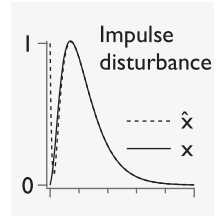
$$\dot{\hat{x}} = -\hat{x} + \ell(y - \hat{y}) = -\hat{x} + \ell(x + \xi_0 \sin t - \hat{x})\,.$$

The error dynamics for $e(t) = x(t) - \hat{x}(t)$ is then

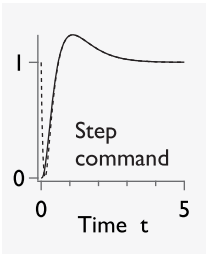
$$\dot{e} = -e - \ell(e + \xi_0 \sin t)$$

$$se(s) - e_0 = -(1 + \ell)e(s) - \frac{\ell\xi_0}{1 + s^2}$$

$$e(s) = -\frac{e_0}{s + 1 + \ell} - \frac{\ell\xi_0}{(1 + s^2)(s + 1 + \ell)}\,.$$



Again, we see qualitatively that for small ℓ, initial errors decay slowly, but larger ℓ increase the effects of measurement noise, making $e \to -\xi_0$.



4.8 Observer-based feedback for the harmonic oscillator. In Example 4.13, we set up a structure for an observer-based control of a harmonic oscillator.

a. Write down the coupled system plus observer. Include a feedforward gain to make a step command go to the right value. Design the controller to have poles $(-2,-2)$ and the observer to have poles $(-10,-10)$. Give numerical values for the controller ($\boldsymbol{K}$), observer ($\boldsymbol{L}$), and feedforward (k_r) gains.

b. Reproduce the numerical responses at left for an impulse disturbance and step command. Use discordant initial conditions: $\hat{\boldsymbol{x}}(0) = \binom{1}{0}$, but $\boldsymbol{x}(0) = \binom{0}{0}$

Solution.

a. We reproduce the state-observer equations from Eq. (4.68) for the combined state vector $\binom{\boldsymbol{x}}{\hat{\boldsymbol{x}}}$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\boldsymbol{x}\\ \hat{\boldsymbol{x}}\end{pmatrix} = \begin{pmatrix}\boldsymbol{A} & -\boldsymbol{BK}\\ \boldsymbol{LC} & (\boldsymbol{A}-\boldsymbol{BK}-\boldsymbol{LC})\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ \hat{\boldsymbol{x}}\end{pmatrix} + \begin{pmatrix}\boldsymbol{B}k_\mathrm{r}\\ \boldsymbol{B}k_\mathrm{r}\end{pmatrix} r + \begin{pmatrix}\boldsymbol{B}\\ \boldsymbol{0}\end{pmatrix} d\,, \quad y = \begin{pmatrix}\boldsymbol{C} & \boldsymbol{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ \hat{\boldsymbol{x}}\end{pmatrix}.$$

In the present case, this gives the four-dimensional dynamics,

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}x_1\\ x_2\\ \hat{x}_1\\ \hat{x}_2\end{pmatrix} = \begin{pmatrix}0 & 1 & 0 & 0\\ -1 & 0 & -k_1 & -k_2\\ \ell_1 & 0 & -\ell_1 & 1\\ \ell_2 & 0 & (-1-k_1-\ell_2) & -k_2\end{pmatrix}\begin{pmatrix}x_1\\ x_2\\ \hat{x}_1\\ \hat{x}_2\end{pmatrix} + \begin{pmatrix}0 & 0\\ k_\mathrm{r} & 1\\ 0 & 0\\ k_\mathrm{r} & 0\end{pmatrix}\begin{pmatrix}r\\ d\end{pmatrix}. \tag{4.88}$$

with

$$\boldsymbol{C}' = \begin{pmatrix}1 & 0 & 0 & 0\end{pmatrix}. \tag{4.89}$$

For controller poles at $(-2,-2)$, observer poles at $(-10,-10)$, we find, using the built-in Mathematica routines for pole placement,

$$\boldsymbol{K} = \begin{pmatrix}3 & 4\end{pmatrix}, \quad \boldsymbol{L} = \begin{pmatrix}20\\ 99\end{pmatrix}, \quad k_\mathrm{r} = 100\,.$$

b. The plots are for $\boldsymbol{u} = \binom{0}{\delta(t)}$ and $\boldsymbol{u} = \binom{\Theta(t)}{0}$.

4.9 Stabilizing an inverted pendulum forced by a torque.

a. For small displacements about the vertical, show that the scaled equations of motion have $\boldsymbol{A} = \left(\begin{smallmatrix}0 & 1\\ 1 & 0\end{smallmatrix}\right)$ and $\boldsymbol{B} = \binom{0}{1}$.

b. Check that $\{\boldsymbol{A}, \boldsymbol{B}\}$ is controllable.

c. Assuming a known state vector, design a feedback law $u = -\boldsymbol{Kx}$ to stabilize the vertical fixed point with eigenvalues $(-1,-1)$. Find the gains K_1 and K_2.

d. If you measure only the position, $\boldsymbol{C} = (1\ 0)$. Show that $\{\boldsymbol{A}, \boldsymbol{C}\}$ is observable.

e. Design an observer with dynamics $(-2,-2)$.

f. Design a combined observer-controller that regulates the system about $\binom{0}{0}$. This *regulator* is a four-dimensional system. Find the state-space matrices $\{\boldsymbol{A}_\mathrm{oc}, \boldsymbol{B}_\mathrm{oc}, \boldsymbol{C}_\mathrm{oc}\}$ using an input torque disturbance $d(t)$ as input and the angle $\theta(t)$ as output.

g. Find the transfer function $K_{ob}(s)$ for output control ($y \to u$).
h. Plot $\{\theta, \hat{\theta}\}$ and $\{\dot{\theta}, \hat{\dot{\theta}}\}$ for an impulse input disturbance [$u(t) = \delta(t)$].

Solution.

a. *Equations of motion for perturbations about vertical.* In Ch. 2, we showed that the scaled equations of motion are

$$\ddot{\theta} + \sin\theta = u(t)\,,$$

Setting $\theta(t) = \pi + x_1(t)$ and $\dot{\theta} = x_2$ and expanding $\sin\theta$ about π, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}}_{\boldsymbol{A}} + \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{\boldsymbol{B}} u(t)\,,$$

b. *Controllability*.

$$\boldsymbol{W}_c = \begin{pmatrix} \boldsymbol{B} & \boldsymbol{AB} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Since $\det \boldsymbol{W}_c \neq 0$, the matrix is full rank and thus $\{\boldsymbol{A}, \boldsymbol{B}\}$ is controllable.

c. *Full-state control.* The matrix $\boldsymbol{A} - \boldsymbol{BK}$ is given by

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix}\begin{pmatrix} K_1 & K_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 - K_1 & -K_2 \end{pmatrix},$$

which has eigenvalues that are the roots of

$$s^2 + K_2 s + K_1 - 1 = 0\,.$$

We want to choose K_1 and K_2 so that the eigenvalues are $(-1, -1)$. Expanding the desired characteristic equation of $(s+1)^2 = 0$, we have

$$s^2 + 2s + 1 = 0\,,$$

and matching coefficients gives $K_1 = K_2 = 2$.

d. *Observability*.

$$\boldsymbol{W}_0 = \begin{pmatrix} \boldsymbol{C} \\ \boldsymbol{CA} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

which has det = 1 and thus is full rank. Hence, $\{\boldsymbol{A}, \boldsymbol{C}\}$ is observable.

e. *Observer*. The matrix $\boldsymbol{A} - \boldsymbol{LC}$ is given by

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} - \begin{pmatrix} \ell_1 \\ \ell_2 \end{pmatrix}\begin{pmatrix} 1 & 0 \end{pmatrix} = \begin{pmatrix} -\ell_1 & 1 \\ 1 - \ell_2 & 0 \end{pmatrix},$$

which has eigenvalues that are the roots of

$$s^2 + \ell_1 s + \ell_2 - 1 = 0\,.$$

We want observer roots $(-2, -2)$ and thus to match

$$s^2 + 4s + 4 = 0\,,$$

which gives $\ell_1 = 4$ and $\ell_2 = 5$.

f. *Observer-controller*. We have

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \boldsymbol{B} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad \boldsymbol{C} = \begin{pmatrix} 1 & 0 \end{pmatrix}, \qquad \boldsymbol{L} = \begin{pmatrix} 4 \\ 5 \end{pmatrix}, \qquad \boldsymbol{K} = \begin{pmatrix} 2 & 2 \end{pmatrix}.$$

From Eq. (4.68), the combined observer-controller has a 4×4 closed-loop dynamical matrix given by

$$\boldsymbol{A}_{\text{oc}} = \begin{pmatrix} \boldsymbol{A} & -\boldsymbol{BK} \\ \boldsymbol{LC} & (\boldsymbol{A} - \boldsymbol{BK} - \boldsymbol{LC}) \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & -2 & -2 \\ 4 & 0 & -4 & 1 \\ 5 & 0 & -6 & -2 \end{pmatrix}$$

$$\boldsymbol{B}_{\text{oc}} = \begin{pmatrix} \boldsymbol{B} \\ \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \boldsymbol{C}_{\text{oc}} = \begin{pmatrix} \boldsymbol{C} & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}.$$

g. *Controller transfer function*. From Eq. (4.69), the transfer function $K_{\text{ob}}(s)$ is given by

$$K_{\text{ob}}(s)(s) = \left[\boldsymbol{K} \left(s\mathbb{I} - \boldsymbol{A} + \boldsymbol{BK} + \boldsymbol{LC} \right)^{-1} \boldsymbol{L} \right] = \frac{18(1+s)}{14 + 6s + s^2},$$

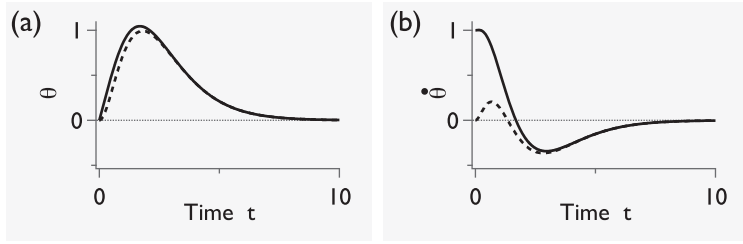
as found using a computer-algebra program. We would use this one in a "real" controller, as it takes directly the output y and produces the input u to be supplied to the physical system. (We would convert to the time domain, of course.)

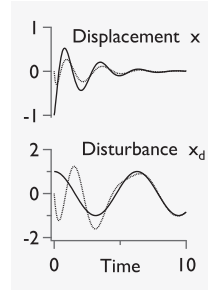
h. *Impulse response*. See below for the responses. The system responses are the solid lines, and their estimates are denoted by dashed lines. (a) records $\theta(t)$ and $\hat{\theta}(t)$. (b) plots $\dot{\theta}(t)$ and $\dot{\hat{\theta}}(t)$. We assume that the initial conditions are identical for system and estimate (as they would be, approximately, if the estimator had been running a while before the disturbance hits). Notice in (a) that $\hat{\theta}$ initially lags the true θ. This makes sense: it takes a while for the estimator to realize "where θ has gone, and then it revises the estimate to track it. In (b), the initial disagreement is far worse. There is no way for the estimator to realize, at first, that the velocity has instantaneously changed. Again, it quickly "figures it out" and the estimate also converges to the true angular velocity.

Normally, these calculations should be done numerically. This problem turns out to be simple enough that there are reasonable analytic expressions. Indeed, using a computer-algebra program, we find, for $t > 0$,

$$\theta(t) = \mathrm{e}^{-2t}(6t + 14) + \mathrm{e}^{-t}(9t - 14) \qquad \hat{\theta}(t) = \mathrm{e}^{-2t}(5t + 14) + \mathrm{e}^{-t}(9t - 14)$$
$$\dot{\theta}(t) = \mathrm{e}^{-2t}(5t + 14) + \mathrm{e}^{-t}(9t - 14) \qquad \dot{\hat{\theta}}(t) = \mathrm{e}^{-2t}(14t + 23) + \mathrm{e}^{-t}(9t - 23),$$

and these are what we actually plot below.





4.10 Canceling a sinusoidal disturbance. Following Section 4.3.3, explore a sinusoidal disturbance that affects a first-order system. Calculate the displacement x and the disturbance position x_d and their estimates $\hat{x}$ and $\hat{x}_\mathrm{d}$. Remember that Eq. (4.79) describes a 6×6 matrix. Reproduce the plot at right, using $\boldsymbol{A} = -1$, $\boldsymbol{B} = \boldsymbol{C} = 1$, $\boldsymbol{A}_\mathrm{d} = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, $\boldsymbol{B}_\mathrm{d} = \left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right)$, $\boldsymbol{C}_\mathrm{d} = (1 \;\; 0)$ and $K = 2$, $L = 4$. Choose $\boldsymbol{L}_\mathrm{d}$ so that the poles of the disturbance observer are at $(-4, -4)$. Initial conditions are $x(0) = -1$, $\boldsymbol{x}_\mathrm{d} = \left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right)$. The initial conditions for the estimators $\hat{x}$ and $\hat{\boldsymbol{x}}_\mathrm{d}$ are zero.

Solution.

Choosing $\boldsymbol{L}_\mathrm{d}$ so that the poles of the disturbance observer are at $(-4, -4)$ leads to $\boldsymbol{L}_\mathrm{d} = \left(\begin{smallmatrix} 8 \\ 15 \end{smallmatrix}\right)$. We can get this directly from standard pole-placement routines or matching coefficients, but they are easy to find directly:

$$\boldsymbol{A}_\mathrm{d} - \boldsymbol{L}_\mathrm{d}\boldsymbol{C}_\mathrm{d} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} - \begin{pmatrix} \ell_1 \\ \ell_2 \end{pmatrix} \begin{pmatrix} 1 & 0 \end{pmatrix} = \begin{pmatrix} -\ell_1 & 1 \\ -1-\ell_2 & 0 \end{pmatrix}.$$

The characteristic equation is

$$(\lambda + \ell_1)\lambda + (1 + \ell_2) = \lambda^2 + \ell_1\lambda + \ell_2 + 1 = 0\,,$$

which has roots

$$\lambda = -\frac{1}{2}\left[\ell_1 \pm \sqrt{\ell_1^2 - 4(\ell_2 + 1)}\right].$$

For $\ell_1 = 8$ and $\ell_2 = 15$, the roots are at $(-4, -4)$, as desired.

$$\boldsymbol{A}_\mathrm{big} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B}\boldsymbol{C}_\mathrm{d} & -\boldsymbol{B}\boldsymbol{K} & -\boldsymbol{B}\boldsymbol{C}_\mathrm{d} \\ \boldsymbol{0} & \boldsymbol{A}_\mathrm{d} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{L}\boldsymbol{C} & \boldsymbol{0} & (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K} - \boldsymbol{L}\boldsymbol{C}) & \boldsymbol{0} \\ \boldsymbol{L}_\mathrm{d}\boldsymbol{C} & \boldsymbol{0} & -\boldsymbol{L}_\mathrm{d}\boldsymbol{C} & \boldsymbol{A}_\mathrm{d} \end{pmatrix}$$

$$= \begin{pmatrix} -1 & 1 & 0 & -2 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & -7 & 0 & 0 \\ 8 & 0 & 0 & -8 & 0 & 1 \\ 15 & 0 & 0 & -15 & -1 & 0 \end{pmatrix}$$

The initial condition is $x(0) = -1$ and $x_\mathrm{d}(0) = 1$, with all others zero. In the big state space, this is

$$\boldsymbol{x}_0 = \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Note that there are no inputs: the initial condition $\boldsymbol{x}(0)$ is enough to generate the non-trivial dynamics shown in the graph via $\boldsymbol{x}(t) = \mathrm{e}^{\boldsymbol{A}_\mathrm{big} t}\, \boldsymbol{x}(0)$.

If the observer poles are moved too far over, the response can develop undamped poles that prevent convergence of $\hat{x}(t)$ to $x(t)$.

Discrete-Time Systems 5

Problems

5.1 **Analog low-pass filters**. Low-pass filters with faster fall-off than a first-order system can remove high-frequency components of a signal before digitization. Recall the form $G(s) = (1 + 2\zeta s + s^2)^{-1}$ of the scaled transfer function of an arbitrary second-order system. Let us explore properties of different choices for the damping ζ.

a. *Butterworth filter*. For a given filter order, the Butterworth filter has the flattest magnitude response. Specifically, the nth-order Butterworth filter is defined as

$$|G(\mathrm{i}\omega)| = \frac{1}{\sqrt{1 + \omega^{2n}}} \approx 1 - \tfrac{1}{2}\omega^{2n} + O(\omega^{4n}) .$$

The lack of low-order terms leads to a flat response. Show that, for $n = 2$, the Butterworth filter corresponds to $\zeta = \frac{1}{2}\sqrt{2} \approx 0.71$.

b. *Bessel filter*. For a given filter order, the Bessel filter has the closest approximation to a linear phase response, $\varphi = -\tau\omega$, which corresponds to a delay in the signal by τ. At order n, the best approximation is to have $\varphi(\omega) = -\tau^*\omega + O(\omega^{2n})$, with τ^* the approximate delay. Show that the $n = 2$ Bessel filter has $\tau^* = \sqrt{3}$ and $\zeta = \frac{1}{2}\sqrt{3} \approx 0.87$, which is slightly more damped than the Butterworth filter.

c. Make Bode plots of the frequency response of the $n = 2$ Butterworth and Bessel filters. Compare the response of a naive cascade of first-order elements, $G_{\mathrm{naive}}(s) = (1 + s)^{-2}$, corresponding to $\zeta = 1$. For the Bessel filter, plot also phase vs. frequency on a linear scale, to see how well it approximates a linear phase response. Finally, plot the step response of all three filters. Why is G_{naive} not a great choice?

The Bessel filter has a "nicer" time response, the Butterworth a nicer frequency response. The differences are more striking for higher-order filters. Software packages can generate these, as well as variants such as elliptic and Chebyshev filters (Smith, 1999). Some Bessel filters are defined using a frequency scale that sets $\tau^* = 1$.

Solution.

a. *Butterworth*:

$$\begin{aligned} |G(\mathrm{i}\omega)| &= \left| \frac{1}{1 + 2\,\mathrm{i}\zeta\omega - \omega^2} \right| \\ &= \left[1 - 2\omega^2 + \omega^4 + 4\zeta^2\omega^2 \right]^{-1/2} \\ &= \left[1 + (4\zeta^2 - 2)\omega^2 + \omega^4 \right]^{-1/2} , \end{aligned}$$

which has the desired form $G(s) = \frac{1}{\sqrt{1+\omega^4}}$ if $4\zeta^2 - 2 = 0$. That is, if $\zeta = \frac{1}{2}\sqrt{2}$.

b. *Bessel*: We need the Taylor expansion for the arc tangent. Using

$$\tan\varphi = \frac{\operatorname{Im} G(\mathrm{i}\omega)}{\operatorname{Re} G(\mathrm{i}\omega)} = \frac{-2\zeta\omega}{1 - \omega^2} \equiv x ,$$

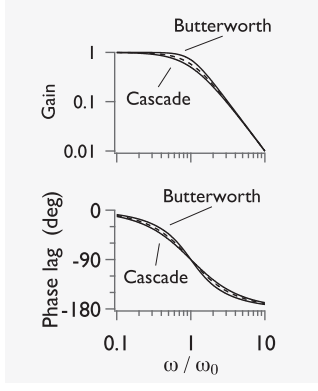
we have

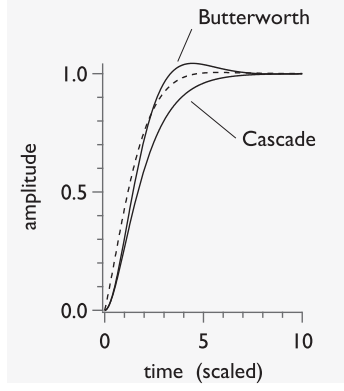
$$\begin{aligned} \varphi = \tan^{-1}(x) &\approx x - \tfrac{1}{3}x^3 + O\left(x^5\right) \\ &= \frac{-2\zeta\omega}{1 - \omega^2} - \frac{1}{3}\frac{-8\zeta^3\omega^3}{(1 - \omega^2)^3} + O\left(\omega^5\right) \\ &= -2\zeta\omega - 2\zeta\left(1 - \frac{4}{3}\zeta^2\right)\omega^3 + O\left(\omega^5\right) . \end{aligned}$$

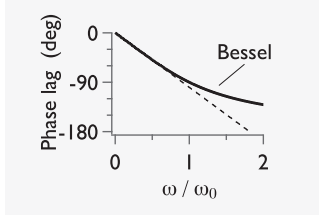
For $\zeta = \frac{1}{2}\sqrt{3}$, the third-order term vanishes, and the phase lag is

$$\varphi = -\sqrt{3}\omega + O\left(\omega^5\right) .$$

c. *Plots*. Left: Bode plots for naive (cascade), Butterworth, and Bessel second-order filters. The Bessel filter uses the dashed line. Right: corresponding step responses. Bottom: Bessel phase response vs. linear approximation. Higher-order Bessel filters would follow the linear phase lag up to higher frequencies.







Notice how the Butterworth filter has the flattest frequency response and how the Bessel filter has the closest approximation to a square-wave time response. The differences are small for 2nd-order filters but become more important at higher orders. The naive (cascade) filter has little to recommend it beyond simplicity. The bandwidth is lower, the phase not linear, and the step response is quadratic, rather than linear, at short times. Still, it is commonly used, as it can be constructed from two independent first-order blocks.

5.2 **Dithering details**. Why are some choices for random dither ξ better than others? Define $\langle x\rangle \equiv \int_{-\infty}^{\infty} \mathrm{d}\xi\, p(\xi)\, Q(x_0+\xi)$ and $\sigma \equiv \sqrt{\mathrm{Var}}$, with $\mathrm{Var} = \int_{-\infty}^{\infty} \mathrm{d}\xi\, p(\xi)\,[Q(x_0+\xi) - \langle x\rangle]^2$. Here, $p(\xi)$ is the probability density function of the added dither, and $Q(x)$ is the quantization nonlinearity, defined as rounding x to the nearest integer.

a. Consider uniform noise, $p(\xi) = 1$ for $(-\frac{1}{2}, +\frac{1}{2})$, or $\sqcap\!\sqcup$. Show that $\langle x\rangle = x_0$, so that there is no bias. Then show that $\mathrm{Var} = (\delta x_0)(1-\delta x_0)$, where δx_0 is the fractional part of x_0. (That is, if $x_0 = 3.1$, then $\delta x_0 = 0.1$.)
b. For triangular noise, $p(\xi) = 1 - |\xi|$, for $|\xi| < 1$ (and 0 for $|\xi| > 1$), or $\_/\!\!\diagup\!\!\diagdown\_$, show that $\langle x\rangle = x_0$ and $\mathrm{Var} = \frac{1}{4}$. There is again no bias, and $\sigma = \frac{1}{2}$ is independent of x_0.
c. Investigate a Gaussian dither of standard deviation σ_0 numerically. Plot both the bias of $\langle x\rangle$ (deviation from the mean) and its variance as a function of x_0. Investigate for $\sigma_0 = 0.4$, 0.5, and 0.6. Is there an optimal value for σ_0?
d. *Subtractive dithering*. Let $x \equiv Q(x_0+\xi) - \xi$. Show that $\langle x\rangle = x_0$ and $\mathrm{Var} = \frac{1}{12}$ for this new x. Thus, with uniform noise, the standard deviation is not only independent of x_0, it is $\sqrt{3}$ lower than using triangular noise. Why doesn't everyone use subtractive dithering? The catch is the need to subtract the exact analog noise value added to the analog signal. Usually, this value is hard to know.

Solution.

This problem involves several simple but "annoying" integrals. We do the first and then turn to computer algebra programs for the remainder.

a. *Uniform dither*. Let $Q(x)$ be the rounding (quantization) function and $p(\xi)$ is a uniform distribution from $(-\frac{1}{2}, +\frac{1}{2})$. Then the mean value is

$$\langle x\rangle = \int_{-\infty}^{\infty} \mathrm{d}\xi\, p(\xi)\, Q(x_0+\xi) = \int_{-1/2}^{1/2} \mathrm{d}\xi\, Q(x_0+\xi)\,,$$

Let $x_0 = \lfloor x_0\rfloor + \delta x_0$, where $\lfloor x_0\rfloor$ is the floor function (integer part of x_0, truncated) and δx_0 is the fractional part. Then, by inspection, for $0 < \delta x_0 < \frac{1}{2}$, we have

$$Q(x_0) = \begin{cases} \lfloor x_0\rfloor & -\frac{1}{2} < \xi < \left(\frac{1}{2} - \delta x_0\right) \\ \lfloor x_0\rfloor + 1 & \left(\frac{1}{2} - \delta x_0\right) < \xi < \frac{1}{2}\,. \end{cases}$$

and

$$\int_{-1/2}^{1/2} d\xi\, Q(x_0+\xi) = \lfloor x_0 \rfloor (1-\delta x_0) + (\lfloor x_0 \rfloor + 1)\delta x_0$$
$$= \lfloor x_0 \rfloor - \cancel{\delta x_0 \lfloor x_0 \rfloor} + \cancel{\delta x_0 \lfloor x_0 \rfloor} + \delta x_0 = x_0 \,.$$

The case $\frac{1}{2} < \delta x_0 < 1$ is similar.
For the variance,

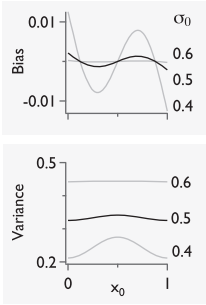
$$\text{Var} = \left\langle (x - x_0)^2 \right\rangle = \int_{-\infty}^{\infty} d\xi\, p(\xi)\, [Q(x_0+\xi) - x_0]^2 = \int_{-1/2}^{1/2} d\xi\, [Q(x_0+\xi) - x_0]^2 \,,$$

Using a similar notation as for the mean and also doing the case $0 < \delta x_0 < \frac{1}{2}$, we have

$$= \int_{-1/2}^{1/2-\delta x_0} d\xi\, (\lfloor x_0 \rfloor - x_0)^2 + \int_{1/2-\delta x_0}^{1/2} d\xi\, (\lfloor x_0 \rfloor + 1 - x_0)^2$$
$$= (\delta x_0)^2\, (1-\delta x_0) + (1-\delta x_0)^2\, (\delta x_0)$$
$$= (\delta x_0)\, (1-\delta x_0)$$

The other case is similar.

b. *Triangular dither*. This is very similar to the uniform-noise case, but the math is uglier. See the Mathematica code, which uses the `Round` function.



c. *Numerical simulations of Gaussian dither*. See graphs at left. We see that as the standard deviation of the Gaussian dither σ_0 increases, the bias decreases, and the variance increases. Choosing $\sigma_0 = 0.5$ seems a reasonable compromise, but for a particular application, one might favor a higher or lower σ_0. Note that with $\sigma_0 = 0.5$, the variance is about 0.33, compared to 0.25 with triangular noise. In addition, there is a slight dependence of both bias and variance on x_0 for Gaussian dither, whereas the bias is zero and the variance independent of x_0 for triangular dither. The latter is thus a better choice, but the practical consequences of the difference is slight, and Gaussian noise is often present "for free" in your measuring system.... (Manufacturers of some data-acquisition boards design the noise level with this issue in mind.) Thus, in practice, the most commonly encountered dithering scheme is with Gaussian noise, with a standard deviation equal to roughly half a quantization level.

d. *Subtractive dither*. This is again similar to triangular dither but with even more complications. The Mathematica code is straightforward and verifies the claimed results.

5.3 Compressed sensing and the counterfeit coin.

a. For the seven-coin / one-fake problem with the three measurements given, verify explicitly that each possibility leads to a unique pattern of measurements. To guide our intuition, change "coordinates" in the manner suggested in the text, so that each genuine mass has $f_i = 0$, while the fake coin has mass $f_j = 1$.

b. Now assume that there are one or two (identical) fakes. Find an explicit counterexample where inferring which masses are fakes is impossible.

Solution.

a. We consider explicitly the predicted measurements for each case. We define Case 1 to be the case where the "mass deviation" vector $\boldsymbol{f}_1$ is given by

$$\boldsymbol{f}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Similar definitions apply to the other six cases.
Recall from Eq. (5.10) that

$$\begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \\ f_7 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}.$$

The measurement vectors $\boldsymbol{y}_i$ for each case are then given by the columns of $\boldsymbol{\Phi}$. Since, by inspection, each column has a distinct pattern, all we have to do is match the measurement vector $\boldsymbol{y}$ to one of the columns to identify the fake mass. (More generally, the measurement vector is proportional to the mass difference times the corresponding column.) Since each column is different in the chosen $\boldsymbol{\Phi}$, a unique reconstruction is possible.

b. Now consider the case where $\boldsymbol{f}$ has exactly two non-zero elements (both equalling one, for simplicity). With two non-zero elements, $\boldsymbol{y}$ is the sum of the two corresponding columns. We then see that

$$\boldsymbol{f} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{f} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

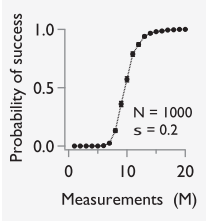
both lead to

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{f} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

It is then impossible to infer from $\boldsymbol{y}$ which $\boldsymbol{f}$ is responsible. Reconstruction thus fails for this particular $\boldsymbol{\Phi}$. We note that there are ${}_7C_1 + {}_7C_2 = 7 + 21 = 28$ possibilities for the 1-or-2-fakes scenario, suggesting the need for 5 measurements to guarantee success ($2^5 = 32 > 21$).

Notice, in this problem, that the a priori knowledge of the number of fakes (e.g., one in part (a)) is translated to a sparsity condition and allows the improvement over an algorithm where nothing was known about the number of fake coins.

5.4 **Phase transition in compressed sensing**. Consider the N-coin / 1-fake problem. Generate the $M \times N$-dimensional measurement matrix $\boldsymbol{\Phi}$ numerically by letting each element be 0 or 1 with 50% probability. Then write a code to do the reconstruction numerically. Use a brute-force algorithm that examines each of the N possibilities explicitly, predicts the outcome $\boldsymbol{y}$ based on the choice of $\boldsymbol{f}$, and calculates the ℓ_2 norm of the error, $\|\boldsymbol{y} - \boldsymbol{y}_0\|_2$. Then select the $\boldsymbol{f}$ that minimizes this error. For a given m, repeat enough times to estimate the probability P of identifying the correct nonzero element of $\boldsymbol{f}$. Then vary M to estimate $P(M)$. You should find something resembling the graph at left. Add measurement noise, with $\boldsymbol{y}_0 = \boldsymbol{\Phi}\boldsymbol{f} + \boldsymbol{\xi}$, where $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathcal{I})$. Confirm that the reconstruction algorithm is robust against moderate noise levels.



Surprisingly, the probability for successful reconstruction rises sharply at $M^* \approx \ln N$: there is a *phase transition* in reconstruction probability, controlled by the relative measurement number M/N and sparseness S/N, with a qualitative difference between a low-data "phase" where reconstruction is impossible and a high-data "phase" where it succeeds almost always. This phase transition is universal: many choices for $\boldsymbol{\Phi}$ give the same reconstruction thresholds, or phase boundaries. See Donoho and Tanner (2009) and Krzakala et al. (2012).

Solution.

The program is relatively straightforward. The algorithm consists of the following steps:

- Pick the "true" (sparse) vector $\boldsymbol{f}$. $N - 1$ elements equal zero; the remaining one equals 1.
- Pick a measurement matrix of random 1's and 0's. (Use sign of a random number that is symmetrically distributed about 0.)
- Each component of $\boldsymbol{f}$ picks out a column of $\boldsymbol{\Phi}$. Thus, make an $M \times N$ matrix by repeating the M-dimensional measurement vector $\boldsymbol{y}$ N times.
- Calculate the ℓ_2 norm of each column.
- Select the minimum. Because of the measurement noise, there will be no ties.

- Repeat many times to get an average success rate.
- Repeat for different measurement numbers M.

Note that an essentially identical curve is seen in the noiseless case ($\sigma = 0$).

5.5 **Final Value Theorem for Z-transform**. Show that $\lim_{k\to\infty} f_k = \lim_{z\to 1}[(z-1)f(z)]$. Hint: Take $\mathcal{Z}(f_{k+1} - f_k)$, and write the infinite sum as a limit $k \to \infty$ of a finite sum.

Solution.

Following the hint,

$$\begin{aligned}
\mathcal{Z}[f_{k+1} - f_k] &= \sum_{n=0}^{\infty} (f_{n+1} - f_n)\, z^{-n} \\
&= \lim_{k\to\infty} \sum_{n=0}^{k} (f_{n+1} - f_n)\, z^{-n} \\
&= \lim_{k\to\infty} \left[(f_1 - f_0) + (f_2 - f_1)z^{-1} + (f_3 - f_2)z^{-2} + \cdots + (f_{k+1} - f_k)z^{-k} \right] \\
&= \lim_{k\to\infty} \left[-f_0 + (1 - z^{-1})f_1 + (1 - z^{-1})z^{-1} f_2 + \cdots + (1 - z^{-1})z^{-k+1} f_k + f_{k+1} \right] .
\end{aligned}$$

We then take the limit $z \to 1$. Assuming that $\lim_{k\to\infty} f_k$ exists, then the Z-transform converges, and we can interchange the limits. First, we set $z \to 1$ and then $k \to \infty$:

$$\lim_{z\to 1} \mathcal{Z}[f_{k+1} - f_k] = -f_0 + \lim_{k\to\infty} f_{k+1} .$$

Alternatively, we can use the shift theorem:

$$\mathcal{Z}[f_{k+1} - f_k] = [zf(z) - f_0] - f(z) = (z-1)f(z) - zf_0 .$$

Taking the $z \to 1$ limit gives

$$\lim_{z\to 1} \mathcal{Z}[f_{k+1} - f_k] = \lim_{z\to 1}[(z-1)f(z) - zf_0] = \lim_{z\to 1}[(z-1)f(z)] - f_0 .$$

Finally, equating the two expressions, we have the *final value theorem*:

$$\lim_{z\to 1}[(z-1)f(z)] = \lim_{k\to\infty} f_{k+1} = \lim_{k\to\infty} f_k .$$

We did not ask for the proof of the *initial value theorem*, but it is simple:

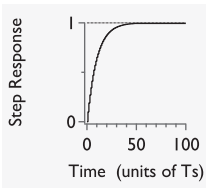
$$\lim_{z\to\infty} f(z) = \lim_{z\to\infty} \sum_{k=0}^{\infty} f_k z^{-k} = \lim_{z\to\infty} \left(f_0 + f_1 z^{-1} + f_2 z^{-2} + \cdots \right) = f_0 .$$

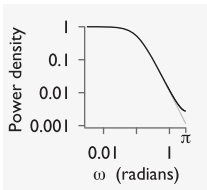
5.6 **IIR vs. FIR low-pass filter**. In Example 5.2, we claimed that the IIR filter

$$y_k = a y_{k-1} + (1-a)u_{k-1} , \qquad 0 < a < 1$$

is equivalent to the IIR filter

$$y_k = A(u_k + a u_{k-1} + a^2 u_{k-2} + \cdots + a^n u_{k-n}) ,$$





where A is a normalization constant chosen to make the DC gain equal to 1. For large n, the two transfer functions become very similar. To see this:

a. By Z-transformation, derive the form of the transfer function for both filters.
b. Show that the IIR and FIR transfer functions are identical for $n \to \infty$.
c. Show that the corner frequency of the equivalent continuous system is $\omega_0 = \frac{1-a}{\sqrt{a}}$.
d. Reproduce the step and frequency response graphs shown at left.

Solution.

a. For the IIR filter, we take the Z-transform:

$$y = az^{-1}y + (1-a)z^{-1}u\,,$$

which implies

$$\frac{y}{u} = \frac{(1-a)z^{-1}}{1-a/z} = \frac{1-a}{z-a}\,.$$

For the FIR filter, the Z-transform gives

$$\frac{y}{u} = A\left[1 + (a/z) + (a/z)^2 + \cdots + (a/z)^n\right].$$

At $\omega = 0$ (or $z = 1$), we have

$$A(1 + a + a^2 + \cdots + a^n) = A\left(\frac{1-a^{n+1}}{1-a}\right) = 1\,,$$

which implies

$$A = \left(\frac{1-a}{1-a^{n+1}}\right).$$

b. In the $n \to \infty$ limit, $A \to 1 - a$. For $|a| < 1$, we can also rewrite

$$\left[1 + (a/z) + (a/z)^2 + \cdots + (a/z)^n\right] = \frac{1-\left(\frac{a}{z}\right)^{n+1}}{1-\left(\frac{a}{z}\right)} \to \frac{1}{1-\frac{a}{z}}\,.$$

When $n \to \infty$, we have $A \to 1 - a$ and, thus,

$$\frac{y}{u} = \frac{1-a}{1-\left(\frac{a}{z}\right)}\,.$$

The two filters then have the same transfer function up to a delay of one unit.

c. From (a), the Power density for the IIR filter $P(\omega)$ is

$$\begin{aligned} P(\omega) &= \left|\frac{1-a}{z-a}\right|^2_{z=e^{i\omega}} \\ &= \frac{(1-a)^2}{(e^{i\omega}-a)(e^{-i\omega}-a)} \end{aligned}$$

$$= \frac{(1-a)^2}{1+a^2-2a\cos\omega}$$
$$= \frac{(1-a)^2}{(1-a)^2+2a(1-\cos\omega)}\,.$$

Taylor expanding gives $1-\cos\omega \approx \frac{1}{2}\omega^2$ and, hence,

$$P(\omega) = \frac{(1-a)^2}{(1-a)^2+a\omega^2}$$
$$= \frac{1}{1+\left(\frac{\omega}{\omega_c}\right)^2}\,,$$

with

$$\omega_c = \left(\frac{1-a}{\sqrt{a}}\right)\frac{1}{T_s}\,.$$

The sampling frequency, $1/T_s$, appears if we redo the problem in dimensional units, substituting $z = e^{i\omega T_s}$, rather than $z = e^{i\omega}$.

d. See the book website for code. To go from the transfer function to the frequency response, we substitute $z = e^{i\omega T_s}$ and then take the magnitude squared and phase of $y(z)/u(z)$.

5.7 FIR filter with linear phase response. For $n = 2N+1$ odd, consider the FIR filter

$$y_k = B_0u_k + B_1u_{k-1} + \cdots + B_{n-1}u_{k-n+1}\,,$$

a. Show that if $B_m = B_{n-1-m}$, then the complex frequency response $y(\omega)$ has a linear phase. That is, show that $y(\omega) = \tilde{y}(\omega)\,e^{-i\omega\tau}$, where $\tilde{y}(\omega)$ and τ are real.

b. An ideal low-pass filter would have a frequency response that is 1 for $\omega < \omega_c$ and 0 for $\omega_c < \omega < (\pi/T_s)$. Show we can realize the filter via FIR coefficients

$$(B_m)_{\text{acausal}} = \left(\frac{\omega_c}{\pi}\right)\left(\frac{\sin m\omega_cT_s}{m\omega_cT_s}\right), \qquad -\infty < m < \infty\,,$$

where m is integer. For negative m, the ideal low-pass filter is *acausal* and cannot be implemented in real time, since it needs future information.

c. To make a realizable filter, truncate to $n = 2N+1$ terms and then delay each component to make it causal. The resulting filter has

$$B_m = \left(\frac{\omega_c}{\pi}\right)\left(\frac{\sin(m-N)\omega_cT_s}{(m-N)\omega_cT_s}\right), \qquad 0 < m < 2N\,,$$

Verify that this filter is linear phase and plot the magnitude of the frequency response for $n = 101$ and 1001. Note and explain the *Gibb's phenomenon*.

d. Discuss the effects of multiplying the FIR coefficients by a *Hamming window*,

$$B_m \to B_m \times \left[0.54 - 0.46\cos\left(\frac{\pi m}{N}\right)\right], \qquad 0 < m < 2N\,.$$

Solution.

a. Recall that the frequency response of y_k is given by Eqs. (5.35) and (5.35). That is, we take the Z-transform

$$y = B_0 + B_1 z^{-1} + \cdots + B_{n-2} z^{-n+2} + B_{n-1} z^{-n+1} .$$

Assume $B_m = B_{n-1-m}$. Then

$$\begin{aligned} y &= B_0(1 + z^{-n+1}) + B_1(z^{-1} + z^{-n+2}) + \cdots + B_N z^{-n+N+1} \\ &= B_0(1 + z^{-2N}) + B_1(z^{-1} + z^{-2N+1}) + \cdots + B_N z^{-N} \\ &= z^{-N}\left[B_0(z^N + z^{-N}) + B_1(z^{N-1} + z^{-N+1}) + \cdots + B_N\right] \\ &= 2\,\mathrm{e}^{-\mathrm{i}N\omega T_\mathrm{s}}\left\{B_0 \cos N\omega T_\mathrm{s} + B_1 \cos[(N-1)\omega T_\mathrm{s}] \cdots + B_N\right\} , \end{aligned}$$

where we substitute $z = \mathrm{e}^{\mathrm{i}\omega T_\mathrm{s}}$ to calculate the frequency response. Thus, the frequency response has a linear phase lag ($\tau = NT_\mathrm{s}$). The real function $\tilde{y}(\omega)$ is given by

$$\tilde{y}(\omega) = 2\left\{B_0 \cos N\omega T_\mathrm{s} + B_1 \cos[(N-1)\omega T_\mathrm{s}] + \cdots + B_N\right\}$$

b. We recall the calculation of the time response of a function that passes frequencies in the range $-\omega_\mathrm{c} < \omega < +\omega_\mathrm{c}$. We need the negative as well as positive frequencies. By the inverse Fourier transform, we have

$$B(t) = \frac{1}{2\pi}\int_{-\omega_\mathrm{c}}^{\omega_\mathrm{c}} \mathrm{d}\omega\ \mathrm{e}^{\mathrm{i}\omega t} = \frac{1}{2\pi}\frac{\mathrm{e}^{\mathrm{i}\omega t}}{it}\bigg|_{-\omega_\mathrm{c}}^{\omega_\mathrm{c}} = \frac{\omega_\mathrm{c}}{\pi}\left(\frac{\sin \omega_\mathrm{c} t}{\omega_\mathrm{c} t}\right) .$$

Then evaluating at times mT_s gives $B_m = B(mT_\mathrm{s})$:

$$B_m = \frac{\omega_\mathrm{c}}{\pi}\left(\frac{\sin \omega_\mathrm{c} mT_\mathrm{s}}{\omega_\mathrm{c} mT_\mathrm{s}}\right) .$$

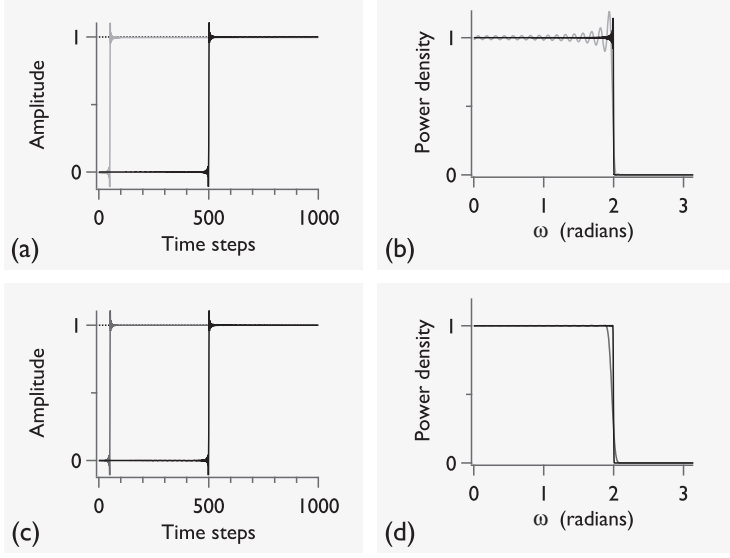
As discussed, we need to consider all integer values of m, including negative values, to have an ideal filter.

c. We truncate to $2N + 1$ terms and delay the filter by NT_s by setting $m \to m - N$. The coefficients are then, as claimed,

$$B_m = \left(\frac{\omega_\mathrm{c}}{\pi}\right)\left(\frac{\sin(m-N)\omega_\mathrm{c} T_\mathrm{s}}{(m-N)\omega_\mathrm{c} T_\mathrm{s}}\right) , \qquad 0 < m < 2N ,$$

Such a filter has linear phase, since evaluating $B_{n-1-m} = B_{2N-m}$ amounts to letting $(m - N) \to [(2N - m) - N] = -(m - N)$. That is, we multiply the numerator and denominator by -1, which leaves the coefficients unchanged. We thus verify that $B_{2N-m} = B_m$. See the plots below, which show the step and frequency response of an FIR low-pass filter. (a) Step response for 101 (gray line) and 1001 (black line) coefficients of a truncated sinc(·) FIR filter. (b) Corresponding frequency response. The Gibbs phenomenon results because the jump discontinuity needs all frequencies to represent accurately. (It's a longer story, but this is enough for now!)

d. Windowing smooths the jump discontinuity caused by the truncation of filter coefficients, at the cost of distorting their values somewhat. Below, we graphically show that the frequency response becomes much flatter when using a Hamming window. On the other hand, the step response still shows ringing. More sophisticated techniques can minimize this ringing.



5.8 Discrete Parseval's theorem.

a. Using the definitions of the discrete time Fourier transform (DTFT) given in Section 5.2.3, derive Parseval's Theorem, Eq. (5.37).

b. By integrating around the unit circle, derive an alternate form of the theorem,

$$\sum_{k=0}^{\infty} f_k^2 = \frac{1}{2\pi \mathrm{i}} \oint \frac{\mathrm{d}z}{z} f(z)\, f(z^{-1}) \,.$$

c. Show that Parseval's Theorem works explicitly for the transform pair $f_k = a^k \theta_k$, with $|a| < 1$ and $f(z) = \frac{z}{z-a}$, with $z = \mathrm{e}^{\mathrm{i}\omega}$. Here, $\theta_k = 1$ for $n \geq 0$ and 0 otherwise.

Solution.

a. Recall from Eq. (5.36) that the DFT transform pair is

$$f(\omega) \equiv \sum_{k=-\infty}^{\infty} f_k \, \mathrm{e}^{-\mathrm{i}\omega k} \qquad \Longleftrightarrow \qquad f_k = \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} f(\mathrm{e}^{\mathrm{i}\omega}) \, \mathrm{e}^{\mathrm{i}\omega k} \,.$$

Then

$$\begin{aligned} \sum_{k=0}^{\infty} f_k^2 &= \sum_{k=0}^{\infty} \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \left[f(\omega) \, \mathrm{e}^{\mathrm{i}\omega k} \right] f_k \\ &= \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} f(\omega) \left(\sum_{k=0}^{\infty} f_k \, \mathrm{e}^{\mathrm{i}\omega k} \right) \end{aligned}$$

$$= \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \, [f(\omega) f(\omega)^*]$$
$$= \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \, |f(\omega)|^2 \, .$$

b. The proof just parallels the one for the DFT. Recall from Eq. (5.32) that the inverse Z-transform is

$$f_k = \mathcal{Z}^{-1}[f(z)] = \frac{1}{2\pi i} \oint_C \mathrm{d}z \, f(z) \, z^{k-1} \, .$$

Then

$$\sum_{k=0}^{\infty} f_k^2 = \sum_{k=0}^{\infty} \frac{1}{2\pi i} \oint_C \mathrm{d}z \left[f(z) \, z^{k-1} \right] f_k$$
$$= \frac{1}{2\pi i} \oint_C \frac{\mathrm{d}z}{z} \, f(z) \left(\sum_{k=0}^{\infty} f_k \, z^k \right)$$
$$= \frac{1}{2\pi i} \oint_C \frac{\mathrm{d}z}{z} \, f(z) \, f(z^{-1}) \, ,$$

where we substitute the forward Z-transform,

$$f(z) = \sum_{k=0}^{\infty} f_k \, z^{-k} \, .$$

c. We check an explicit calculation. First, for $k \geq 0$, we have $f_k = a^k$, and

$$\sum_{k=0}^{\infty} f_k^2 = \sum (a^2)^k = \frac{1}{1-a^2} \, ,$$

for $|a| < 1$. To calculate this in the Z-domain, we write

$$\frac{1}{2\pi i} \oint \frac{\mathrm{d}z}{z} \left(\frac{z}{z-a} \right) \left(\frac{1/z}{1/z-a} \right)$$
$$= \frac{1}{2\pi i} \oint \frac{\mathrm{d}z}{(z-a)(1-az)}$$
$$= \frac{\cancel{2\pi i}}{\cancel{2\pi i}} \frac{1}{1-a^2} \, ,$$

using the residue theorem and evaluating the simple pole at $z = a$, which is inside the unit circle since $|a| < 1$.

5.9 **Discretization of a zero-order hold**. To find the discrete matrices $\boldsymbol{A}_\mathrm{d}$ and $\boldsymbol{B}_\mathrm{d}$ from Eq. (5.40) in one step and without inverting $\boldsymbol{A}$, show that $\exp[T_\mathrm{s} \left(\begin{smallmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{0} \end{smallmatrix}\right)] = \left(\begin{smallmatrix} \boldsymbol{A}_\mathrm{d} & \boldsymbol{B}_\mathrm{d} \\ \boldsymbol{0} & \mathbb{I} \end{smallmatrix}\right)$.

Solution.

By direct calculation,

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}^n = \begin{pmatrix} \boldsymbol{A}^n & \boldsymbol{A}^{n-1}\boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} .$$

Thus,

$$\exp\left[T_\mathrm{s}\begin{pmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\right] = \begin{pmatrix}\mathbb{I} & \boldsymbol{0}\\ \boldsymbol{0} & \mathbb{I}\end{pmatrix} + T_\mathrm{s}\begin{pmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix} + \frac{T_\mathrm{s}^2}{2!}\begin{pmatrix}\boldsymbol{A}^2 & \boldsymbol{AB}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix} + \cdots + \frac{T_\mathrm{s}^n}{n!}\begin{pmatrix}\boldsymbol{A}^n & \boldsymbol{A}^{n-1}\boldsymbol{B}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix} + \cdots$$

$$= \begin{pmatrix}\mathrm{e}^{\boldsymbol{A}T_\mathrm{s}} & \boldsymbol{A}^{-1}\left(\mathrm{e}^{\boldsymbol{A}T_\mathrm{s}} - \mathbb{I}\right)\boldsymbol{B}\\ \boldsymbol{0} & \mathbb{I}\end{pmatrix}$$

$$= \begin{pmatrix}\boldsymbol{A}_\mathrm{d} & \boldsymbol{B}_\mathrm{d}\\ \boldsymbol{0} & \mathbb{I}\end{pmatrix}.$$

Note that even if $\boldsymbol{A}$ is not invertible, we can still directly evaluate the matrix exponential and identify $\boldsymbol{B}_\mathrm{d}$ as the upper right block.

5.10 ZOH discretization. For a continuous function $u(t)$, its zero-order-hold staircase function $u_k(t)$ is defined in Eq. (5.38). Show that

a. The Laplace transform of the zero-order hold is given by $\mathcal{L}[u_\mathrm{ZOH}] = (\frac{1-\mathrm{e}^{-sT_\mathrm{s}}}{s})\mathcal{Z}[u]$, where $\mathcal{Z}[u]$ is the Z-transform of the sequence $u_k = u(kT_\mathrm{s})$.
b. The ZOH discrete transfer function $G_\mathrm{d}(z) = (1 - z^{-1})\,\mathcal{Z}\{\mathcal{L}^{-1}[\frac{G(s)}{s}]\}$, with $G(s)$ a continuous system transfer function and $\mathcal{Z}\{\mathcal{L}^{-1}[\cdot]\}$ the Z-transform of the time domain signal from the inverse Laplace transform, sampled at times kT_s.
c. The 1st-order system $G(s) = \frac{1}{1+s}$ implies that $G_\mathrm{d}(z) = \frac{1-\mathrm{e}^{-T_\mathrm{s}}}{z-\mathrm{e}^{-T_\mathrm{s}}}$ (cf. Eq. 5.42).
d. The 2nd-order system $G(s) = \frac{1}{1+s^2}$ implies that $G_\mathrm{d}(z) = \frac{(1-\cos T_\mathrm{s})(z+1)}{(z-\mathrm{e}^{iT_\mathrm{s}})(z-\mathrm{e}^{-iT_\mathrm{s}})}$. The *sampling zero* at $z = -1$ arises solely from the sampling process; $G(s)$ has no zero.

Solution.

a.

$$\mathcal{L}[u_\mathrm{ZOH}] = u(0)\int_0^{T_\mathrm{s}} \mathrm{d}t\,\mathrm{e}^{-st} + u(T_\mathrm{s})\int_{T_\mathrm{s}}^{2T_\mathrm{s}} \mathrm{d}t\,\mathrm{e}^{-st} + \ldots$$

$$= u(0)\frac{1}{s}\left(1 - \mathrm{e}^{-sT_\mathrm{s}}\right) + u(T_\mathrm{s})\frac{1}{s}\left(\mathrm{e}^{-sT_\mathrm{s}} - \mathrm{e}^{-2sT_\mathrm{s}}\right) + \ldots$$

$$= \left(\frac{1 - \mathrm{e}^{-sT_\mathrm{s}}}{s}\right)\left[u(0) + u(T_\mathrm{s})\,\mathrm{e}^{-sT_\mathrm{s}} + \ldots\right]$$

$$= \left(\frac{1 - z^{-1}}{s}\right)\left[u_0 + u_1\,z^{-1} + \ldots\right]$$

$$= \left(\frac{1 - z^{-1}}{s}\right)\mathcal{Z}[u]\,.$$

It is good to understand the meaning of this formula more intuitively. The Laplace transform of a step function $\theta(t)$ is just $\frac{1}{s}$. If the response is delayed

by T_s it is $e^{-sT_s}(\frac{1}{s})$. Thus, a pulse of duration T_s in the time domain has a Laplace-domain representation of

$$\left(\frac{1-e^{-sT_s}}{s}\right).$$

b. Here is an informal derivation, which should be made more rigorous!

$$y(t) = G(t) * u_k(t),$$

with $u_k(t)$ the staircase time function constructed from the sequence u_k. Laplace transforming and using "operator overload" notation $y(s)$ for $\mathcal{L}[y](s)$, etc., we have

$$\begin{aligned} y(s) &= G(s) * u_k(s) \\ &= G(s)\left(\frac{1-z^{-1}}{s}\right) u(z) \\ &= \left(\frac{G(s)}{s}\right)\left(1-z^{-1}\right) u(z). \end{aligned}$$

We then inverse-Laplace transform to return to the time domain:

$$y(t) = \mathcal{L}^{-1}\left(\frac{G(s)}{s}\right)\left(1-z^{-1}\right) u(z),$$

where we can "leave" the z terms because they are not written in terms of s. (This step is not rigorous!) Then Z-transforming gives

$$y(z) = \mathcal{Z}\left[\mathcal{L}^{-1}\left(\frac{G(s)}{s}\right)\right]\left(1-z^{-1}\right) u(z),$$

which implies a ZOH-discretized transfer function of

$$G_d(z) = \frac{y(z)}{u(z)} = \left(1-z^{-1}\right)\mathcal{Z}\left[\mathcal{L}^{-1}\left(\frac{G(s)}{s}\right)\right].$$

c. We apply these ideas to a first-order system:

$$\frac{G(s)}{s} = \frac{1}{s(s+1)} = \frac{1}{s} - \frac{1}{s+1}.$$

Then

$$\mathcal{L}^{-1}\left(\frac{G(s)}{s}\right) = \theta(t) - e^{-t}\,\theta(t) = 1 - e^{-kT_s},$$

using the discretization at times $t > 0$. Taking the Z-transform then gives

$$\frac{z}{z-1} - \frac{z}{z-e^{-T_s}} = \frac{z(1-e^{-T_s})}{(z-1)(z-e^{-T_s})}.$$

Finally, noting that $1 - z^{-1} = \frac{z-1}{z}$, we have

$$G_d(z) = \left(\frac{z-1}{z}\right)\frac{z(1-e^{-T_s})}{(z-1)(z-e^{-T_s})} = \frac{1-e^{-T_s}}{z-e^{-T_s}}.$$

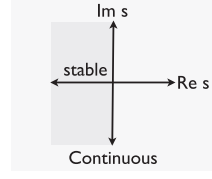
d. For the second-order system (undamped oscillator) given by $G(s) = \frac{1}{1+s^2}$, the zero-hold discretization is

$$\begin{aligned}
G_{\mathrm{d}}(z) &= \left(1 - z^{-1}\right) \mathcal{Z}\left\{\mathcal{L}^{-1}\left[\frac{G(s)}{s}\right]\right\} \\
&= \left(\frac{z-1}{z}\right) \mathcal{Z}\left\{\mathcal{L}^{-1}\left[\frac{1}{s(s^2+1)}\right]\right\} \\
&= \left(\frac{z-1}{z}\right) \mathcal{Z}\left\{\mathcal{L}^{-1}\left[\frac{1}{s} - \frac{s}{s^2+1}\right]\right\} \\
&= \left(\frac{z-1}{z}\right) \mathcal{Z}\left\{\theta(t) - \cos t\right\}_{t \to kT_{\mathrm{s}}} \\
&= \left(\frac{z-1}{z}\right) \mathcal{Z}\left\{1 - \cos kT_{\mathrm{s}}\right\} \\
&= \left(\frac{z-1}{z}\right) \left[\frac{z}{z-1} - \frac{z(z - \cos T_{\mathrm{s}})}{z^2 - 2(\cos T_{\mathrm{s}})z + 1}\right] \\
&= 1 - \frac{(z-1)(z - \cos T_{\mathrm{s}})}{z^2 - 2(\cos T_{\mathrm{s}})z + 1} \\
&= \frac{(1 - \cos T_{\mathrm{s}})(z+1)}{z^2 - 2(\cos T_{\mathrm{s}})z + 1} \\
&= \frac{(1 - \cos T_{\mathrm{s}})(z+1)}{(z - \mathrm{e}^{\mathrm{i}T_{\mathrm{s}}})(z - \mathrm{e}^{-\mathrm{i}T_{\mathrm{s}}})} \,.
\end{aligned}$$

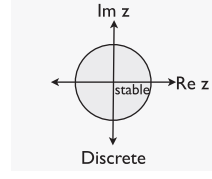
Note that the poles of $G_{\mathrm{d}}(z)$ are on the unit circle, at $p_{\pm} = \mathrm{e}^{\pm \mathrm{i}T_{\mathrm{s}}}$ and that there is a sampling zero at $z = -1$.

In the last two sections, we include so much detail to show explicitly how the calculations work. Normally, one can simply use control software with built-in functions that give directly the desired discretization, especially for numerical calculations (with a numerical value of T_{s}).

5.11 Mapping s to z. Show the following:



a. The change of variable $z = \mathrm{e}^{sT_{\mathrm{s}}}$ maps Re $s < 0$ to $|z| < 1$ (see right).
b. The same mapping is valid for the Tustin transformation: $s = \frac{2}{T_{\mathrm{s}}} \frac{z-1}{z+1}$. Thus, if the continuous system is stable, so too is its Tustin discretization.
c. The backward Euler rule for $s \to z$ gives the *Euler* approximation to an integral, while the Tustin transformation gives the *trapezoidal* algorithm.



Solution.

a. Let $s = s' + is''$. Then $z = \mathrm{e}^{sT_{\mathrm{s}}} = \mathrm{e}^{s'T_{\mathrm{s}}}\, \mathrm{e}^{\mathrm{i}s''T_{\mathrm{s}}} \quad \Longrightarrow \quad |z| = \mathrm{e}^{s'T_{\mathrm{s}}}$, and

$$|z| < 1 \quad \Longrightarrow \quad s' = \mathrm{Re}\, s < 0 \,.$$

b. This one is more intuitive in the reverse direction. Let $z = r\,\mathrm{e}^{\mathrm{i}\theta}$ be a point inside the unit disk (i.e., $r < 1$). Then, from Eq. (5.48)

$$\begin{aligned} s &= \frac{2}{T_\mathrm{s}}\left(\frac{1-z^{-1}}{1+z^{-1}}\right) \sim \left(\frac{1-z^{-1}}{1+z^{-1}}\right) = \left(\frac{z-1}{z+1}\right) \\ &= \left(\frac{r\,\mathrm{e}^{\mathrm{i}\theta}-1}{r\,\mathrm{e}^{\mathrm{i}\theta}+1}\right)\left(\frac{r\,\mathrm{e}^{-\mathrm{i}\theta}+1}{r\,\mathrm{e}^{-\mathrm{i}\theta}+1}\right) \\ &= \left(\frac{r^2-1+2ir\sin\theta}{r^2+1+2r\cos\theta}\right) \end{aligned}$$

and, thus,

$$\mathrm{Re}\, s = -\left(\frac{1-r^2}{r^2+1+2r\cos\theta}\right).$$

For $0 < r < 1$, the numerator of the fraction is positive. The denominator is, too: over the range $0 < \theta < 2\pi$, the denominator is in the positive range $(1-r)^2$ to $(1+r)^2$. Thus

$$0 < r < 1 \quad \leftrightarrow \quad \mathrm{Re}\, s < 0\,.$$

In other words, the mapping $s \leftrightarrow z$ also maps the left-hand part of the s-plane to the interior of the unit disk.

The mapping, of course, is a different mapping from $z = \exp(sT_\mathrm{s})$.

c. Consider an integral in the Laplace domain:

$$I(t) = \int \mathrm{d}t'\, e(t') \quad \rightarrow \quad I(s) = \left(\frac{1}{s}\right)e(s)\,.$$

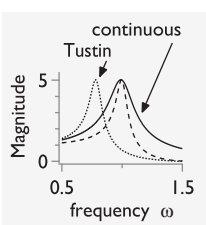
We can view converting $s \rightarrow z$ as approximating the continuous integral with a discrete one, going back from z to the discrete time domain via z^{-1} equalling "delay by T_s:

$$\text{Euler:} \qquad \frac{1}{s} = \frac{T_\mathrm{s}}{1-z^{-1}} \qquad \Longrightarrow\ I_k = I_{k-1} + T_\mathrm{s} e_k\,.$$

$$\text{Tustin:} \qquad \frac{1}{s} = \frac{T_\mathrm{s}}{2}\frac{z+1}{z-1} \qquad \Longrightarrow\ I_{k+1} = I_k + \frac{T_\mathrm{s}}{2}(e_k + e_{k+1})\,.$$

5.12 Tustin transformation and frequency warping.

a. Show that that the Tustin transformation, $s \rightarrow \frac{2}{T_\mathrm{s}}\frac{z-1}{z+1}$ distorts frequencies so that a frequency ω in the continuous system maps to a frequency $\omega' = \frac{2}{T_\mathrm{s}}\tan(\omega T_\mathrm{s}/2)$.



b. Find a value λ to rescale, or "prewarp," the Tustin transformation ($s \rightarrow s' = \lambda s$), so that its frequency response matches the continuous system at $\omega = \omega'$.

c. Plot $|G(s)| = \left|\frac{1}{1+2\zeta s+s^2}\right|$ with $\zeta = 0.1$, its discrete Tustin approximation $|G_\mathrm{Tustin}(z)|$ for $T_\mathrm{s} = 2$, and its prewarped version, matched at $\omega = \omega' = 1$. At left, the dashed line represents the prewarped approximation.

Solution.

a. The frequency response of a continuous system is given by $G(s = \mathrm{i}\omega)$. Let $G_{\mathrm{Tustin}}(z)$ be the corresponding discrete transfer function resulting from the Tustin transformation

$$s \to \left(\frac{2}{T_\mathrm{s}}\right)\left(\frac{z-1}{z+1}\right).$$

Then its frequency response is given by $z \to \mathrm{e}^{\mathrm{i}\omega T_\mathrm{s}}$, so that

$$\left(\frac{z-1}{z+1}\right)\Bigg|_{z=\mathrm{e}^{\mathrm{i}\omega T_\mathrm{s}}} = \left(\frac{\mathrm{e}^{\mathrm{i}\omega T_\mathrm{s}}-1}{\mathrm{e}^{\mathrm{i}\omega T_\mathrm{s}}+1}\right) = \left(\frac{\mathrm{e}^{\mathrm{i}\omega T_\mathrm{s}/2}-\mathrm{e}^{-\mathrm{i}\omega T_\mathrm{s}/2}}{\mathrm{e}^{\mathrm{i}\omega T_\mathrm{s}/2}+\mathrm{e}^{-\mathrm{i}\omega T_\mathrm{s}/2}}\right) = \mathrm{i}\tan(\omega T_\mathrm{s}/2).$$

Thus, we have that

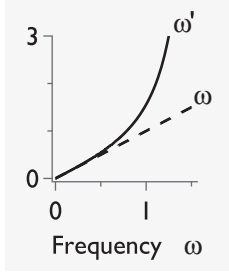
$$s = \mathrm{i}\omega \to s = \mathrm{i}\frac{2}{T_\mathrm{s}}\tan(\omega T_\mathrm{s}/2) \equiv \mathrm{i}\omega'$$

and

$$\omega' = \frac{2}{T_\mathrm{s}}\tan(\omega T_\mathrm{s}/2).$$

The fact that $\omega \neq \omega'$ reflects the distorted frequency response. The frequency-scale distortion vanishes ($\omega' \approx \omega$) for low frequencies $\ll T_\mathrm{s}^{-1}$. For $T_\mathrm{s} = 2$, $\omega' = \tan\omega$ (see below). The Taylor expansion about $\omega = 0$ is

$$\omega' = \omega + \tfrac{1}{3}\omega^3 + O(\omega^5).$$



b. Now let consider a rescaled Tustin transformation, which is stretched along the frequency axis so that the frequency response of the discrete filter matches the frequency response of the corresponding continuous filter at one particular frequency ω'. Let

$$s' = \lambda\left(\frac{z-1}{z+1}\right).$$

Then evaluating at $z = \exp(i\omega T_\mathrm{s})$ leads to

$$s' = \lambda\ \mathrm{i}\tan(\omega T_\mathrm{s}/2).$$

The two response functions are then matched ($s = s' = \mathrm{i}\omega'$) for the frequency $\omega = \omega'$ when we choose

$$\lambda = \frac{\omega'}{\tan(\omega' T_\mathrm{s}/2)}.$$

The technique is known in the literature as "prewarping." I do not like the name very much, as it suggests a nonlinear correction. The correction is linear. Of course, this transformation works only at the one frequency ω', and the frequency response remains distorted at other frequencies. Still, many filters have just one characteristic frequency, and if you need that frequency to be a substantial fraction of T_s and to be accurate, then this technique can be useful.

c. The continuous system, an underdamped oscillator when $\zeta = 0.1$, has a transfer function

$$G(s) = \frac{1}{1 + 2\zeta s + s^2}\,.$$

For the normal Tustin discretization with $T_s = 2$, we have $s \to (\frac{z-1}{z+1})$ and then

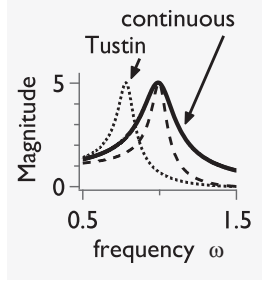
$$G_{\text{Tustin}}(z) = \frac{1}{1 + 2\zeta s + s^2} = \frac{1}{(\frac{z-1}{z+1})^2 + 2\zeta(\frac{z-1}{z+1}) + 1}$$
$$= \frac{(z+1)^2}{2[z^2 + 1 + \zeta(z^2 - 1)]}\,,$$

which is then evaluated for $z = \exp[-\mathrm{i}\omega T_s]$.

For the prewarped Tustin discretization with $T_s = 2$ and matched at $\omega' = 1$, we have $s' = \frac{\omega'}{\tan(\omega')}\left(\frac{z-1}{z+1}\right) = \frac{1}{\tan 1}\left(\frac{z-1}{z+1}\right)$ and then

$$G_{\text{prewarp}}(z) = \frac{1}{\frac{1}{(\tan 1)^2}(\frac{z-1}{z+1})^2 + 2\zeta\frac{1}{\tan 1}(\frac{z-1}{z+1}) + 1}\,.$$

Note that $(\tan 1)^{-1} \approx 0.642093$ and $(\tan 1)^{-2} \approx 0.412283$. See the book website for code to produce the magnitude response plots. The plots are reproduced below, for convenience. We emphasize that the large differences between the normal Tustin response and the continuous (and prewarped) versions results from the fact that the sampling time T_s is such a large fraction of the period of the oscillator. Indeed, the ratio is $2/(2\pi) \approx 0.3$.



5.13 PID discretization: simpler can be better. Sometimes, the simple backward Euler discretization works best. Consider PI control of the continuous system $G(s) = \frac{1}{1+s}$, with $K(s) = 1 + \frac{1}{s}$. For $T_s = 0.1$, find the ZOH discretization $G_d(z)$.

a. Discretize the PI controller using backward Euler, $s \to \frac{1-z^{-1}}{T_s}$, and Tustin, $s \to \frac{2}{T_s}\frac{1-z^{-1}}{1+z^{-1}}$. Plot the step response for all three closed-loop systems.
b. Now add derivative control, $K(s) \to 1 + \frac{1}{s} + 0.1s$. Show that the backward-Euler controller is little changed, but something goes wrong for the Tustin controller.

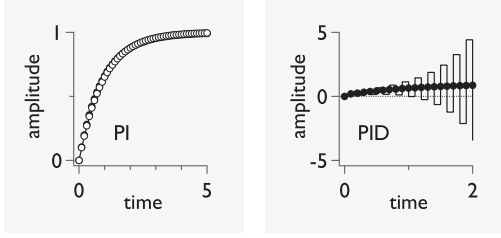
Solution.

a. For $G(s) = \frac{1}{1+s}$ and $K(s) = 1+\frac{1}{s} = \frac{1+s}{s}$, the complementary sensitivity function giving the transfer function from reference to output is given by

$$T(s) = \frac{GK}{1+GK} = \frac{1}{1+(KG)^{-1}} = \frac{1}{1+s},$$

which is just the same as $G(s)$. (So why bother with control? The closed-loop system rejects disturbances, but the open-loop system without $K(s)$ does not.) The graphs below (left) show exponential rise expected for the step response and that both the backward-Euler and Tustin methods give close approximations when sampled at $T_s = 0.1$.
b. For PID control $K(s) \to 1 + \frac{1}{s} + 0.1s$, the backward-Euler continues to track the continuous system well, but we see an oscillatory instability for the Tustin discretization of the controller. Since the derivative term is improper ($\sim s$), Tustin gives a controller pole at $z_p = -1$, which is marginal and wildly oscillatory. Then the closed-loop system perturbs this to an unstable pole at $z_p = -1.2$, outside the unit disk.

Conclusion: stick to backward Euler for PID.



5.14 Controllability of a discrete system. Section 4.1.1 for continuous systems mostly carries over to discrete systems. But let us distinguish the *reachable set* of states $\mathcal{R}_k$ that may be reached from $\boldsymbol{x}_0$ in k steps from the *controllable set* of states C_k, the $\boldsymbol{x}_k$ that maybe brought to $\mathbf{0}$ in k steps. A system is *reachable* if $\mathcal{R}_k = \mathbb{R}^n$ for all $k \geq n$.

a. Prove that a discrete SISO system is reachable if $\boldsymbol{W}_c = (\boldsymbol{B}\ \boldsymbol{AB}\ \boldsymbol{A}^2\boldsymbol{B}\ \cdots\ \boldsymbol{A}^{n-1}\boldsymbol{B})$ is invertible. As part of the proof, show that reachability requires n time steps (*deadbeat control*). Hint: look explicitly at a sequence of iterates of $\boldsymbol{x}_0$.
b. Show that controllability is equivalent to reachability if $\boldsymbol{A}^{-1}$ exists. Thus, not all reachable discrete systems are controllable. Contrast with continuous systems.

c. For the undamped oscillator $\boldsymbol{A} = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, $\boldsymbol{B} = \left(\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right)$, show that the continuous system and its ZOH are controllable, except at $T_\mathrm{s} = m\pi$, for positive integer m. Why does controllability fail at these values of T_s? Why is it harder to control the oscillator when $T_\mathrm{s} = 2\pi, 4\pi, \ldots$ than when $T_\mathrm{s} = \pi, 3\pi, \ldots$?

Solution.

a. For a discrete, linear, time-invariant system with dynamical matrix $\boldsymbol{A}$ and input coupling $\boldsymbol{B}$, we can iterate explicitly from the initial condition $\boldsymbol{x}_0$, given inputs $\mathcal{U}_k \equiv \{\boldsymbol{u}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{k-1}\}$:

$$\begin{aligned}
\boldsymbol{x}_1 &= \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{B}\boldsymbol{u}_0 \\
\boldsymbol{x}_2 &= \boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{B}\boldsymbol{u}_1 = \boldsymbol{A}^2\boldsymbol{x}_0 + \boldsymbol{A}\boldsymbol{B}\boldsymbol{u}_0 + \boldsymbol{B}\boldsymbol{u}_1 \\
\boldsymbol{x}_3 &= \boldsymbol{A}\boldsymbol{x}_2 + \boldsymbol{B}\boldsymbol{u}_2 = \boldsymbol{A}^3\boldsymbol{x}_0 + \boldsymbol{A}^2\boldsymbol{B}u_0 + \boldsymbol{A}\boldsymbol{B}\boldsymbol{u}_1 + \boldsymbol{B}\boldsymbol{u}_2 \\
\boldsymbol{x}_k &= \boldsymbol{A}^k\boldsymbol{x}_0 + \boldsymbol{A}^{k-1}\boldsymbol{B}\boldsymbol{u}_0 + \boldsymbol{A}^{k-2}\boldsymbol{B}\boldsymbol{u}_1 + \cdots + \boldsymbol{B}\boldsymbol{u}_{k-1} \\
&= \boldsymbol{A}^k\boldsymbol{x}_0 + \sum_{i=0}^{k-1} \boldsymbol{A}^{k-i-1}\boldsymbol{B}\boldsymbol{u}_i \\
&= \boldsymbol{A}^k\boldsymbol{x}_0 + \boldsymbol{W}_\mathrm{k}\mathcal{U}_k\,,
\end{aligned}$$

where

$$\boldsymbol{W}_k = \begin{pmatrix} \boldsymbol{A}^{k-1}\boldsymbol{B} & \boldsymbol{A}^{k-2}\boldsymbol{B} & \cdots & \boldsymbol{A}\boldsymbol{B} & \boldsymbol{B} \end{pmatrix}.$$

If we first specialize to the case $\boldsymbol{x}_0 = \boldsymbol{0}$ and choose $k = n$, we see that an arbitrary state $\boldsymbol{x}_n$ can be reached if $\boldsymbol{W}_n$ is invertible. Indeed, the explicit n-dimensional input sequence required is just

$$\mathcal{U}_n = \boldsymbol{W}_n^{-1}\boldsymbol{x}_n\,.$$

If we iterate for $k > n$, then we can use the Cayley-Hamilton theorem to express all higher powers $\boldsymbol{A}^k$ as linear combinations of powers of $\boldsymbol{A}$ that are $\leq n$. Thus, the matrix $\boldsymbol{W}_k$ is a $k \times n$ matrix that has full rank (rank $= n$) for $n > k$, assuming that $\boldsymbol{W}_n$ is invertible. We are just adding further columns that are linear combinations. Adding these columns cannot reduce the rank and, since the new columns are linear combinations of the old, cannot extend it, either. Thus, rank $\boldsymbol{W}_k = n$ for all $k \geq n$. As a consequence, we can speak of *the* reachability matrix $\boldsymbol{W}_n \equiv \boldsymbol{W}_\mathrm{r}$.

Of course, for $k < n$, the rank cannot exceed k and is $< n$. This is a distinction from continuous systems and implies that full reachability (and controllability) requires n time steps. A control that achieves this in the minimum number of steps n is called *deadbeat* control.

Our proof implicitly assumed a SISO system with scalar u_k, even though we wrote it in vector notation, $\boldsymbol{u}_k$. To extend to a MIMO system, we note that the condition of invertibility will simply be replaced by the condition

to have full rank (n). This extension parallels the previous discussion of continuous systems.

Finally, we extend our proof to non-zero initial conditions $\boldsymbol{x}_0$. Assume that the system is reachable from $\boldsymbol{x}_0 = \boldsymbol{0}$. Then, for non-zero initial condition, we can still invert the expression for $\boldsymbol{x}_n$:

$$\mathcal{U}_n = \boldsymbol{W}_n^{-1} \left(\boldsymbol{x}_n - \boldsymbol{A}^n \boldsymbol{x}_0\right) .$$

Thus, the same reachability condition applies.

b. To see why controllability and reachability are not quite the same concepts for discrete systems, we note that if the target state is $\boldsymbol{x}_n = \boldsymbol{0}$, then we write

$$-\boldsymbol{A}^n \boldsymbol{x}_0 = \boldsymbol{W}_\mathrm{r} \mathcal{U}_n .$$

If we assume reachability, then $\boldsymbol{W}_\mathrm{r}$ is invertible. But what if $\boldsymbol{A}$ is not invertible? Then there exist non-zero initial conditions $\boldsymbol{x}_0$ such that $\boldsymbol{A}\boldsymbol{x}_0 = \boldsymbol{0}$. For those initial conditions, we need to solve

$$\boldsymbol{W}_\mathrm{r} \mathcal{U}_n = \boldsymbol{0} .$$

Because $\boldsymbol{W}_\mathrm{r}$ is invertible, the only solution has zero inputs, $\mathcal{U}_n = \boldsymbol{0}$, which obviously will not make a non-zero $\boldsymbol{x}_0$ (in the null space of $\boldsymbol{A}$) reach the origin after n time steps. Of course, if $\boldsymbol{A}$ is invertible, then we can solve, as before, for the needed non-zero input.

Why doesn't the non-invertability of $\boldsymbol{A}$ destroy reachability? In the discussion of reachability from non-zero initial conditions, a non-invertible $\boldsymbol{A}$ would also mean that there are non-zero initial conditions for which $\boldsymbol{A}\boldsymbol{x}_0 = \boldsymbol{0}$. However, we also implicitly assumed that the target state $\boldsymbol{x}_n$ was not the origin. In that case, the equation $\mathcal{U}_n = \boldsymbol{W}_n^{-1}\boldsymbol{x}_n \neq \boldsymbol{0}$. The problem arises only when $\boldsymbol{A}$ is not invertible *and* the target state is the origin.

Does this distinction carry over to the continuous case? The difference there is that an initial condition is propagated using $\mathrm{e}^{\boldsymbol{A}t}$, which always has an inverse, $\mathrm{e}^{-\boldsymbol{A}t}$. Thus, this distinction between reachability and controllability does not arise in the continuous case. In Chapter 4, we applied the term "controllability" to both situations.

c. For the undamped harmonic oscillator,

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \boldsymbol{B} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \Longrightarrow \quad \boldsymbol{W}_\mathrm{c} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

which is invertible. The system is controllable.

For the zero-order hold discretization,

$$\boldsymbol{A}_\mathrm{d} = \begin{pmatrix} \cos T_\mathrm{s} & \sin T_\mathrm{s} \\ -\sin T_\mathrm{s} & \cos T_\mathrm{s} \end{pmatrix}, \boldsymbol{B}_\mathrm{d} = \begin{pmatrix} 2\sin^2(T_\mathrm{s}/2) \\ \sin T_\mathrm{s} \end{pmatrix},$$

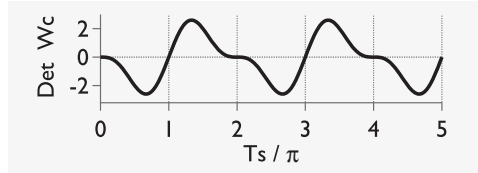
which leads to a controllability matrix,

$$\boldsymbol{W}_{\mathrm{c}} = \begin{pmatrix} 1 - \cos T_{\mathrm{s}} & \sin T_{\mathrm{s}} \\ \cos T_{\mathrm{s}} - \cos 2T_{\mathrm{s}} & (2\cos T_{\mathrm{s}} - 1)\sin T_{\mathrm{s}} \end{pmatrix}.$$

To analyze the controllability, we first compute the determinant of the controllability matrix:

$$\det \boldsymbol{W}_{\mathrm{c}} = -4\sin^2(T_{\mathrm{s}}/2)\sin T_{\mathrm{s}}\,.$$

Let's plot this as a function of sampling time:



Notice that for almost all T_{s}, the determinant is non-zero, indicating generic controllability. However it equals zero for $T_{\mathrm{s}} = m\,\pi$, for m a positive integer. (We start at time zero, although this can be dropped.) The zeros for odd m are first order: the function crosses zero with finite derivative. It is easy to Taylor expand the determinant to see that, in the vicinity of an odd integer $m = 1, 3, \ldots$, that

$$\det \boldsymbol{W}_{\mathrm{c}} = 4(T_{\mathrm{s}} - m\pi) + O(T_{\mathrm{s}} - m\pi)^3\,.$$

Near even integers $m = 2, 4, \ldots$, the expansion is, by contrast,

$$\det \boldsymbol{W}_{\mathrm{c}} = -(T_{\mathrm{s}} - m\pi)^3 + O(T_{\mathrm{s}} - m\pi)^5\,.$$

For odd m the input coupling vector is given by

$$\boldsymbol{B}_{\mathrm{d}} = \begin{pmatrix} 2\sin^2(T_{\mathrm{s}}/2) \\ \sin T_{\mathrm{s}} \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix},$$

while for even m,

$$\boldsymbol{B}_{\mathrm{d}} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

For even m, the coupling thus completely vanishes, making the system obviously uncontrollable. Physically, we inject energy at the resonance frequency, meaning that we always add energy and, thus, that we cannot force the system to an arbitrary state (remember, there is no damping).

For odd m, we look at the dynamical matrix, which is $\boldsymbol{A}_{\mathrm{d}} = -\mathbb{I}$ for such sampling times. Thus, the input coupling affects only the position and cannot control the velocity of the oscillator. At this frequency, we are forcing an even multiple of times per period.

As the dynamical matrix $\boldsymbol{A}_{\mathrm{d}}$ is just a rotation matrix, it is never singular.

5.15 Prediction observers. For the prediction observer:

a. Let the estimation error $\boldsymbol{e}_k^- \equiv \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k^-$. Show that $\boldsymbol{e}_{k+1}^- = (\boldsymbol{A} - \boldsymbol{LC})\,\boldsymbol{e}_k^-$.
b. Show that the dynamics of the observer error and the physical system decouple (*separation principle*), in analogy with Eq. (4.65).

Solution.

a.

$$\begin{aligned}
\boldsymbol{e}_{k+1}^- &= \boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}}_{k+1}^- \\
&= (\boldsymbol{A}\boldsymbol{x}_k + \cancel{\boldsymbol{B}\boldsymbol{u}_k}) - [\boldsymbol{A}\hat{\boldsymbol{x}}_k^- + \cancel{\boldsymbol{B}\boldsymbol{u}_k} + \boldsymbol{L}(y_k - \boldsymbol{C}\hat{\boldsymbol{x}}_k^-)] \\
&= \boldsymbol{A}\boldsymbol{e}_k^- + \boldsymbol{LC}\underbrace{\hat{\boldsymbol{x}}_k^-}_{\boldsymbol{x}_k - \boldsymbol{e}_k^-} - \boldsymbol{L}y_k \\
\Longrightarrow \quad &= (\boldsymbol{A} - \boldsymbol{LC})\,\boldsymbol{e}_k^- + \boldsymbol{L}\underbrace{\boldsymbol{C}\boldsymbol{x}_k}_{y_k} - \boldsymbol{L}y_k \\
&= (\boldsymbol{A} - \boldsymbol{LC})\,\boldsymbol{e}_k^-
\end{aligned}$$

b. The dynamics is

$$\boldsymbol{x}_{k+1} = \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}u_{\mathrm{k}}\,,$$

with feedback $u_k = -\boldsymbol{K}\hat{\boldsymbol{x}}_k^- = -\boldsymbol{K}(\boldsymbol{x}_k - \boldsymbol{e}_k^-)$. This gives coupled equations

$$\boldsymbol{x}_{k+1} = \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}(\boldsymbol{K}\boldsymbol{e}_k - \boldsymbol{K}\boldsymbol{x}_k) = (\boldsymbol{A} - \boldsymbol{BK})\boldsymbol{x}_k + \boldsymbol{BK}\boldsymbol{e}_k^-\,.$$

Using the result from part (a) then gives,

$$\begin{aligned}
\boldsymbol{x}_{k+1} &= (\boldsymbol{A} - \boldsymbol{BK})\boldsymbol{x}_k + \boldsymbol{BK}\boldsymbol{e}_k^- \\
\boldsymbol{e}_{k+1}^- &= (\boldsymbol{A} - \boldsymbol{LC})\,\boldsymbol{e}_k^-\,.
\end{aligned}$$

Putting the two equations into a single matrix notation gives

$$\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{e}^- \end{pmatrix}_{k+1} = \begin{pmatrix} \boldsymbol{A} - \boldsymbol{BK} & \boldsymbol{BK} \\ \boldsymbol{0} & \boldsymbol{A} - \boldsymbol{LC} \end{pmatrix} \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{e}^- \end{pmatrix}_k\,.$$

As in the discussion in Chapter 4, the characteristic equation is given by

$$\det\,(s\mathbb{I} - \boldsymbol{A} + \boldsymbol{BK})\,\det\,(s\mathbb{I} - \boldsymbol{A} + \boldsymbol{LC}) = 0\,,$$

which means that the individual determinants should each vanish separately. The first term represents feedback for the system dynamics, the second the observer dynamics. We see here that they can be designed separately (independent choice of $\boldsymbol{K}$ and $\boldsymbol{L}$).

5.16 Current vs. prediction observers. The prediction observer state vector, $\hat{\boldsymbol{x}}_{k+1}^-$, is based on observations up to time k (Eq. (4.65)). Here, we construct an estimator that is based on observations up to time $k + 1$. First: Given the old estimate $\hat{\boldsymbol{x}}_k$, we predict the next state: $\hat{\boldsymbol{x}}_{k+1}^- = \boldsymbol{A}\hat{\boldsymbol{x}}_k + \boldsymbol{B}u_k$. Then we correct the estimate using the difference between the new observation y_{k+1} and its predicted value,

$\hat{y}_{k+1} = \boldsymbol{C}\hat{\boldsymbol{x}}_{k+1}^-$. Thus, $\hat{\boldsymbol{x}}_{k+1} = \hat{\boldsymbol{x}}_{k+1}^- + \boldsymbol{L}(y_{k+1} - \hat{y}_{k+1})$, with $\boldsymbol{L}$ the observer gain. For the current observer,

a. Show that $\begin{pmatrix}\boldsymbol{x}\\ \hat{\boldsymbol{x}}\end{pmatrix}_{k+1} = \begin{pmatrix}\boldsymbol{A} & \boldsymbol{0}\\ \boldsymbol{LCA} & \boldsymbol{A}-\boldsymbol{LCA}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ \hat{\boldsymbol{x}}\end{pmatrix}_k + \begin{pmatrix}\boldsymbol{B}\\ \boldsymbol{B}\end{pmatrix}u_k$.

b. Define the error $\boldsymbol{e}_k \equiv \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k$, and show that $\boldsymbol{e}_{k+1} = (\boldsymbol{A} - \boldsymbol{LCA})\,\boldsymbol{e}_k$.

c. Reproduce the margin plots for y_k and $\hat{y}_k$ in Section 5.4.2 using the parameters given in the caption. Plot $\hat{y}_k$ for both current and prediction observers. Explore the output behavior for different estimator gains $\boldsymbol{L}$. Why are there more problems with large gains for the prediction observer than for the current observer?

Solution.

a. The dynamical equation for the state vector is

$$\boldsymbol{x}_{k+1} = \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}u_k\,.$$

Then

$$\begin{aligned}\hat{\boldsymbol{x}}_{k+1} &= \hat{\boldsymbol{x}}_{k+1}^- + \boldsymbol{L}(y_{k+1} - \boldsymbol{C}\hat{\boldsymbol{x}}_{k+1}^-)\\ &= (\mathbb{I} - \boldsymbol{LC})\hat{\boldsymbol{x}}_{k+1}^- + \boldsymbol{L}y_{k+1}\\ &= (\mathbb{I} - \boldsymbol{LC})(\boldsymbol{A}\hat{\boldsymbol{x}}_k + \boldsymbol{B}u_k) + \boldsymbol{LC}\boldsymbol{x}_{k+1}\\ &= (\boldsymbol{A} - \boldsymbol{LCA})\hat{\boldsymbol{x}}_k + (\boldsymbol{B} - \cancel{\boldsymbol{LCB}})u_k + \boldsymbol{LC}(\boldsymbol{A}\boldsymbol{x}_k + \cancel{\boldsymbol{B}u_k})\\ &= (\boldsymbol{A} - \boldsymbol{LCA})\hat{\boldsymbol{x}}_k + \boldsymbol{B}u_k + \boldsymbol{LCA}\boldsymbol{x}_k\,.\end{aligned}$$

Putting these two equations together in matrix form gives

$$\begin{pmatrix}\boldsymbol{x}\\ \hat{\boldsymbol{x}}\end{pmatrix}_{k+1} = \begin{pmatrix}\boldsymbol{A} & \boldsymbol{0}\\ \boldsymbol{LCA} & \boldsymbol{A}-\boldsymbol{LCA}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ \hat{\boldsymbol{x}}\end{pmatrix}_k + \begin{pmatrix}\boldsymbol{B}\\ \boldsymbol{B}\end{pmatrix}u_k\,.$$

b.

$$\begin{aligned}\boldsymbol{e}_{k+1} &= \boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}}_{k+1}\\ &= (\boldsymbol{A}\boldsymbol{x}_k + \cancel{\boldsymbol{B}u_k}) - [\boldsymbol{A}\hat{\boldsymbol{x}}_l + \cancel{\boldsymbol{B}u_k} + \boldsymbol{L}(y_{k+1} - \boldsymbol{C}\underbrace{\hat{\boldsymbol{x}}_{k+1}^-}_{\boldsymbol{A}\hat{\boldsymbol{x}}_k + \boldsymbol{B}u_k})\,]\\ &= \boldsymbol{A}\boldsymbol{e}_k + \boldsymbol{LC}(\boldsymbol{A}\underbrace{\hat{\boldsymbol{x}}_k}_{\boldsymbol{x}_k - \boldsymbol{e}_k} + \boldsymbol{B}u_k) - \boldsymbol{L}y_{k+1}\\ \Longrightarrow\ &= (\boldsymbol{A} - \boldsymbol{LCA})\,\boldsymbol{e}_k + \boldsymbol{L}\underbrace{\boldsymbol{C}(\boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}u_k)}_{y_{k+1}} - \boldsymbol{L}y_{k+1}\\ &= (\boldsymbol{A} - \boldsymbol{LCA})\,\boldsymbol{e}_k\end{aligned}$$

Because this is the same formula that we derived for the prediction observer in Problem 5.15, we find in this case, too, that the separation principle holds. We can decouple the design of the controller from that of the observer:

$$\begin{aligned}\boldsymbol{x}_{k+1} &= (\boldsymbol{A} - \boldsymbol{BK})\boldsymbol{x}_k + \boldsymbol{BK}\boldsymbol{e}_k\\ \boldsymbol{e}_{k+1} &= (\boldsymbol{A} - \boldsymbol{LC})\,\boldsymbol{e}_k\,.\end{aligned}$$

Putting the two equations into a single matrix notation gives

$$\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{e} \end{pmatrix}_{k+1} = \begin{pmatrix} \boldsymbol{A} - \boldsymbol{BK} & \boldsymbol{BK} \\ \boldsymbol{0} & \boldsymbol{A} - \boldsymbol{LC} \end{pmatrix} \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{e} \end{pmatrix}_{k},$$

so that the state dynamics are influenced by $\boldsymbol{K}$ and estimator convergence by $\boldsymbol{L}$.

c. The prediction observer has a delay, which leads to instability at high gains.

5.17 Fractional delays. For a linear system, fractional delays affect two neighboring time points. For example, consider $\dot{x} = -x(t) + u(t - \tau)$, with $0 < \tau < T_s$. Let the input $u_{k-\tau}(t)$ be a staircase ZOH signal delayed by τ with respect to the state x_k.

a. Draw a timing diagram for $x(t)$ and $u(t)$. Indicate x_k and $u_{k-\tau}$.

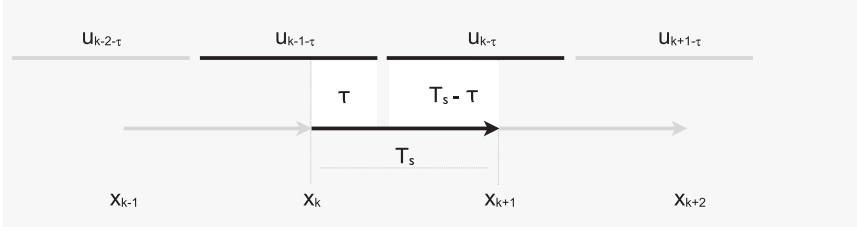
b. Show that the discrete dynamics have the form $x_{k+1} = Ax_k + B_1 u_{k-1} + B_0 u_k$. Find A, B_1, and B_0. Check the limits $\tau \to 0$ and $\tau \to T_s$. Hint: Split the integral.

c. Redo (b) assuming $T_s < \tau < 2T_s$. (Hint: the coefficients are almost the same.)

Other delays can be treated similarly (e.g., a delay between two subsystems can be analyzed as a delayed input to the second subsystem). Finally, another approach uses the *modified Z-transform*, $F(z, m) \equiv \sum_{k=0}^{\infty} f[(k + m - 1)T_s]z^{-k}$, with $0 < m < 1$.

Solution.

a. In the diagram below, we see that the forces during the interval from kT_s to $(k + 1)T_s$ are u_{k-1} for the first part and u_k for the second.



b. We integrate $\dot{x} = -x(t) + u(t - \tau)$ from kT_s to $(k + 1)T_s$:

$$x_{k+1} = \mathrm{e}^{(-1)T_s} x_k + \int_{kT_s}^{(k+1)T_s} \mathrm{d}t\, \mathrm{e}^{-[(k+1)T_s - t]}\, u(t - \tau),$$

so that $A = \mathrm{e}^{-T_s}$. To find the discrete input, we write the second term as

$$\begin{aligned} &= \int_{kT_s}^{kT_s+\tau} \mathrm{d}t\, \mathrm{e}^{-[(k+1)T_s - t]}\, u_{k-1} + \int_{kT_s+\tau}^{(k+1)T_s} \mathrm{d}t\, \mathrm{e}^{-[(k+1)T_s - t]}\, u_k, \\ &= \mathrm{e}^{-T_s} \int_0^{\tau} \mathrm{d}t\, \mathrm{e}^{t}\, u_{k-1} + \mathrm{e}^{-(T_s - \tau)} \int_0^{T_s - \tau} \mathrm{d}t\, \mathrm{e}^{t}\, u_k \\ &= \mathrm{e}^{-T_s}(\mathrm{e}^{\tau} - 1)\, u_{k-1} + (1 - \mathrm{e}^{-(T_s - \tau)})\, u_k. \end{aligned}$$

Thus, $B_1 = \mathrm{e}^{-T_\mathrm{s}}(\mathrm{e}^{\tau} - 1)$ and $B_0 = (1 - \mathrm{e}^{-(T_\mathrm{s}-\tau)})$. Then,

- Taking the limit $\tau \to 0$, we have $B_1 = 0$ and $B_0 = 1 - \mathrm{e}^{-T_\mathrm{s}}$.
- Taking the limit $\tau \to T_\mathrm{s}$, we have $B_1 = 1 - \mathrm{e}^{-T_\mathrm{s}}$ and $B_0 = 0$.

c. For $T_\mathrm{s} < \tau < 2T_\mathrm{s}$, we write $\tau = T_\mathrm{s} + \tau'$, with $0 < \tau' < T_\mathrm{s}$. Then we do the same calculation to find

$$\int_{kT_\mathrm{s}}^{(k+1)T_\mathrm{s}} \mathrm{d}t\, \mathrm{e}^{-[(k+1)T_\mathrm{s}-t]}\; u(t - T_\mathrm{s} - \tau') = B_1\, u_{k-2} + B_0\, u_{k-1}\,,$$

as before, with $B_1 = \mathrm{e}^{-T_\mathrm{s}}(\mathrm{e}^{\tau'} - 1)$ and $B_0 = (1 - \mathrm{e}^{-(T_\mathrm{s}-\tau')})$. We get the same fractional contributions from the old and the new. We just have to displace the inputs to the appropriate integer delays (here 2 and 1 instead of 1 and 0).

5.18 Delays and predictive feedback.

a. For $x_{k+1} = ax_k + u_k$, with $u_k = -K_\mathrm{p}x_k$, find the range of K_p that stabilizes $x = 0$.
b. For delayed proportional feedback $u_k = -K_\mathrm{p}x_{k-1}$, show that $a > 2$ implies that no value of K_p can stabilize $x = 0$.
c. Show that $u_k = -K_\mathrm{p}x_k^{\mathrm{pred}}$, with $x_k^{\mathrm{pred}} = ax_{k-1} + u_{k-1}$, can stabilize $x = 0$ for all a.

Solution.

a. The closed-loop dynamics are

$$x_{k+1} = ax_k + u_k = (a - K_\mathrm{p})x_k\,,$$

and the range of stability is

$$-1 < (a - K_\mathrm{p}) < +1\,, \quad \Longrightarrow \quad (a-1) < K_\mathrm{p} < (a+1)\,.$$

b. The closed-loop, delayed-feedback dynamics are

$$x_{k+1} = ax_k + u_k = ax_k - K_\mathrm{p}x_{k-1}\,,$$

Taking the Z-transform gives, for $x(z)$,

$$zx = ax - K_\mathrm{p}z^{-1}x\,, \quad \Longrightarrow \quad z^2 - az + K_\mathrm{p} = 0\,.$$

Solving the quadratic equation, we find roots at

$$z_\pm = \tfrac{1}{2}\left(a \pm \sqrt{a^2 - 4K_\mathrm{p}}\right).$$

The condition for stability for this discrete system is $|z| < 1$. If $a^2 - 4K_\mathrm{p} > 0$, then there are two real roots, one greater than $\frac{1}{2}a$ and one less than this value. So for $a > 2$, one of the z roots must have magnitude greater than 1.

If $a^2 - 4K_\mathrm{p} < 0$, then the roots are a complex-conjugate pair, with real part $\frac{1}{2}a$. Again, $a < 2$, and no value of K_p stabilizes $x = 0$ when $a > 2$.

c. We now have to solve two coupled difference equations:

$$x_{k+1} = ax_k + u_k\,, \qquad u_{k+1} = -K_\mathrm{p}(\underbrace{ax_k + u_k}_{x_{k+1}^{\mathrm{pred}}})\,.$$

Taking the Z-transform gives, for $x(z)$ and $u(z)$,

$$zx = ax + u\,, \qquad zu = -K_\mathrm{p}\,(ax + u)\ .$$

Let's put these together in matrix form:

$$\begin{pmatrix} z-a & -1 \\ K_\mathrm{p}a & z+K_\mathrm{p} \end{pmatrix}\begin{pmatrix} x \\ u \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

The determinant of the matrix then must be zero:

$$(z-a)(z+K_\mathrm{p}) + K_\mathrm{p}a = z^2 - \left(a - K_\mathrm{p}\right)z - \cancel{K_\mathrm{p}a} + \cancel{K_\mathrm{p}a} = 0\,,$$

which implies

$$z = 0,\, \left(a - K_\mathrm{p}\right)\ .$$

The latter is just the condition that we had for a system without delay, implying that, for any $a > 1$, we can choose $a - 1 < K_\mathrm{p} < a + 1$. Note that, in comparing to the continuous system, choosing a for fixed discrete delay T_s is equivalent to having a growth exponent scaled to 1 and varying the continuous delay τ.

5.19 Deadbeat control of an undamped oscillator. Derive the deadbeat controller $K_\mathrm{d}(z)$ described in Section 5.4.2 and Example 5.10. Reproduce the graphs in the example. Hint: for a step, $r(z) = \frac{z}{z-1}$. Use the final value theorem for y_k (or inverse transform).

Solution.

a. We begin by deriving the deadbeat controller for an nth-order, SISO linear system. We proceed in three steps:

i. For $y(z) = T_\mathrm{d}(z)\,r(z)$, let us assume that the complementary sensitivity function is of the form $T_\mathrm{d}(z) = \frac{f(z)}{z^n}$. Then, for a step command,

$$y(z) = T_\mathrm{d}(z)\,r(z) = \frac{f(z)}{z^n}\,\frac{z}{z-1} = \frac{f(z)}{(z-1)z^{n-1}}\,.$$

The final value theorem then gives

$$\lim_{k\to\infty} y_k = \lim_{z\to 1}(z-1)y(z) = \lim_{z\to 1}\frac{f(z)}{z^{n-1}} = f(1)\,.$$

We can also look at the initial value theorem:

$$y_0 = \lim_{z\to\infty} y(z) = \lim_{z\to\infty}\frac{f(z)}{z^n}\,.$$

Because our initial condition is $y_0 = 0$, we can infer that if $f(z)$ is a polynomial, its order cannot be higher than $n - 1$.

ii. Next, we look at the input u_k that is generated by the reference command. From the usual block-diagram manipulations, [direct / (1+loop)], we have

$$u(z) = \frac{K_\mathrm{d}}{1 + K_\mathrm{d} G_\mathrm{d}}\, r(z) = T_\mathrm{d} G_\mathrm{d}^{-1}\, r(z)\,.$$

Now we apply this to deadbeat control, with r a step command:

$$u(z) = T_\mathrm{d} G_\mathrm{d}^{-1}\, r(z) = \frac{f(z)}{z^n}\,\frac{D(z)}{N(z)}\,\frac{z}{z-1}\,.$$

As suggested in the text, we choose $f(z) = N(z)/N(1)$. The normalization $N(1)^{-1}$ is needed to make $f(1) = 1$. In this case,

$$u(z) = \frac{1}{z-1}\,\frac{D(z)}{N(1) z^{n-1}}\,.$$

The final value theorem then implies that

$$\lim_{k\to\infty} u_k = \frac{D(1)}{N(1)} = \frac{1}{G_\mathrm{d}(1)}\,,$$

which is just what we expect: at zero frequency, the steady-state output to a constant input $u = 1/G_\mathrm{d}(1)$ generates the output

$$y = G_\mathrm{d}(1)\, u = G_\mathrm{d}(1) \frac{1}{G_\mathrm{d}(1)} = 1\,.$$

iii. Finally, we derive the controller. From $T_\mathrm{d} = \frac{K_\mathrm{d} G_\mathrm{d}}{1 + K_\mathrm{d} G_\mathrm{d}}$, we have

$$K_\mathrm{d}(z) = \frac{G_\mathrm{d}^{-1}}{T_\mathrm{d}^{-1} - 1} = \frac{D(z)}{N(z)}\,\frac{1}{\frac{z^n N(1)}{N(z)} - 1} = \frac{D(z)}{z^n N(1) - N(z)}\,.$$

b. From Problem 5.10, the ZOH of the second-order system $G(s) = \frac{1}{s^2+1}$ is

$$G_\mathrm{d}(z) = \frac{(1 - \cos T_\mathrm{s})(z+1)}{(z - \mathrm{e}^{\mathrm{i}T_\mathrm{s}})(z - \mathrm{e}^{-\mathrm{i}T_\mathrm{s}})}\,.$$

For $T_\mathrm{s} = 0.2$, we have

$$G_\mathrm{d}(z) \approx 0.0199 \frac{z+1}{z^2 - 1.960 z + 1}\,,$$

with poles at $p_\pm = \cos T_\mathrm{s} \pm \mathrm{i} \sin T_\mathrm{s} \approx 0.980 \pm 0.199 i$, on the unit circle. In the notation of the previous section,

$$N(z) = (1 - \cos T_\mathrm{s})(z+1) \quad \text{and} \quad N(1) = 2(1 - \cos T_\mathrm{s})\,,$$

and the denominator is $D(z) = z^2 - 2\cos T_\mathrm{s} z + 1$. The discrete transfer function leads to a controller

$$K_\mathrm{d}(z) = \frac{D(z)}{z^2 N(1) - N(z)} = \left(\frac{1}{1 - \cos T_\mathrm{s}}\right) \frac{z^2 - 2\cos T_\mathrm{s} z + 1}{(2z+1)(z-1)}\,.$$

We use this controller, with $T_\mathrm{s} = 0.2$, to calculate the response in Example 5.10.

5.20 Feedforward gain. For a steady-state output $y = r$, you can add an offset $k_\mathrm{r} r$ to the input u. Show that $k_\mathrm{r} = [\boldsymbol{C}(\mathbb{I} - \boldsymbol{A} + \boldsymbol{B}\boldsymbol{K})^{-1}\boldsymbol{B}]^{-1}$ for a discrete system. Cf. Eq. 4.47.

Solution.

$$\boldsymbol{x}_{k+1} = (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K})\boldsymbol{x}_k + \boldsymbol{B}k_\mathrm{r} r_k$$

Then, at steady state $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k \equiv \boldsymbol{x}$, and

$$\boldsymbol{x}(\mathbb{I} - \boldsymbol{A} + \boldsymbol{B}\boldsymbol{K}) = \boldsymbol{B}k_\mathrm{r} r_k \,,$$

and

$$\boldsymbol{x} = (\mathbb{I} - \boldsymbol{A} + \boldsymbol{B}\boldsymbol{K})^{-1}\boldsymbol{B}k_\mathrm{r} r_k \,.$$

Finally,

$$\begin{aligned} y_k = \boldsymbol{C}\boldsymbol{x} &= \boldsymbol{C}(\mathbb{I} - \boldsymbol{A} + \boldsymbol{B}\boldsymbol{K})^{-1}\boldsymbol{B}k_\mathrm{r} r_k = r_k \,, \\ \implies k_\mathrm{r} &= \frac{+1}{\boldsymbol{C}(\mathbb{I} - \boldsymbol{A} + \boldsymbol{B}\boldsymbol{K})^{-1}\boldsymbol{B}} \,. \end{aligned}$$

5.21 Feedforward control of an oscillator. For the feedforward filter in Figure 5.13,

a. Implement numerically the feedforward control and reproduce the nine graphs. Design a feedforward filter by inverting the denominator and adding poles at zero. Produce the continuum response using the hybrid procedure of Section 5.4.3. For $T_\mathrm{s} = \frac{\pi}{2}$ and $\frac{1}{2}$, find $G_\mathrm{d}(z)$, the parameter λ, and the feedforward $F(z)$.
b. The z^n in the denominator of a feedforward filter means that we can write it as an FIR filter with delay: $F(z) = F_0 + F_1 z^{-1} + F_2 z^{-2} \cdots$. In the time domain, this is $u_k = F_0 r_k + F_1 r_{k-1} + F_2 r_{k-2} \cdots$. Put the transfer functions from (a) in this form, and show that they transform the reference r_k into the desired "shaped-input" u_k.
c. Show that the input amplitude $\sim [4\sin^2(\omega T_\mathrm{s}/2)]^{-1}$, where $\omega = 1$ is the angular frequency of the oscillator. The factor is $\sim (\omega T_\mathrm{s})^{-2}$ as $T_\mathrm{s} \to 0$. The ω^{-2} dependence mirrors the high-frequency response of the original transfer function.
d. Investigate numerically the impact of oscillator damping. Plot the shaped inputs and both discrete and continuous outputs for a reference step, for $\zeta = \{0, 0.4, 1\}$.

Solution.

a. From Problem 5.10, the ZOH discretization of $G(s) = \frac{1}{s^2+1}$ is

$$G_\mathrm{d}(z) = \frac{(1 - \cos T_\mathrm{s})(z + 1)}{z^2 - 2(\cos T_\mathrm{s})z + 1}$$

For $T_s = \frac{\pi}{2}$, we have

$$G_d(z) = \frac{z+1}{z^2+1}, \qquad \lambda = 2, \qquad F(z) = \frac{z^2+1}{2z^2}.$$

For $T_s = \frac{1}{2}$, we have (to two decimal places)

$$G_d(z) = \frac{0.12\,(z+1)}{z^2 - 1.76\,z + 1}, \qquad \lambda = 0.24, \qquad F(z) = \frac{z^2 - 1.76\,z + 1}{0.24\,z^2}.$$

Notice that in going from $T_s = \pi/2 = 1.57$ to 0.5, the scale factor in the feedforward filter increased a factor $2/0.24 \approx 8.2$.
See Mathematica notebook for more details.

b. *FIR forms*. From part (a), we read off, for $T_s = \pi/2$,

$$u_k = 0.5 r_k + 0.5 r_{k-2}$$

and, for $T_s = 0.5$,

$$u_k = 4.1 r_k - 7.2 r_{k-1} + 4.1 r_{k-2}$$

We verify that these produce the desired u_k in Figure 5.13 and that these system inputs lead to the correct outputs y_k.

c. *Amplitude-speed tradeoff.* Going back from scaled units, we have $T_s \to \omega T_s$ and

$$\begin{aligned} G_d(z) &= (1 - z^{-1})\left(\frac{1}{\omega^2}\right)\mathcal{Z}\,(1 - \cos \omega k T_s) \\ &= (1 - z^{-1})\left(\frac{1}{\omega^2}\right)\left[\frac{1}{(1 - z^{-1})} - \frac{1 - z^{-1}\cos \omega T_s}{1 - 2z^{-1}\cos \omega T_s + z^{-2}}\right] \\ &= \left(\frac{1}{\omega^2}\right)\frac{(1 - \cos \omega T_s)(1+z)}{1 + z^2 - 2z\cos(\omega T_s)}. \end{aligned}$$

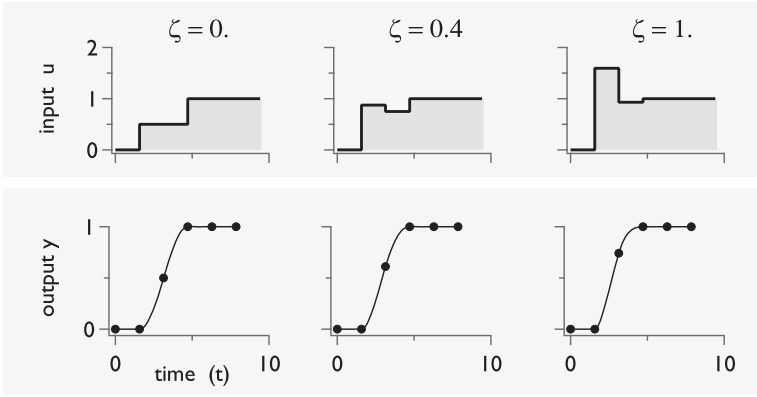
In constructing a feedforward filter, our rule is to extract the denominator of $\text{den}(z) = G_d(z)$ and evaluate $\lambda = \text{den}(z = 1)$. The scale factor is the inverse of λ. Here, this gives us

$$\lambda^{-1} = \left[1 + z^2 - 2z\cos(\omega T_s)\right]_{z=1} = 2[1 - \cos(\omega T_s)] = 4\sin^2\left(\frac{\omega T_s}{2}\right) \approx (\omega T_s)^2.$$

Thus, the required scaling amplitude diverges as $\omega T_s \to 0$. This increase in amplitude is the price we pay for making our system go faster than it "wants to." (The easiest frequency, of course, is the natural frequency of the system, $\omega = 2\pi f$.)

d. *Damping*. It is easy to add damping (numerically). We simply go through the above procedure with $G(s) = \frac{1}{1+2\zeta s+s^2}$. Qualitatively, damping increases the needed amplitudes and breaks the symmetry of the pulse waveforms that you may have observed in part (a).

The three damping cases for $r(t) = \theta(t - T_s)$, with $T_s = \pi/2$, are shown below.



5.22 Deadbeat control of an undamped oscillator in one time step? Try to make a one-step deadbeat controller for $G(s) = \frac{1}{s^2+1}$ by enforcing $T_d(z) = \frac{1}{z}$.

a. Show that the required controller has the form $K_d(z) = \frac{1}{1-\cos T_s} \frac{z^2-2\cos T_s z+1}{(z+1)(z-1)}$.
b. Reproduce the margin figure in Section 5.4.3.
c. The oscillatory step response arises because a controller pole cancels the sampling zero at -1. To see the problem in a simpler context, compare transfer functions $G_1(z) = 1$ and $G_2(z) = \frac{z-a}{z-a}$. Compute the output y_k given an initial condition y_0.

Solution.

a. We recall again the ZOH discretization of the oscillator:

$$G_d(z) = \frac{(1 - \cos T_s)(z + 1)}{z^2 - 2(\cos T_s)z + 1}$$

With $T_d(z) = \frac{1}{z}$, the discrete controller is

$$K_d(z) = \frac{G_d^{-1}}{T_d^{-1} - 1} = \frac{z^2 - 2(\cos T_s)z + 1}{(1 - \cos T_s)(z + 1)} \left(\frac{1}{z - 1}\right) = \left(\frac{1}{1 - \cos T_s}\right) \frac{z^2 - 2\cos T_s z + 1}{z^2 - 1}.$$

The system has poles on the unit circle (cf. Problem 5.10d) that the controller tries to "cancel out."

b. The controller leads to the dynamics sketched in the problem, for $T_s = 0.2$. See code on book website.

c. In the time domain, $G_1(z) = 1$ converts to the time-domain equation $y_k = u_k$ (for no input). For $G_2(z) = \frac{z-a}{z-a}$, we have

$$y_{k+1} - a y_k = u_{k+1} - u_k \,.$$

The dynamics in case 1 is simply $y_k = y_0 + u_k$. But for the nominally identical case 2, it is

$$y_k = a^k y_0 + u_k \,.$$

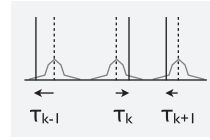
For $|a| < 1$, the term due to the initial value y_0 decays to zero, and the two dynamics are the same. For $|a| = 1$ (applied to the one step deadbeat control, where the pole-zero combination is at -1), the initial condition will be preserved in amplitude (and oscillate with period two when $= -1$). For $|a| > 1$, the output will diverge.

The conclusion is that one should avoid controllers that add poles to cancel system zeros, especially when they are on or outside the unit circle. Notice that both the deadbeat controller attempt given here and the "true" one given earlier introduce a pole at $z = +1$. This leads to problems for input disturbances at $\omega = 1/T_s$, which are then not damped.

System Identification 6

Problems

6.1 **Timing jitter**. Fluctuations in sampling a signal lead to low-pass filtering and distort the apparent transfer function. The phenomenon is similar to the Debye-Waller factor for X-ray scattering from crystal lattices at finite temperatures. To see this, we follow Souders et al. (1990) and consider timing fluctuations $\tau_k \sim p(\tau)$ at time step k, as illustrated at right. Define the jitter signal $f_j(t) \equiv \langle f(t+\tau) \rangle$, where the angle brackets denote an ensemble average over $p(\tau)$, which we assume to be even in τ.



a. Show that jitter acts as a convolution and hence that the continuous-time Fourier transform $f_j(\omega) = f(\omega)\,\varphi_\tau(\omega)$, where $f(\omega)$ is the Fourier transform of the original signal and $\varphi_\tau(\omega)$ is the characteristic function (Fourier transform) of $p(\tau)$.

b. Consider a measurement with sampling at nominal times kT_s. Because of jitter, the actual measurement times are at $kT_s + \alpha_k T_s$, where $\alpha_k \sim \mathcal{N}(0, \alpha^2)$. Show that timing jitter limits the bandwidth to $\omega_b = \omega_s/(\sqrt{2}\pi\alpha)$, with $\omega_s = 2\pi/T_s$.

Solution.

a. The jitter function $f_j(t)$ is defined to be

$$\begin{aligned} f_j(t) &= \int_{-\infty}^{\infty} \mathrm{d}\tau\, f(t+\tau)\, p(\tau) \\ &= \int_{\infty}^{-\infty} \mathrm{d}(-\tau)\, f(t-\tau)\, p(-\tau)\,, \qquad \text{substituting } \tau \to -\tau \\ &= \int_{-\infty}^{\infty} \mathrm{d}\tau\, f(t-\tau)\, p(\tau)\,, \qquad \text{since } p(\tau) = p(-\tau) \\ &= [f * \varphi_\tau](t)\,. \end{aligned}$$

Thus, the expected effect of jitter is to act as a convolution on the original function $f(t)$. Then, the convolution theorem for Fourier transforms gives

$$f_j(\omega) = f(\omega)\,\varphi_\tau(\omega)\,.$$

b. For Gaussian jitter with standard deviation $\tau = \alpha T_s$, the characteristic function of the normal distribution $\mathcal{N}(0, \tau^2)$ is given by

$$\begin{aligned}\frac{1}{\sqrt{2\pi\tau^2}} \int_{-\infty}^{\infty} d\tau_k \exp\left(-\tfrac{\tau_k^2}{2\tau^2}\right) e^{i\omega\tau_k} &= \exp\left\{\left(-\tfrac{1}{2}\tau^2\omega^2\right)\right\} \\ &= \exp\left\{\left(-\tfrac{1}{2}\alpha^2 T_s^2\omega^2\right)\right\} \\ &= \exp\left\{\left(-\frac{\alpha^2 2\pi^2\omega^2}{\omega_s^2}\right)\right\}.\end{aligned}$$

The values of the sampled spectrum with no jitter are then multiplied by the number $\exp\left\{\left(-\frac{\alpha^2 2\pi^2\omega^2}{\omega_s^2}\right)\right\}$, which reduces the amplitude at a given frequency and thereby restricts the bandwidth to roughly $\omega_b = \omega_s / \left(\sqrt{2}\pi\alpha\right) \approx \omega_s/(4.4\alpha)$.

As hinted in the problem statement, timing jitter is a kind of one-dimensional version of the problem of X-ray scattering from crystal lattices. The result gives a useful perspective on the common interpretation of the Debye-Waller factor as implying that thermal motion of atoms broadens diffraction peaks in X-ray scattering. It doesn't. Rather, we see here that the proper statement is that the amplitude of the peaks is lowered, while the width remains unchanged. This is the low-pass filtering due to the "jitter" of atomic motion.

6.2 **Crest factor**. The *crest factor* $\mathrm{Cr}[u(t)]$ measures the maximum amplitude of a signal for a given RMS power. For a waveform of period τ, it is defined as

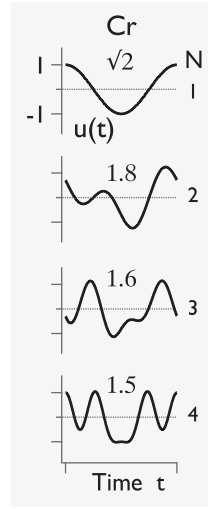
$$\mathrm{Cr}[u] = \frac{u_{\max}}{u_{\mathrm{rms}}}, \qquad u_{\max} = \max_{0\le t\le\tau} |u(t)| \qquad u_{\mathrm{rms}} = \sqrt{\frac{1}{\tau}\int_0^\tau dt\, u^2(t)}\,.$$

A good input signal should have a small crest factor, to inject power into a system while keeping the maximum amplitude low enough to avoid a nonlinear response.

a. Elementary cases: Show that a square wave has Cr = 1, a single sine has Cr = $\sqrt{2}$, and a Dirac delta function has Cr = ∞.
b. Multisine signals. Consider periodic signals $u_N(t)$ of period T that are the sum of N harmonics. With $\omega_n = 2\pi n f_0 = 2\pi n/\tau$, we have $u_N(t) = \sum_{n=1}^N A_n \cos(\omega_n t + \varphi_n)$. Show that $u_{\mathrm{rms}}^2 = \frac{1}{2}\sum_{n=1}^N A_n^2$, independent of the value of the phases φ_n. Set $A_n = 1/\sqrt{N}$, so that $u_{\mathrm{rms}} = 1/\sqrt{2}$, and the crest function depends only on $u_{\max}$. Set also $\varphi_1 = 0$ by overall translational invariance and $f_0 = \tau = 1$ for convenience.
c. Fast numerical calculations. Show that you can vastly speed up the explicit sum for $u_N(t)$ by defining the signal in the Fourier domain and taking the inverse Fourier transform. Choose N_s, the total number of points in a period of the waveform, to be a power of 2. Why is the Fourier-method much faster than the time-domain sum? Write a program to calculate the waveform using the two methods. The two waveforms should agree to within machine precision. For $N_s = 1024$ and $N = 255$, show that the speed-up is $\approx$ 100-fold.

d. N cosines, done wrong. For N cosines, the worst choice is to set $\varphi_n = 0$. Plot the $N = 2, 3, 4$ cases and show that $\mathrm{Cr} = \sqrt{2N}$.

e. Small number of harmonics. We can use brute-force numerics to find the optimum phases. Search an $N - 1$ dimensional grid for all possible phase values. Obviously, the time to solve the problem grows exponentially, but small problems can be solved easily on a laptop. Your solutions for $N \leq 4$ should resemble the graphs at right. Note how, for $N \geq 2$, the optimal Cr decreases with N. Give the phases of the waves and state Cr with more precision.



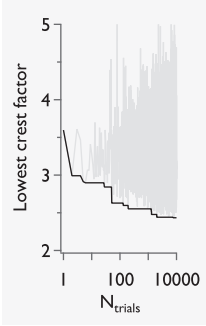
f. Random phases: In the limit of a large number of harmonics, $N \to \infty$, one idea is to choose phases randomly from a uniform distribution between 0 and 2π. In this part, we investigate the properties of the average crest factor. In particular, we will see that $\langle \mathrm{Cr} \rangle \approx \sqrt{2 \ln(2N)}$, a number that is 3–4 for typical values N (say 100 to 1000) and varies little with N in this range.

   i. The $N = 1$ case corresponds to a single cosine. Clearly, for $N = 1$ and $A_1 = 1$, we have $u_{\max} = 1$ for all choices of φ_1. But if we choose φ_1 randomly and look at a random time t, what is the probability density $p_1(u)$? Argue that this is equivalent to picking a random angle θ from $(0, 2\pi)$. Use the change-of-variables formalism for probability distributions to show $p_1(u) = \frac{1}{\pi} \frac{1}{1-u^2}$. Verify that $\langle u \rangle = 0$ and $\langle u^2 \rangle = \frac{1}{2}$, consistent with the result in Figure 6.3b.

   ii. For N harmonics of amplitude $1/\sqrt{N}$, use the Central Limit Theorem (Appendix A.7.3) to show that $\lim_{N\to\infty} p_N(u) \sim \mathcal{N}(0, \frac{1}{2})$ (Figure 6.3b).

   iii. *Extreme Value Statistics*. If we draw M times from a probability distribution $p(u)$, what is the typical value for $u_{\max}$? Let $F(u) = \int_{-\infty}^{u} \mathrm{d}u'\ p(u')$ be the cumulative distribution for $p(u)$. The probability to draw a value greater than u from $p(u)$ is $1 - F(u)$. Argue that the typical largest absolute value $u_{\max}$ in M draws from $p(u)$ is given by the solution to the equation $2M[1 - F(u_{\max})] = 1$. Derive the transcendental equation for M (involves erfc). What value to take for M? Since $u_N(t)$ has frequencies up to Nf_0, we must sample several times the fastest period in order to see the maximum (16 samples per period determines the maximum to better than 1%). But if we sample too often, the maximum value will saturate, meaning that not all draws are independent. Empirically, the maximum number of effectively independent draws is about $M = 5N$.

   iv. Confirm via simulation the results of these calculations shown below. In Figure 6.3d, we simulate, using 10 000 runs, a histogram of the distribution of crest factors (for $N = 255$ and $N_s = 4096$). The histogram follows a *Gumbel* distribution, as expected for the maximum of M draws from a parent distribution that decays exponentially or faster. The theorem is very much analogous to the Central Limit Theorem. The black curve in Figure 6.3d is calculated given the normal distribution from this problem (Gumbel, 1958).



g. Not-so-random phases. Do random phases produce the lowest crest factor for large N? No! Figure 6.4c shows a deterministic choice that gives a crest factor of ≈ 1.66 but works only when all harmonics up to a maximum value are selected. Here is a simple way to lower the crest factor that works for any choice of harmonics: Generate N_{trials} random-phase multisine waveforms and select the one with the lowest crest factor. For 100 harmonics and 1000 points per fundamental period, your plot should resemble the one at left. Crest factors $\lesssim 2.5$ are readily obtained by this method. Using numerical optimization techniques to adjust systematically the phases can further lower the crest factor to ≈ 1.4 (Schoukens et al., 2012).

Solution.

a. Square wave. Let the square wave values be $\pm A$. Then clearly $u_{\max} = A$. Since $u(t)^2 = A^2$, we have $u_{\text{rms}} = A$ and Cr = 1.
Sine wave. Let $u = A\cos 2\pi t$. Then $u_{\max} = A$ and $u_{\text{rms}}^2 = A^2 \int_0^1 \mathrm{d}t \cos^2 2\pi t = A^2/2$, so that Cr $= \sqrt{2}$.
Delta function. Let $u(t) = \lim_{A\to\infty} A$ for $0 < t < 1/A$, which goes to a unit-amplitude delta function as $A \to \infty$. Then $u_{\max} = A$ and $u_{\text{rms}} = \sqrt{A}$ if we take the period of the waveform to be unity. Thus, Cr $= \sqrt{A}$, which diverges.

b. We write

$$u_N^2 = \left(\sum_{n=1}^{N} A_n \cos(\omega_n t + \varphi_n)\right)^2$$
$$= \sum_{n=1}^{N} A_n^2 \cos^2(\omega_n t + \varphi_n) + \sum_{m\neq n} A_m A_n \cos(\omega_m t + \varphi_m)\, \cos(\omega_n t + \varphi_n)\,.$$

When we average u_N^2 over the period τ, the direct terms give

$$\frac{1}{\tau}\int_0^\tau \mathrm{d}t \cos^2(\omega_n t + \varphi_n) = \frac{1}{\tau}\int_0^\tau \mathrm{d}t \cos^2(\omega_n t) = \frac{1}{\omega_n}\frac{\omega_n}{2\pi}\int_0^{2\pi} \mathrm{d}\theta \cos^2\theta = \frac{1}{2}\,,$$

whose sum gives us the required identity. On the other hand, the cross terms can be rewritten as

$$\cos(\omega_m t + \varphi_m)\, \cos(\omega_n t + \varphi_n)$$
$$= \tfrac{1}{2}\left\{\cos\left[(\omega_m + \omega_n)t + \varphi_m + \varphi_n\right] + \cos\left[(\omega_m - \omega_n)t + \varphi_m - \varphi_n\right]\right\}\,.$$

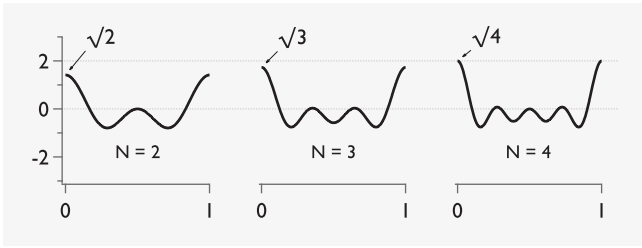
Because $\omega_m \pm \omega_n = \omega_{m\pm n}$, the sums and differences in the cross terms are themselves sines and cosines with harmonic frequencies (integer multiples of f_0). These terms then average to zero when integrated over a full period τ.

c. Speed up for different cases:

N	N_s	speedup
$2^8 - 1 = 255$	$2^{11} = 1024$	100
$2^{10} - 1 = 1023$	$2^{12} = 4096$	520
$2^{12} - 1 = 4095$	$2^{14} = 16\,384$	2100

d. For $\varphi_n = 0$, the graphs tend to a delta function as $N \to \infty$. The maximum amplitude is at $t = 0, 1, \ldots$. We have $u(0) = \frac{1}{\sqrt{N}} \sum_{n=1}^{N}(1) = \sqrt{N}$, so that Cr $= \sqrt{N}/(1/\sqrt{2}) = \sqrt{2N}$.

e. For plots, here are $N = 2, 3, 4$:



N	Cr	Phase
1	$\sqrt{2}$	—
2	1.76	1.58
3	1.62	2.85, 2.16
4	1.48	0, π, 0

f. We have

i. We change variables to $u = \cos\varphi$:

$$p(u) = \frac{1}{2\pi} \sum \frac{1}{\mathrm{d}u/\mathrm{d}\varphi} = \frac{1}{2\pi}(2)\frac{1}{\sqrt{1-u^2}}\,,$$

where the factor of 2 comes from the $\pm u$ roots. The sum of N random variables of mean 0 and variance $\frac{1}{2N}$ then tends towards a Gaussian of mean 0 and variance $N \times \frac{1}{2N} = \frac{1}{2}$. Note that the variance with $A = 1/\sqrt{N}$ is $1/2N$.

ii. From the previous part, we have the sum of N random variables of mean 0 and variance $\frac{1}{2}$. Summing N of these with an amplitude $1/\sqrt{N}$ (variance $1/N$), by the CLT approaches a Gaussian of mean 0 and variance $N(1/2)(1/N) = \frac{1}{2}$.

iii. The probability that one point drawn at random from $u_N(t)$ exceeds the value u is $1-\Phi(u)$. If we pick M independent points in $u_N(t)$, the probability that all of them are less than u is

$$\{1 - [1 - \Phi(u)]\}^M \approx \mathrm{e}^{-M[1-\Phi(u)]}\,,$$

The scale value is e^{-1}. Taking logarithms, we have

$$M[1 - \Phi(u)] \approx 1\,.$$

In this problem, either tail ($\pm u$) is ok, so $[1 - \Phi(u)] \to 2[1 - \Phi(u)]$. Thus,

$$2M[1 - \Phi(u)] = 1\,,$$

which gives the condition to determine $u_{\max}$.

Since $u_N \sim \mathcal{N}(0, \frac{1}{2})$, we have

$$p(u_N) = \frac{1}{\sqrt{2\pi\sigma^2}} \, \mathrm{e}^{-u_N^2/2\sigma^2} = \frac{1}{\sqrt{\pi}} \, \mathrm{e}^{-u_N^2}$$

and

$$1 - \Phi(u_N) = \tfrac{1}{2}\mathrm{erfc}(u_N / \sqrt{2\sigma^2}) = \tfrac{1}{2}\mathrm{erfc}(u_N) \, ,$$

with erfc the complementary error function. We then determine $u_{\max}$ numerically by solving the transcendental equation

$$\mathrm{erfc}(u_{\max}) = \frac{1}{M}$$

for $u_{\max}(M)$.

Note that it is tempting to think that, since $M \gg 1$, we can try an asymptotic expansion for erfc. The large-u expansion is

$$\mathrm{erfc}(u) \sim \left(\frac{\mathrm{e}^{-u^2}}{u\sqrt{\pi}} \right) \left(1 + O(u^{-2}) \right) \, .$$

Dropping any higher-order terms leads to another relation for $u(M)$:

$$\frac{\mathrm{e}^{-u^2}}{u\sqrt{\pi}} = \frac{1}{M} \, .$$

The temptation is to take logs, ignore the denominator terms, and conclude that

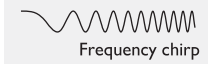
$$u_{\max} \sim \sqrt{\ln M} \, ,$$

a simple, elegant, but not terribly accurate result. For example, even with $M = 10^6$, the equation $\mathrm{erfc}(u) = 1/M$ gives $u \approx 3.45$, while $u \approx \sqrt{\ln M}$ gives 3.72. Empirically, setting $M = 2N$ (with N the number of frequencies) in the approximation gives a value for $u_{\max}$ and Cr that is accurate to better than 1% for $N > 100$. This leads to the expression $\mathrm{Cr} = \sqrt{2\ln(2N)}$ in the figure caption in the problem.

iv. See the book for the graphical results.

g. See graph in book.

6.3 **Frequency chirp**. As illustrated at left, a frequency chirp consists of a sinusoid of continuously varying frequency. It "runs together" the individual sinusoids of the frequency-domain method. We retain the advantage of probing (almost) frequency by frequency but set aside the need to wait for the transients to die away. We also easily control the amplitude of each frequency, boosting in regions where the output is weak, if needed.


Frequency chirp

a. Find an analytic form for a chirp that sweeps from a frequency f_1 to f_2 in a time τ.

b. One disadvantage is that the power spectrum differs from the ideal "brick-wall" of a multisine. Plot the power spectrum numerically for a chirp that goes from 1 to 2 Hz over a time τ = 10, 100, and 1000 s.
c. Compute the fraction of power that falls outside the 1–2 Hz range, as a function of τ. How does the error decrease with τ?

Solution.

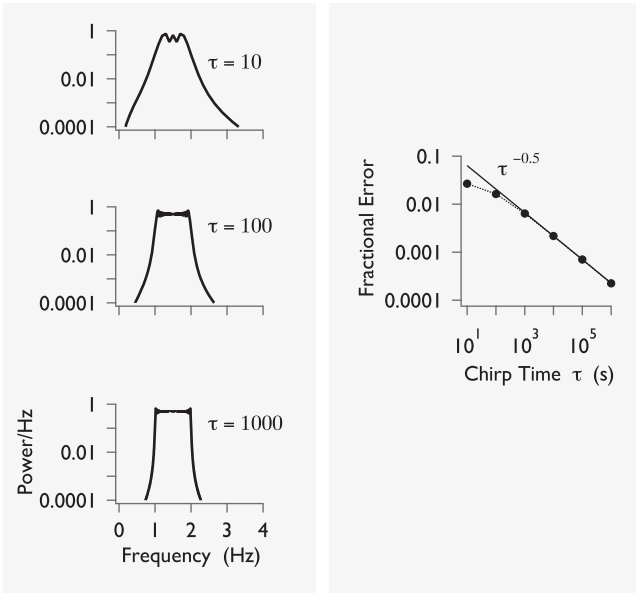
a. Analytic form for a chirp:

$$u(t) = u_0 \cos\left(\frac{\mathrm{d}\varphi}{\mathrm{d}t}\right), \qquad \varphi(t) = f_0 t + \frac{t^2}{2\tau}(f_1 - f_0)\,.$$

Taking the derivative of $\varphi(t)$ gives a time-dependent frequency of

$$\frac{\mathrm{d}\varphi}{\mathrm{d}t} \equiv f(t) = f_0 + \frac{t}{\tau}(f_1 - f_0)\,.$$

b. Power spectrum. See left plots.
c. The error decreases as $\tau^{-1/2}$, which is quite slow.



6.4 **Discrete random binary sequence (DRBS)**. Binary signals have a crest factor of 1 (see Problem 6.2) and thus inject the most power for a given input range. To define a DRBS, at every time interval kT_s, choose $\pm u_0$, with equal probability for $+u_0$ and $-u_0$, as shown at right. Here, we explore a number of properties of these signals.

a. Show – perhaps handwaving is good enough – that the autocorrelation function $R_{uu}(\tau) = 1 - \frac{\tau}{T_s}$ for $|\tau| < T_s$ and that it vanishes for larger $|\tau|$.
b. From the autocorrelation function, find the power spectrum.

c. Write a program to input a DRBS to a harmonic oscillator. Extract the transfer function and compare to the expected form. Be careful about aliasing. Show that oversampling – sampling the input and output at rates that are integer multiples of the original sampling frequency – helps.

Another binary signal variant is the *pseudo-random binary sequence* (PRBS), a *deterministic* sequence generated by a combination of shift registers and XOR logical operations. Its spectrum again is close to white. Its chief advantages are (i) because it is deterministic, its autocorrelation function may be calculated exactly, with no extra statistical uncertainty due to finite lengths of records. (ii) You can measure the signal repeatedly and average the output, reducing the effects of measurement noise. (iii) Because the sequence is periodic, the amount of power that "leaks" outside the desired band is much less than a DRBS signal.

Finally, because they probe just two values, binary sequences do *not* help to detect nonlinearity. It is then better to use an input that explores all levels.

Solution.

a. For $\tau = 0$, the statement is fairly obvious. The product $u(t)\,u(t+\tau) = u(t)^2 = u_0^2$. Then

$$R_{uu}(\tau = 0) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \mathrm{d}t\, u_0^2 = u_0^2\,.$$

For $|\tau| < T_s$, we note that a linear fraction, $1 - |\tau|/T_s$ is "coherent" in this way, whereas the rest is incoherent and averages to zero, in an infinite sample. Similarly, for $|\tau| > T_s$, every point is incoherent and thus averages to zero. Putting all this together gives the desired identity.

b. Power spectrum From the Wiener-Khintchine relation, the power spectrum is

$$S_{uu}(\omega) = \tau\,\mathrm{sinc}^2\left(\frac{\omega\tau}{2}\right).$$

c. Notes on program. Maybe add some typical output?

6.5 Noise and Fourier transforms. Consider a sampled time series of observation noise ξ_k that is white, with $\langle\xi_k\rangle = 0$ and $\langle\xi_k\,\xi_{k'}\rangle = \xi^2\delta_{kk'}$ and with $0 \le k \le N-1$.

a. Show that the discrete Fourier transform $\xi(\ell)$ satisfies $\langle\xi(\ell)\rangle = 0$ and $\langle\xi(\ell)\,\xi(\ell')^*\rangle = N\xi^2\delta_{\ell\ell'}$. Here, $0 \le \{\ell, \ell'\} \le N-1$.

b. Why is each Fourier component statistically independent, even though it is built up from the entire time series?

c. The time series has N components, but the DFT has $2N$ components, since $\xi(\ell)$ is complex. Further, the real and imaginary parts of $\xi(\ell)$ are statistically independent. How can N independent noise components lead to $2N$ Fourier components?

d. Show that Parseval's theorem is satisfied and explain the physical significance:

$$\sum_{k=0}^{N-1} \left\langle \xi_k^2 \right\rangle = \frac{1}{N} \sum_{\ell=0}^{N-1} \left\langle |\xi(\ell)^2| \right\rangle$$

Note: For colored time-domain noise, $\xi_k = \sum_n h_{k-n}\, e_n$, for white noise $e_n \sim \mathcal{N}(0,1)$. Then a similar argument to (a) gives $\langle \xi(\ell)\, \xi(\ell')^* \rangle = N|H(\mathrm{e}^{-2\pi\,\mathrm{i}\ell/N})|^2\, \delta_{\ell\ell'}$, where $H(\cdot)$ is the Z-transform of h_k. The Fourier components remain independent complex, zero-mean Gaussian variables, but with frequency-dependent variances.

Solution.

a. Noting that the times series coefficients ξ_k are real, we have

$$\begin{aligned}
\langle \xi(\ell)\, \xi(\ell')^* \rangle &= \sum_{k,k'=0}^{N-1} \left\langle \xi_k\ \mathrm{e}^{\frac{-2\pi\,\mathrm{i}k\ell}{N}}\ \xi_{k'}\ \mathrm{e}^{\frac{2\pi\,\mathrm{i}k'\ell'}{N}} \right\rangle \\
&= \sum_{k,k'=0}^{N-1} \langle \xi_k\, \xi_{k'} \rangle\, \mathrm{e}^{\frac{-2\pi\,\mathrm{i}k\ell}{N}}\ \mathrm{e}^{\frac{2\pi\,\mathrm{i}k'\ell'}{N}} \\
&= \xi^2 \sum_{k,k'=0}^{N-1} \delta_{k,k'}\ \mathrm{e}^{\frac{-2\pi\,\mathrm{i}k\ell}{N}}\ \mathrm{e}^{\frac{2\pi\,\mathrm{i}k'\ell'}{N}} \\
&= \xi^2 \sum_{k=0}^{N-1} \mathrm{e}^{\frac{-2\pi\,\mathrm{i}k(\ell-\ell')}{N}} \\
&= N\xi^2 \delta_{\ell\ell'}\,.
\end{aligned}$$

b. Even though each Fourier component is built up of all the time series elements, they are all orthogonal by construction and thus are independent variables. It is as if we take a certain amount of randomness and apportion it to different basis vectors.

c. The $2N$ Fourier variables are not all independent. Because ξ_k are real, the DFT components satisfy $\xi(N-\ell) = \xi(\ell)^*$, giving N independent components.

d. Mathematically, Parseval's theorem is a simple consequence of part (a). Setting $\ell = \ell'$, we have $\langle |\xi(\ell)|^2 \rangle = N\xi^2$, so that $\langle \xi_k^2 \rangle = (1/N)\langle |\xi(\ell)|^2 \rangle = \xi^2$. Physically, the noise power is the same, whether summed up in the time domain or in the frequency domain.

6.6 Aliasing, two ways. We can calculate the power spectrum of a sampled signal via the sampling theorem (Chapter 5) or directly from the discrete dynamics (Chapter 6). In a simple case, the two approaches give the same answer: Consider noise-free observations of a 1d Brownian particle, where $\gamma \dot{x}(t) = \xi_F(t)$. The noise $\langle \xi_F(t)\, \xi_F(t') \rangle = \sqrt{2D}\, \gamma\, \delta(t-t')$, and the power spectral density is $\langle |x|^2 \rangle(\omega) = 2D/\omega^2$. Now sample $x(t)$ at intervals T_s, giving x_k. Calculate its power spectrum two ways:

a. Discretize the continuous equations by integrating over T_s. Take the Z-transform and calculate the magnitude.

b. Use the sampling theorem, Eq. (5.3), and $\sum_{n=-\infty}^{\infty} \frac{1}{(\omega - n\omega_\mathrm{s})^2} = \frac{\pi^2}{\omega_\mathrm{s}^2} \csc(\pi\omega/\omega_\mathrm{s})^2$. Derive the required identity by applying Parseval's Theorem to $f(t) = \mathrm{e}^{\mathrm{i}\omega t}$,

assuming the function to be periodic with period T_s and to have jump discontinuities (Stone and Goldbart, 2009, Section 2.2.3).

Verify that as $\omega \to 0$, your discrete power spectrum approaches the continuous one.

Solution.

The goal is to show that, for a signal sampled at intervals T_s, the power spectrum is

$$\left\langle |x|^2 \right\rangle(\omega) = \frac{DT_s^2}{1 - \cos \omega T_s} .$$

Before starting, note the units: DT_s^2 has units of $\ell^2 \cdot t$. Evaluating the integral $\int \mathrm{d}\omega \left\langle |x|^2 \right\rangle(\omega)$ over a range of frequencies thus gives a length squared, as it should.

a. *Discretization and the Z-transform*: We discretize exactly by integrating over an interval T_s, giving the discrete dynamical system,

$$x_{k+1} = x_k + \xi_k , \qquad \langle \xi_n \rangle = 0 , \qquad \langle \xi_k \, \xi_\ell \rangle = 2DT_s \, \delta_{k\ell} .$$

Taking the Z-transform gives

$$(z - 1)\, x(z) = \xi(z) \qquad \Longrightarrow \qquad |x|^2(z) = \frac{|\xi|^2}{|z - 1|^2} .$$

We then evaluate at $z = \mathrm{e}^{\mathrm{i}\omega T_s}$, which implies that

$$|z - 1|^2 = \left(\mathrm{e}^{\mathrm{i}\omega T_s} - 1\right) \left(\mathrm{e}^{-\mathrm{i}\omega T_s} - 1\right) = (1 - 2\cos \omega T_s + 1) = 2(1 - \cos \omega T_s) .$$

Using the relation $\langle |\xi|^2 \rangle = (2DT_s)\, T_s$, we then find

$$\left\langle |x|^2 \right\rangle(\omega) = \frac{DT_s^2}{1 - \cos \omega T_s} .$$

b. Alternatively, we recall from the sampling theorem, Eq. (5.3), that we can write the Fourier transform of the continuous signal $x(t)$ as

$$x_s(\omega) = \frac{1}{T_s} \sum_{n=-\infty}^{\infty} x(\omega - n\omega_s) .$$

From $x(\omega) = -\mathrm{i}\xi_v(\omega)/\omega$, we have

$$x_s(\omega) = \frac{-\mathrm{i}}{T_s} \sum_{n=-\infty}^{\infty} \frac{\xi_v(\omega - n\omega_s)}{\omega - n\omega_s} .$$

Since $\xi_v(\omega)$ represents white noise, the ensemble average is

$$\langle \xi_v(\omega)\, \xi_v(\omega') \rangle = 2DT_s^2 \, \delta(\omega - \omega')$$

and, hence,

$$\left\langle |x_s|^2 \right\rangle(\omega) = 2D \sum_{n=-\infty}^{\infty} \frac{1}{(\omega - n\omega_s)^2} ,$$

where $\omega_s = 2\pi/T_s$. We then use the identity

$$\sum_{n=-\infty}^{\infty} \frac{1}{(\omega - n\omega_s)^2} = \frac{\pi^2}{\omega_s^2} \csc\left(\frac{\pi\omega}{\omega_s}\right)^2 = \frac{T_s^2}{4} \frac{1}{\sin\left(\frac{\omega T_s}{2}\right)^2} = \frac{T_s^2}{2} \frac{1}{1 - \cos\omega T_s}$$

to conclude that

$$\left\langle |x_s|^2 \right\rangle(\omega) = \frac{DT_s^2}{1 - \cos\omega T_s} .$$

Here, the Fourier transform of the sampled-signal is denoted $x_s(\omega)$, to distinguish it from the Fourier transform of the continuous signal, $x(\omega)$.

To establish the identity

$$\sum_{n=-\infty}^{\infty} \frac{1}{(\omega - n\omega_s)^2} = \frac{\pi^2}{\omega_s^2} \csc\left(\frac{\pi\omega}{\omega_s}\right)^2 ,$$

Fourier expand $f(t) = \mathrm{e}^{\mathrm{i}\omega t}$, assuming $f(t)$ to be periodic, with period T_s:

$$f(t) = \mathrm{e}^{\mathrm{i}\omega t} = \sum_{n=-\infty}^{\infty} c_n \, \mathrm{e}^{\mathrm{i}n\omega_s t} .$$

The Fourier series coefficients c_n are given by

$$\begin{aligned} c_n &= \frac{1}{T_s} \int_0^{T_s} \mathrm{d}t \, f(t) \, \mathrm{e}^{-\mathrm{i}n\omega_s t} \\ &= \frac{\omega_s}{2\pi} \int_0^{T_s} \mathrm{d}t \, \mathrm{e}^{\mathrm{i}(\omega - n\omega_s)t} \\ &= \left(\frac{\omega_s}{2\pi}\right) \frac{2\sin\frac{\pi}{\omega_s}(\omega - n\omega_s)}{\omega - n\omega_s} \\ &= \left(\frac{\omega_s}{\pi}\right) \frac{\sin\frac{\pi\omega}{\omega_s}}{\omega - n\omega_s} , \end{aligned}$$

where the last identity uses $\sin(\theta - n\pi) = \sin\theta$. Finally, Parseval's theorem equates $\|f(t)\|^2 = (1/T_s)\int_0^{T_s} \mathrm{d}t \, f(t) \, f^*(t) = \sum_n |c_n|^2$. Since $\|f(t)\|^2 = 1$, we have

$$1 = \left(\frac{\omega_s}{\pi}\right)^2 \sin^2\left(\frac{\pi\omega}{\omega_s}\right)^2 \sum_{n=-\infty}^{\infty} \frac{1}{(\omega - n\omega_s)^2} .$$

Rearranging gives the desired identity. More general cases can be treated via the Poisson summation formula. See Problem A.4.2.

Finally, for small ω, the sampled-signal power spectrum is

$$\left\langle |x|^2 \right\rangle(\omega) \approx \frac{DT_s^2}{1 - \left[1 - \frac{1}{2}(\omega T_s)^2\right]} = \frac{2D}{\omega^2} ,$$

which completes the "loop" back from the sampled signal power spectrum to that of the original continuous signal.

6.7 **Bias of transfer function estimates**. The simple estimate of a transfer function as the ratio of two noisy DFT variables is biased. That is, $\langle G \rangle = \left\langle \frac{y+\xi_y}{u+\xi_u} \right\rangle \neq \frac{y}{u}$, where we drop the frequency dependence on all quantities. Assuming that $\langle \xi_y \xi_u^* \rangle = 0$, the bias arises entirely from the fluctuations in the input, ξ_u. Thus, we simplify by setting $\xi_y = 0$. Scaling by y/u and setting $z = \xi_u/u$, we can study the bias by comparing $\langle b(z) \rangle \equiv \left\langle \frac{1}{1+z} \right\rangle$ to 1. Here, $z = x + iy$, with $x, y \sim \mathcal{N}(0, \sigma^2/2)$ and $\sigma = 1/\mathrm{SNR}_u$.

a. Taylor expand to show that $\langle |b(z)|^2 \rangle = 1 + \sigma^2 + O(\sigma^4)$. Intuitively, negative fluctuations increase $|b|^2$ more than positive fluctuations decrease it. Indeed, if z can take the value ≈ -1, the corresponding fluctuation in b will be very large.
b. Since z is complex, its statistics are tricky. Show that $\langle z^2 \rangle = 0$.
c. Expand $\langle b(z) \rangle = \left\langle \frac{1}{1+z} \right\rangle$ in a full Taylor series and use the result from the previous part to conclude, incorrectly, that $\langle b(z) \rangle = 1$. Where is the flaw in the argument?
d. Calculate the bias directly, by integrating $b(z)$ over the probability distribution for z. By evaluating the integral first in terms of the real and imaginary components (x and y) and then converting to polar coordinates, show that $\langle b(z) \rangle = 1 - \mathrm{e}^{-1/\langle |z|^2 \rangle}$.

Solution.

a. We have

$$\begin{aligned}
\langle |b|^2 \rangle &= \left\langle \left| \frac{1}{1+z} \right|^2 \right\rangle \\
&= \left\langle \frac{1}{(1+x)^2 + y^2} \right\rangle \\
&\approx \left\langle 1 - 2x - x^2 - y^2 + 4x^2 + \cdots \right\rangle \\
&= 1 - 0 - \frac{\sigma^2}{2} - \frac{\sigma^2}{2} + 4\frac{\sigma^2}{2} + \cdots \\
&= 1 + \sigma^2 + O(\sigma^4) .
\end{aligned}$$

b. We have $\langle z^2 \rangle = \langle (x + \mathrm{i}y)^2 \rangle = \langle x^2 - y^2 + 2\,\mathrm{i}\,xy \rangle = \frac{\sigma^2}{2} - \frac{\sigma^2}{2} + 2i(0) = 0$.
c. Since $\langle z \rangle = \langle z^2 \rangle = 0$, the higher-order moments are also zero. This can be seen by noting that the moment-generating function is $M(k) = \mathrm{e}^{\langle z \rangle k + \frac{1}{2}\langle z^2 \rangle k^2} = \mathrm{e}^0 = 1$ and that the mth moment is given by the mth derivative of $M(k)$, evaluated at $k = 1$.
   You might then conclude that $\langle b(z) \rangle = 1 - \langle z \rangle + \langle z^2 \rangle - \langle z^3 \rangle + \cdots = 1$. The conclusion is false, however, because in order for the Taylor series to be valid, z must be in the region of convergence in the complex plane ($|z| < 1$). But fluctuations in z are unbounded, and a Taylor series is not allowed.

d. We have

$$\begin{aligned}
\langle b(z)\rangle &= \int \mathrm{d}z\, \frac{1}{1+z}\, p(z) \\
&= \int \mathrm{d}z\, \frac{1}{1+z} \left(\frac{1}{2\pi(\sigma^2/2)}\, \mathrm{e}^{\frac{-|z|^2}{\sigma^2}} \right) \\
&= \frac{1}{\pi\sigma^2} \iint_{\mathcal{R}^2} \mathrm{d}x\, \mathrm{d}y\, \frac{1+x-\cancelto{0}{\mathrm{i}y}}{(1+x)^2+y^2}\, \mathrm{e}^{-\frac{x^2+y^2}{\sigma^2}} \\
&= \frac{1}{\pi\sigma^2} \iint_{\mathcal{R}^2} \mathrm{d}x\, \mathrm{d}y\, \frac{1+x}{(1+x)^2+y^2}\, \mathrm{e}^{-\frac{x^2+y^2}{\sigma^2}} \\
&= \frac{2}{\sigma^2} \int_0^\infty \mathrm{d}r\, r\, \mathrm{e}^{-\frac{r^2}{\sigma^2}} \left[\frac{1}{2\pi} \int_0^{2\pi} \mathrm{d}\theta\, \frac{1+r\cos\theta}{1+r^2+2r\cos\theta} \right] \\
&= \frac{2}{\sigma^2} \int_0^\infty \mathrm{d}r\, r\, \mathrm{e}^{-\frac{r^2}{\sigma^2}} \begin{cases} 1 & 0<r<1 \\ 0 & r>1 \end{cases} \\
&= \frac{2}{\sigma^2} \int_0^1 \mathrm{d}r\, r\, \mathrm{e}^{-\frac{r^2}{\sigma^2}} \\
&= \int_0^{1/\sigma^2} \mathrm{d}u\, \mathrm{e}^{-u} \\
&= 1 - \mathrm{e}^{-1/\sigma^2}\ .
\end{aligned}$$

In the third line, the y numerator-term in the integral vanishes because the integrand is an odd function of y that is integrated over $-\infty < y < \infty$. Notice how the result shows an essential singularity for $\sigma \to 0$. The result is "beyond all orders" of perturbation theory in σ^2.

6.8 **Variance of transfer function estimate**. For high SNR, Problem 6.9 shows that fluctuations about the mean value of a complex number are approximately Gaussian. Using this idea and the result in Problem 6.7b, Taylor expand the transfer function estimate $\delta G = (y_0 + \delta y)/(u_0 + \delta u) - G_0$ to derive Eq. (6.10). First derive the result in terms of the unknown true values u_0 and $\sigma_u^2 = \langle|\delta u|^2\rangle$, etc. and then express in terms of the estimated means and variances given by Eqs. (6.7) and (6.8), for M periods of the input function $u = u_0 + \delta u$. All quantities are functions of the frequency ω_ℓ.

Solution.

For one measurement block, we have

$$\begin{aligned}
\delta G &= \left(\frac{y_0 + \delta y}{u_0 + \delta u} \right) - \left(\frac{y_0}{u_0} \right) \\
&= \frac{y_0}{u_0} \left(\frac{1 + \delta y/y_0}{1 + \delta u/u_0} - 1 \right) \\
&\approx G_0 \left(\frac{\delta y}{y_0} - \frac{\delta u}{u_0} \right),
\end{aligned}$$

where we expand to first order. The variance is then

$$\begin{aligned}\sigma^2_{|G|} &\equiv \left\langle |\delta G|^2 \right\rangle \\ &= |G_0|^2 \left\langle \left(\frac{\delta y}{y_0} - \frac{\delta u}{u_0}\right)\left(\frac{\delta y}{y_0} - \frac{\delta u}{u_0}\right)^* \right\rangle \\ &= |G_0|^2 \left[\frac{\sigma_y^2}{|y_0|^2} + \frac{\sigma_u^2}{|u_0|^2} - 2\,\mathrm{Re}\left(\frac{\sigma_{yu}^2}{y_0\, u_0^*}\right)\right].\end{aligned}$$

Note that we neglect terms such as $\langle(\,\delta u/u_0)^2\rangle$, following the result from Problem 6.7b.

If we then measure M periods of the input, we can replace all quantities by their estimates. The variance of the estimate is also reduced by a factor of M, since we have M realizations of the single period estimates. This gives, finally,

$$\hat{\sigma}^2_{|G|} \approx \left(\frac{|\hat{G}|^2}{M}\right)\left[\frac{\hat{\sigma}_y^2}{|\hat{y}|^2} + \frac{\hat{\sigma}_u^2}{|\hat{u}|^2} - 2\,\mathrm{Re}\left(\frac{\hat{\sigma}_{yu}^2}{\hat{y}\,\hat{u}^*}\right)\right].$$

The approximation is threefold:

a. Replacing exact quantities by the estimates (e.g., $y_0 \to \hat{y}$);
b. ignoring the corrections to Gaussian distributions discussed in Problem 6.7;
c. neglecting correlations between the noise in one input block and its effect on the output of the next block.

6.9 **Amplitude and phase noise**. Complex Gaussian random variables $z = x + iy$ with non-zero mean result from taking a discrete Fourier transform of a finite-amplitude signal. Let $x \sim \mathcal{N}(x_0, \sigma^2)$ and $y \sim \mathcal{N}(y_0, \sigma^2)$ be independent Gaussian random variables. Define fluctuations $\delta x = x - x_0$ and $\delta y = y - y_0$, and define magnitude and phase variables $r = r_0 + \delta r$ and $\theta = \theta_0 + \delta\theta$. Define the signal-to-noise ratio as $\mathrm{SNR}_x = x_0/\sigma$ and $\mathrm{SNR}_y = y_0/\sigma$.[1]

a. For high SNR ($\gg 1$), find δr and $\delta\theta$ to first order. Calculate the mean and variance of each and show that $\langle\, \delta r\, \delta\theta\rangle = 0$. Interpret the result geometrically.
b. Describe the zero SNR case ($r_0 = 0$). Derive (or guess) the radial and angular distributions for this case. Then describe how the low SNR case (δx and δy comparable to r_0) interpolates between the high-SNR and zero-SNR cases. Illustrate three cases (high, low, and zero SNR) by Monte Carlo simulations. For each case, plot $p(r, \theta)$ and the marginal plots $p(r)$ and $p(\theta)$.
c. One subtlety is that the mean of the magnitude is biased. An exact calculation gives $\langle r\rangle = \sqrt{r_0^2 + \sigma^2}$. Interpret this result physically. Derive it approximately by continuing the Gaussian expansion for r to second order in the noise.

This problem asks you to think physically about the distributions $p(r)$ and $p(\theta)$, but they are straightforward to investigate analytically, as well (Goodman, 2007).

[1] Physicists often define signal-to-noise ratios in terms of amplitudes, engineers in terms of power.

For reference, if $\boldsymbol{r}_0 = (r_0\cos\theta_0, r_0\sin\theta_0)$ and $\boldsymbol{r} = (r\cos\theta, r\sin\theta)$, then

$$p(\boldsymbol{r}) = \frac{1}{2\pi\sigma^2}\exp\left\{\left[-\frac{1}{2\sigma^2}|\boldsymbol{r}-\boldsymbol{r}_0|^2\right]\right\}$$
$$p(r,\theta) = \frac{r}{2\pi\sigma^2}\exp\left\{\left\{-\frac{1}{2\sigma^2}\left[r^2+r_0^2-2rr_0\cos(\theta-\theta_0)\right]\right\}\right\}.$$

• Integrating out θ leads to $p(r) = \int_0^{2\pi} \mathrm{d}\theta\, p(r,\theta) = \frac{r}{\sigma^2}\exp\left\{\left(\frac{-(r^2+r_0^2)}{2\sigma^2}\right)\right\} I_0\left(\frac{r\,r_0}{\sigma^2}\right)$ (*Rice distribution*), with $I_0(\cdot)$ a modified Bessel function of the first kind of order zero.
• Integrating out r by $p(\theta) = \int_0^{\infty} \mathrm{d}r\, p(r,\theta)$ gives, with $\bar{r} = r_0/\sigma$,

$$p(\theta) = \frac{1}{2\pi}\exp\left\{\left(-\tfrac{1}{2}\bar{r}^2\right)\right\}\left[1+\left(\frac{\bar{r}\cos\theta}{\sqrt{2}}\right)\exp\left\{\left(\tfrac{1}{2}\bar{r}^2\cos^2\theta\right)\right\}\int_{-\infty}^{\bar{r}\cos\theta}\mathrm{d}r'\,\exp\left\{\left(-\tfrac{1}{2}r'^2\right)\right\}\right].$$

Solution.

a. Expanding $r^2 - r_0^2$ and $\tan\theta - \tan\theta_0$, we have

$$\delta r = \frac{x_0\delta x + y_0\delta y}{r_0}, \qquad \delta\theta = \frac{x_0\delta y - y_0\delta x}{r_0^2}.$$

Since $\langle\,\delta x\rangle = \langle\,\delta y\rangle = 0$, we have $\langle\,\delta r\rangle = \langle\,\delta\theta\rangle = 0$. For the second moments and variances, we use $\langle\,\delta x^2\rangle = \langle\,\delta y^2\rangle = \sigma^2$ and $\langle\,\delta x\,\delta y\rangle = 0$ to show

$$\langle\,\delta r^2\rangle = \frac{x_0^2\sigma^2 + y_0^2\sigma^2}{r_0^2} = \sigma^2$$
$$\langle\,\delta\theta^2\rangle = \frac{x_0^2\sigma^2 + y_0^2\sigma^2}{r_0^4} = \frac{\sigma^2}{r_0^2}$$
$$\langle\,\delta r\,\delta\theta\rangle = \frac{(x_0\delta x + y_0\delta y)\,(x_0\delta y - y_0\delta x)}{r_0^3} = \frac{-x_0y_0\sigma^2 + y_0x_0\sigma^2}{r_0^3} = 0.$$

In polar coordinates, the radial and azimuthal unit vectors ($\hat{r}$ and $\hat{\theta}$) are perpendicular. So the vanishing is essentially for geometrical reasons.

b. For high SNR ratios r_0/σ, the radial and azimuthal coordinates are independent Gaussian random variables. As the signal-to-noise ratio decreases, the angles spread out, reaching the uniform distribution $p(\theta) = 1/(2\pi)$ for SNR = 0. (With pure noise, all phase angles are equally probable.) We illustrate these qualitative ideas for $\mathrm{SNR}_x = 10$, 2, and 0, below. One can derive explicit distributions, but for practical purposes, we want to always be in the high-SNR limit. For example, it is easy to show that for SNR=0, the magnitude distribution for r is the *Rayleigh distribution*

$$p(r) = \frac{r}{\sigma^2}\exp\left\{\left(\frac{-r^2}{2\sigma^2}\right)\right\}.$$

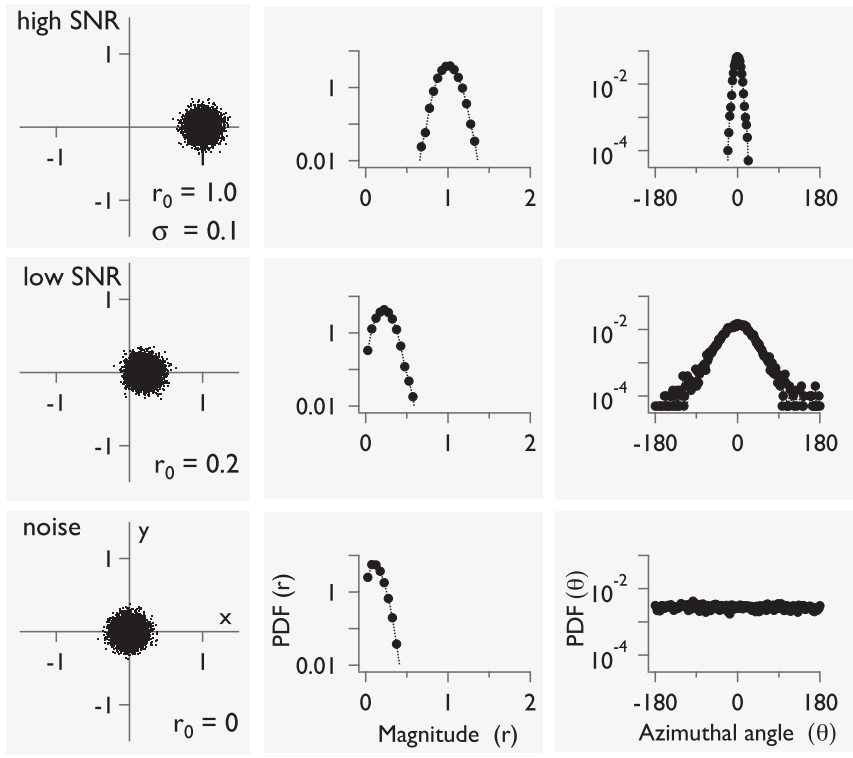
The more general Rice distribution given in the text interpolates between the Rayleigh distribution for $r_0/\sigma \to 0$ and the Gaussian for $r_0/\sigma \to \infty$. In

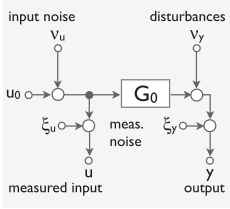
the same limits, the angular distribution $p(\theta)$ interpolates between a uniform distribution and a Gaussian.

c. The mean of the magnitude is biased: $\langle r \rangle = \sqrt{r_0^2 + \sigma^2}$. Physically, the noise power adds to the magnitude. We also see this in the second order expansion of r, whose terms are

$$r_0 + \frac{\sigma^2}{2r_0},$$

which matches the expansion of the square root to lowest order.





6.10 Transfer function bias and multiple noise sources. Consider a system $G_0 = y_0/u_0$ that is probed by noisy signals. In particular, let the input be given by $u = u_0 + \nu_u + \xi_u$, where ν_u represents input noise (e.g., from a power amplifier) to the system and ξ_u the input measurement noise (see the block diagram at left). Let the output be given by $y = y_0 + \nu_y + \xi_y$, where ν_y represents output noise (e.g., thermal fluctuations) and ξ_y the output measurement noise. Assume that the four noise sources are independent. All quantities are frequency-dependent, complex Fourier coefficients.

a. Show that the measured transfer function $G = y/u = G_0/[1 + \frac{\xi_u}{u_0+\nu_u}] + \frac{\nu_y+\xi_y}{u_0+\nu_u+\xi_u}$.
b. Why is G biased? Which noise source is responsible?
c. Explain why increasing input noise can reduce bias.

Solution.

a. The main point is that input noise propagates through the system and contributes $G_0 \nu_u$ to the output signal. Thus, $y_0 = G_0(u_0 + \nu_u)$, and G is given by

$$\begin{aligned} G &= \frac{y}{u} \\ &= \frac{y_0 + \nu_y + \xi_y}{u_0 + \nu_u + \xi_u} \\ &= \frac{G_0(u_0 + \nu_u) + \nu_y + \xi_y}{u_0 + \nu_u + \xi_u} \\ &= \frac{G_0(u_0 + \nu_u)}{u_0 + \nu_u + \xi_u} + \frac{\nu_y + \xi_y}{u_0 + \nu_u + \xi_u} \\ &= G_0 \left(\frac{1}{1 + \frac{\xi_u}{u_0 + \nu_u}} \right) + \frac{\nu_y + \xi_y}{u_0 + \nu_u + \xi_u} . \end{aligned}$$

b. The bias is due to the input measurement noise ξ_u. To see this, we let all other noise sources be zero. Then

$$G = G_0 \left(\frac{1}{1 + \frac{\xi_u}{u_0}} \right),$$

which is the case studied in Problem 6.7. Because the output noise sources are all independent, averaging over them shows that they cannot contribute to the bias.

c. The role of input noise is interesting. From the above formulas, we see that increasing ν_u reduces the bias, since $G \to G_0$. To understand this, we first note that when ν_u is the only noise source, there is no bias:

$$G = \left[\frac{G_0(u_0 + \nu_u)}{u_0 + \nu_u} \right] = G_0 .$$

Intuitively, the input noise cancels out even though we do not know it explicitly. Of course, the cancellation assumes linear dynamics. Increasing ν_u in the presence of other noise sources drives the system to this limit.

6.11 Do not average the magnitude of a Fourier Transform. Averaging the magnitude of multiple Fourier transforms is a poor strategy for reducing noise:

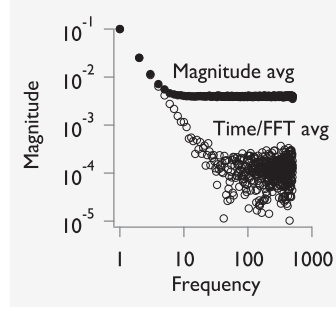
a. Write a program to generate a multifrequency sine wave of period 1 s, sampled 1000 times per period, and repeated for 1000 periods. Let the multifrequency sine wave have harmonics with amplitude $\propto 1/f^2$, with f the frequency, and choose the phases randomly. Calculate the magnitude of the power spectrum three ways: (i) Average each period in the time domain and then compute the DFT magnitude of the time-averaged waveform. (ii) Compute the DFT of each waveform, average the complex waves, and then find the magnitude. (iii) Take the DFT of each waveform, compute the magnitude, and then average.

Plot all three magnitudes on one graph. Why is averaging the magnitude spectra wrong?

b. A less obvious issue is that when the input is noise dominated, the magnitude estimate is biased. To see this, consider an input $x \sim \mathcal{N}(0,1)$ and a transfer function =1. That is, show, both by Monte Carlo simulation and by analytic calculation, that the output has $\sqrt{\langle x^2 \rangle} = 1$ while $\langle |x| \rangle = \sqrt{2/\pi} \approx 0.80$.

Solution.

a. We plot the averaged magnitude (filled markers) and the magnitude of the temporal average (or, equivalently, the FFT average) (hollow markers). The plots of the time-averaged and complex-DFT averaged waveforms are identical. The noise floor of the magnitude average is higher because, in averaging the magnitudes, we lose their phase information. Thus, their power adds incoherently. On the other hand, when averaging in the time or complex Fourier domain, we can average out the noise and lower the noise floor.



b. For the bias, we note that if $x \sim \mathcal{N}(0,1)$, then $\langle x^2 \rangle = 1 = \sqrt{\langle x^2 \rangle}$, but

$$\begin{aligned}
\langle |x| \rangle &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}x\, |x|\, \mathrm{e}^{-x^2/2} \\
&= \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} \mathrm{d}x\, x\, \mathrm{e}^{-x^2/2} \\
&= \sqrt{\frac{2}{\pi}} \int_{0}^{\infty} \mathrm{d}u\, \mathrm{e}^{-u}, \qquad (u = x^2/2) \\
&= \sqrt{\frac{2}{\pi}}.
\end{aligned}$$

This is easily verified by MC simulation. Use 10^6 points to get agreement to 3 decimal places.

6.12 Noisy resistor measurements. Noisy measurements can bias estimates even without dynamics. Let us estimate a resistance by applying a series of noisy currents I_k and measuring noisy voltages V_k (Pintelon and Schoukens, 2012). Let $I_k = 1 + \delta I_k$ and $V_k = 1 + \delta V_k$, with δI_k and $\delta V_k \sim \mathcal{N}(0,1)$ (resistance $R = 1$).

Consider three estimators $\hat{R}$ for the resistance, each minimizing a different cost function J.

a. $J_1(R) \equiv \frac{1}{2}\sum_{k=1}^{N}(R_k - R)^2$, where $R_k = V_k/I_k$. Show that the value of R that minimizes J_1 is given by $\hat{R}_1 = \frac{1}{N}\sum_k R_k \equiv \overline{R_k}$, which is biased (Problem 6.7).

b. $J_2(R) \equiv \frac{1}{2}\sum_k (V_k - RI_k)^2$. Show that choosing $\hat{R}_2 = \overline{V_k\, I_k}/\overline{I_k^2}$ minimizes J_2. Show that $\langle \hat{R}_2 \rangle = 1/(1+\sigma_I^2)$, which implies that estimator is biased.

c. $J_3(R, I, V) \equiv \frac{1}{2}\sum_k \frac{(V_k - V)^2}{\sigma_V^2} + \frac{(I_k - I)^2}{\sigma_I^2}$, with the constraint that $V = RI$. Here, V and I are unknown "true values." Show that choosing $\hat{R}_3 = \overline{V_k}/\overline{I_k}$ minimizes J_3. Hint: use the constraint to eliminate V; then differentiate with respect to both R and I.

d. Write a simulation for $\sigma_I = \sigma_V = 1$ and $N = 500$. Repeat for 10^4 trials and plot histograms of the values $\hat{R}_1$, $\hat{R}_2$, and $\hat{R}_3$. Explain why $\hat{R}_1$ is pathological here but not $\hat{R}_3$. How large a value of N is needed for $\hat{R}_3$ to be ok?

The first estimator is sometimes used in elementary physics laboratory courses, the second in intermediate courses, and the third (hopefully) in more advanced courses. The second is the common, unweighted least squares, assuming no error in the "input" variable (current). The third is the weighted least-squares estimate, taking into account errors in both variables.

Solution.

a.

$$J_1(R) = \frac{1}{2}\sum_{k=1}^{N}(R_k - R)^2 \quad \Longrightarrow \quad \frac{\mathrm{d}J_1}{\mathrm{d}R} = \frac{2}{2}\left[\sum_{k=1}^{N}(R_k) - RN\right](-1) = 0\,,$$

so that $\hat{R}_1 = \overline{R_k}$, as claimed.

b.

$$J_2(R) = \frac{1}{2}\sum_{k=1}^{N}(V_k - RI_k)^2 \quad \Longrightarrow \quad \frac{\mathrm{d}J_2}{\mathrm{d}R} = \left[\sum_{k=1}^{N}(V_k) - RI_k\right](-I_k) = 0\,,$$

so that

$$\hat{R}_2 = \overline{V_k\, I_k}/\overline{I_k^2}\,.$$

Next, we calculate the bias:

$$\langle \hat{R}_2 \rangle = \lim_{N\to\infty} \frac{\sum_k (1+\delta V_k)(1+\delta I_k)}{\sum_k (1 + 2\delta I_k + \delta I_k^2)} \quad \to \quad \frac{N(1+0)}{N(1+\sigma_I^2)} = \frac{1}{1+\sigma_I^2}\,,$$

which shows that the bias is directly linked to input noise in the current (and not to the output, or voltage noise).

c.

$$J_3(R, I, V) = \frac{1}{2}\sum_{k=1}^{N}\frac{(V_k - V)^2}{\sigma_V^2} + \frac{(I_k - I)^2}{\sigma_I^2}\,,$$

subject to the constraint $V = IR$. Putting in the constraint gives

$$J_3(R,I) = \frac{1}{2}\sum_{k=1}^{N} \frac{(V_k - IR)^2}{\sigma_V^2} + \frac{(I_k - I)^2}{\sigma_I^2}\,.$$

Differentiating with respect to R and I then gives

$$\frac{\partial J_3}{\partial R} = \frac{1}{\sigma_V^2}\left[\sum_{k=1}^{N}(V_k) - RIN\right](-I) = 0 \qquad \Longrightarrow \qquad \overline{V_k} = IR$$

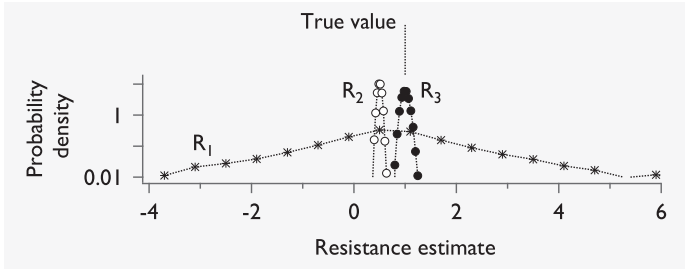
$$\frac{\partial J_3}{\partial I} = \frac{1}{\sigma_V^2}\left[\sum_{k=1}^{N}(V_k) - RIN\right](-R) + \frac{1}{\sigma_I^2}\left[\sum_{k=1}^{N}(I_k) - IN\right](-1) = 0 \qquad \Longrightarrow \qquad \overline{I_k} = I\,.$$

(the first bracketed term is marked "0")

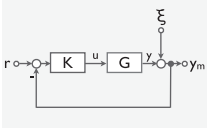
Putting the two results together then gives

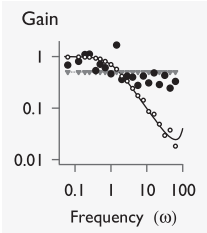
$$\hat{R}_3 = \frac{\overline{V_k}}{\overline{I_k}}\,.$$

d. The estimator $\hat{R}_1$ is pathological because the noise is strong enough that the denominator is nearly zero often. Those trials lead to R_k values that are very large. In the limit of large noise, we have the situation discussed in connection with Eq. (6.9), where the variance diverges. The situation is avoided for $\hat{R}_3$ because the denominator is now an average whose variance decreases as N^{-1}. For large N, there is negligible chance that the denominator will vanish.
The simulations lead to the histograms below. Note the huge spread in $\hat{R}_1$ with its unphysical negative values, and notice the bias in $\hat{R}_2$.



6.13 Measuring a closed-loop transfer function.





a. For the block diagram at left, show that $G_m \equiv \frac{y_m}{u} = \frac{KGr+\xi}{K(r-\xi)}$.
b. Simulate an unstable discrete first-order system that is stabilized by proportional feedback. The dynamics are given by $(y_m)_k = y_k + \xi_k$ and $u_k = -K[r_k - (y_m)_k]$, with $y_{k+1} = (1 + T_s)y_k + T_s u_k$, where the observational noise $\xi_k \sim \mathcal{N}(0, \xi^2)$ and T_s is the sampling time. Use $K = 2$, $\xi^2 = 1$, $T_s = 0.05$ s, and scan frequencies from 0.01 to 10 Hz. Try three different reference signals: $r_k = 0$ (no reference), $r_k \sim \mathcal{N}(0,1)$ (white noise), and r_k a random-phase multisine of RMS amplitude = 1. Run the simulations for 10 periods of 100 s. Discard the first response to eliminate large transients. Your plot should resemble the figure at left, where the solid line is the transfer function of the

noiseless discrete system (u_k to y_k), the triangles the no-reference case, the filled markers the random-phase case, and the white markers the multisine case. Explain the results.

Solution.

a. From $y_m = Gu + \xi$ and $u = K(r - y_m)$, we find

$$y_m = \frac{KGr + \xi}{1 + KG} \quad \text{and} \quad u = \frac{K(r - \xi)}{1 + KG} ,$$

whose ratio gives G_m.

b. First, we calculate the magnitude of the noiseless transfer function. We have

$$y_{k+1} = (1 + T_\mathrm{s})y_k + T_\mathrm{s}\, u_k .$$

We take the Z-transform:

$$(z - 1 - T_\mathrm{s})y = T_\mathrm{s}\, u .$$

Then,

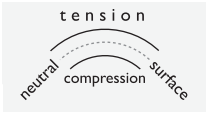
$$\left|\frac{y}{u}\right| = \frac{T_\mathrm{s}}{|z - 1 - T_\mathrm{s}|} ,$$

where $z = \mathrm{e}^{\mathrm{i}\omega T_\mathrm{s}}$. Putting in the values of the constants gives the solid line in the figure. The upturn is due to aliasing: the upper frequency limit slightly exceeds the Nyquist frequency.

- *No reference*: The triangles correspond to the case of no reference. As discussed in the text, the measured transfer function reduces to $-1/K$ in this limit. Thus, for the magnitude, we have $|y/u| = 1/K = 0.5$.
- *White noise*: When we use reference signals drawn from $\mathcal{N}(0, 1)$, there is a great deal of bias at all but the lowest frequencies. We can understand this because we are effectively treating the reference as an unknown random signal. Averaging it is a silly idea. The result is barely passable at low frequencies because G is largest there and $G_m \approx G$ (with noise). At least at these frequencies, we begin to dominate over the observation noise. Increasing the proportional gain would help, but using a multisine is much better.
- *Multisine*: With a periodic multisine that has exactly the same power (RMS=1) as the white-noise reference, the results are much better. Here, because we know the reference and because it is periodic, it does not contribute any statistical error, which diminishes with increasing numbers of measured periods. (With larger gain K, we could get better results with fewer periods.)

6.14 Resonance frequency of a thin plate We fill in some details of Example 6.6. For a careful approach, see Landau et al. (1986) and also Rossing and Russell (1990). Here, in the spirit of making rough approximations, feel free to use handwaving arguments. Consider just one transverse dimension, for simplicity.

a. Let $\psi(x,t)$ be a component of the elastic displacement in a material. Argue that the local kinetic energy per volume is $T = \frac{1}{2}\rho(\psi_t)^2$ and that the local elastic potential energy per volume is $U = \frac{1}{2}E(\psi_x)^2$. Here, ρ is the density and E the *Young's modulus*. The partial derivatives are $\psi_t = \partial_t\psi$ and $\psi_x = \partial_x\psi$.
b. From the Lagrangian $L = T - U$ and the Euler-Lagrange equations for a field, derive a wave equation for ψ. Find the dispersion relation, and show that the expected (longitudinal) sound speed is $c_\mathrm{L} = \sqrt{E/\rho}$. For an object of size ℓ, the expected lowest resonance frequency is $f \approx c_\mathrm{L}/\ell$.
c. Anisotropic objects such as plates can bend with a radius that is much greater than the plate thickness h, leading to lower resonance frequencies. Argue that the bending energy per unit area is of order $U_\mathrm{bend} \sim Eh^3(\psi_{xx})^2$, where h is the plate thickness. Hint: As shown at left, a bent plate of thickness h has one surface under tension and the other under compression. You can also use symmetry, or even a ball-and-spring model to derive the energy.



d. Use the higher-order version of the Lagrangian argument above to show that the resonance is lowered to $f \approx c_\mathrm{L}/(\ell a)$, where the aspect ratio $a = \ell/h$.

Solution.

a. The kinetic energy is $\frac{1}{2}mv^2$, where m is the local mass and v the rate of change of the displacement. Per volume, this is indeed $\frac{1}{2}\rho(\psi_t)^2$, where $\psi_t = \partial_t\psi$. For the potential energy, ψ_x gives the local strain. Notice that if ψ were constant, this would correspond to translating the entire object, which produces zero strain. For a Hooke's law material, we expect $U = \frac{1}{2}E(\psi_x)^2$.
b. The Lagrangian is then

$$L = T - U = \frac{1}{2}\rho\left(\frac{\partial\psi}{\partial t}\right)^2 - \frac{1}{2}E\left(\frac{\partial\psi}{\partial x}\right)^2 .$$

The Euler-Lagrange equations for a field over x and t are

$$\frac{\partial}{\partial t}\frac{\partial L}{\partial\psi_t} + \frac{\partial}{\partial x}\frac{\partial L}{\partial\psi_x} - \frac{\partial L}{\partial x} = 0 .$$

Noting that $\partial_x L = 0$, we have the wave equation,

$$\rho\frac{\partial^2\psi}{\partial t^2} = E\frac{\partial^2\psi}{\partial x^2} ,$$

The dispersion relation is $\omega = c_\mathrm{L}k$, with longitudinal sound speed $c_\mathrm{L} = \sqrt{E/\rho}$.
c. The kinetic energy per area is

$$\int_{-h/2}^{h/2} \mathrm{d}z\, \frac{1}{2}\rho\,(\psi_t)^2 = \tfrac{1}{2}\rho h(\psi_t)^2 .$$

For the potential energy per area, we observe that about the midplane $z = 0$, the strain is zero and it is positive on one side ($z > 0$ for instance) and negative on the other. Taylor expanding then gives

$$U_\mathrm{bend} = \int_{-h/2}^{h/2} \mathrm{d}z\, \frac{1}{2}E(\psi_x)^2$$

$$= \frac{1}{2}E\int_{-h/2}^{h/2} \mathrm{d}z \left(\cancelto{0}{\psi_x|_{z=0}} + z\psi_{xx}|_{z=0} + \cdots\right)^2$$
$$= \frac{1}{24}Eh^3(\psi_{xx})^2 \,.$$

In a more hand-waving argument, we would approximate the integral dimensionally and lose the factor of 1/12. Note that the above argument neglects the coupling between extension in one direction and compression in the perpendicular directions, which is captured by *Poisson's ratio*, ν, which is typically ≈ 0.3. There is then an additional factor of $1-\nu^2$ in the denominator.

d. Since there are second derivatives in the Lagrangian, we integrate by parts one more time in the derivation of the Euler-Lagrange equation and find

$$\frac{\partial}{\partial t}\frac{\partial L}{\partial \psi_t} - \frac{\partial^2}{\partial x^2}\frac{\partial L}{\partial \psi_{xx}} - \frac{\partial L}{\partial x} = 0\,.$$

We then get the equation of motion,

$$\rho h\frac{\partial^2\psi}{\partial t^2} = -\left(\frac{Eh^3}{12}\right)\frac{\partial^4\psi}{\partial x^4}$$
$$\frac{\partial^2\psi}{\partial t^2} = -\left(\frac{Eh^2}{12\rho}\right)\frac{\partial^4\psi}{\partial x^4}\,,$$

with dispersion relation $\omega = c_{\mathrm{L}}hk^2$. Notice that these bending waves are strongly dispersive (phase velocity $\omega/k = c_{\mathrm{L}}hk$). Using $k \sim 1/\ell$ and dropping numerical constants, we have

$$f = \frac{c_{\mathrm{L}}h}{\ell^2} = \frac{c_{\mathrm{L}}}{\ell a}\,,$$

where the aspect ratio $a = \ell/h$.

6.15 Measuring a transfer function with noisy inputs. We explore the implications of input noise on transfer function measurements. The noisy input at frequency ω_ℓ is $u = u_0 + \nu$, where $\nu \sim \mathcal{N}(0, \sigma_u^2)$ and where u_0 is the (unobservable) true value of the input. Similarly, the noisy output is $y = y_0 + \xi$, where $\xi \sim \mathcal{N}(0, \sigma_y^2)$. From M input periods, we can use Eq. (6.7) to estimate the averages $\hat{u}$ and $\hat{y}$ and Eq. (6.8) to estimate the (co)-variances $\hat{\sigma}_u^2$, $\hat{\sigma}_y^2$, and $\hat{\sigma}_{yu}^2$. Note that we have simplified the notation by dropping the ω_ℓ dependence from all quantities.

a. Estimating G via $y = Gu$ is equivalent to fitting data to a straight-line relation between u and y with errors in both variables. Generalizing the Bayesian derivation of the χ^2 statistic in Appendix, Eq. (A.210), define $\boldsymbol{z} = \begin{pmatrix} \hat{y}-y_0 \\ \hat{u}-u_0 \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \hat{\sigma}_y^2 & \hat{\sigma}_{yu}^2 \\ \hat{\sigma}_{yu}^2 & \hat{\sigma}_u^2 \end{pmatrix}$ and show that the best estimate of the transfer function is given by minimizing the *errors-in-variables* cost function $\chi^2 = \sum_{\ell=1}^{N} \boldsymbol{z}^\dagger \boldsymbol{\Sigma}^{-1}\boldsymbol{z}$. If input-output correlations can be neglected, show that the general χ^2 simplifies to $\chi^2 = \sum_{\ell=1}^{N}\left(\frac{|\hat{y}-y_0|^2}{\hat{\sigma}_y^2} + \frac{|\hat{u}-u_0|^2}{\hat{\sigma}_u^2}\right)$. In both cases, we minimize χ^2 with respect to the unknown true input and output values u_0 and y_0, subject to the constraint

(at each frequency ω_ℓ) that $y_0 = G\,u_0$. Recall that the transfer function model $G = G(\mathrm{i}\omega, \boldsymbol{\theta})$, with $\boldsymbol{\theta}$ the fit parameters.

b. We can minimize the above χ^2 with respect to the transfer function parameters $\boldsymbol{\theta}$ and the *nuisance parameters* u_0 and y_0 using Lagrange multipliers to enforce the constraints. Here, in a Bayesian approach that confirms that all distributions are Gaussian, we *marginalize* (integrate out) u_0 and y_0, again limiting ourselves to the $\hat{\sigma}_{yu}^2 = 0$ case for simplicity. To carry this out, assume a uniform prior on the u_0 and y_0 and impose the constraint $y_0 = G\,u_0$. Integrate the probability density $p(G|u_0, y_0)$ over the u_0 to show that the maximum-likelihood solution minimizes $\chi^2 = \sum_{\ell=1}^{N} \frac{|\hat{y}-G\hat{u}|^2}{\hat{\sigma}_y^2+|G|^2\hat{\sigma}_u^2}$ with respect to the parameters $\boldsymbol{\theta}$ in $G(\mathrm{i}\omega_\ell, \boldsymbol{\theta})$. Interpret intuitively the denominator in this expression. The required u_0 integral is a messy version of the identity in Problem A.7.1. You can use a computer-algebra program or try the real case, which has a similar but simpler structure.

Solution.

a. We assume that the fluctuations at each point are independent (as are input and output fluctuations). Here, we write the contribution of a single point, to simplify the notation.

$$
\begin{aligned}
p(G|\hat{u},\hat{y}) &\propto p(\hat{u},\hat{y}|G)\,\underbrace{p(G)}_{\text{uniform}} \\
&= \int_{-\infty}^{\infty}\!\!\int\!\!\int \mathrm{d}u_0\,\mathrm{d}y_0\; p(\hat{u},\hat{y}|u_0,y_0)\,(u_0,y_0|G) \\
&= \int_{-\infty}^{\infty}\!\!\int\!\!\int \mathrm{d}u_0\,\mathrm{d}y_0\; p(\hat{u},\hat{y}|u_0,y_0)\,p(u_0)\,p(y_0|u_0,G) \\
&= \int_{-\infty}^{\infty}\!\!\int\!\!\int \mathrm{d}u_0\,\mathrm{d}y_0\; p(\hat{u},\hat{y}|u_0,y_0)\quad \underbrace{p(u_0)}_{\text{uniform in } u_0}\quad \delta\,(y_0 - G\,u_0) \\
&\propto \int_{-\infty}^{\infty}\!\!\int\!\!\int \mathrm{d}u_0\,\mathrm{d}y_0 \exp\left(-\tfrac{1}{2}\boldsymbol{z}^\dagger\,\boldsymbol{\Sigma}^{-1}\boldsymbol{z}\right)\,\delta\,(y_0 - G\,u_0)\,, \\
&\to \int_{-\infty}^{\infty}\!\!\int\!\!\int \mathrm{d}u_0\,\mathrm{d}y_0 \exp\left(-\frac{|\hat{y}-y_0|^2}{2\hat{\sigma}_y^2}\right)\exp\left(-\frac{|\hat{u}-u_0|^2}{2\hat{\sigma}_u^2}\right)\,\delta\,(y_0 - G\,u_0)\,,
\end{aligned}
$$

where the conditional probability $p(y_0|u_0) = \delta\,(y_0 - G\,u_0)$ expresses the constraint between u_0 and y_0. The last line is the simplification that results when the covariance $\hat{\sigma}_{yu}^2 = 0$. In this case,

$$
\begin{aligned}
\boldsymbol{z}^\dagger\,\boldsymbol{\Sigma}^{-1}\boldsymbol{z} &= \begin{pmatrix}(\hat{y}-y_0)^* & (\hat{u}-u_0)^*\end{pmatrix}\begin{pmatrix}(\hat{\sigma}_y^2)^{-1} & 0\\ 0 & (\hat{\sigma}_u^2)^{-1}\end{pmatrix}\begin{pmatrix}\hat{y}-y_0\\ \hat{u}-u_0\end{pmatrix} \\
&= \frac{|\hat{y}-y_0|^2}{\hat{\sigma}_y^2} + \frac{|\hat{u}-u_0|^2}{\hat{\sigma}_u^2}\,,
\end{aligned}
$$

and the multivariate Gaussian factors into the product of two independent Gaussian distributions.

In Part (b), we will see that $p(G|\hat{u},\hat{y})$ remains Gaussian. Then, instead of averaging over the unknown u_0 and y_0 (subject to the constraint $y_0 = G\,u_0$), we can also recognize that a uniform prior will lead simply to picking the values of u_0 and $y_0 = G\,u_0$ that *maximize* the likelihood. To maximize the likelihood, we thus *minimize*

$$\chi^2 = \sum_{\ell=1}^{N} \boldsymbol{z}^\dagger\,\boldsymbol{\Sigma}^{-1}\boldsymbol{z} \quad \text{or} \quad \chi^2 = \sum_{\ell=1}^{N}\left(\frac{|\hat{y}-y_0|^2}{\hat{\sigma}_y^2} + \frac{|\hat{u}-u_0|^2}{\hat{\sigma}_u^2}\right),$$

with respect to the u_0 and y_0, subject to the constraint that $y_0 = G\,u_0$.

b. Instead of picking u_0 and $y_0 = G\,u_0$ to minimize the χ^2, we can average over the u_0 explicitly, assuming a uniform prior. Substituting $y_0 = G\,u_0$, we have

$$p(G|\hat{u},\hat{y}) \propto \int_{-\infty}^{\infty} \mathrm{d}u_0 \exp\left(-\frac{|\hat{y}-Gu_0|^2}{2\hat{\sigma}_y^2}\right)\exp\left(-\frac{|\hat{u}-u_0|^2}{2\hat{\sigma}_u^2}\right).$$

The integral over u_0 is really a double integral over $\mathrm{Re}(u_0)$ and $\mathrm{Im}(u_0)$. To simplify the calculation and see its basic structure, we assume that all variables are real. Then the product of the exponents has the form $\exp[-\frac{1}{2}(au_0^2+bu_0+c)]$, where

$$a = \frac{1}{\hat{\sigma}_u^2} + \frac{G^2}{\hat{\sigma}_y^2}, \quad b = \frac{\hat{u}}{\hat{\sigma}_u^2} + \frac{G\hat{y}}{\hat{\sigma}_y^2}, \quad c = \frac{\hat{u}^2}{\hat{\sigma}_u^2} + \frac{\hat{y}^2}{\hat{\sigma}_y^2}.$$

We then use the identify (see Problem A.7.1) that

$$\int_{-\infty}^{\infty} \mathrm{d}u_0 \exp\left[-\frac{1}{2}\left(au_0^2 + bu_0 + c\right)\right] = \sqrt{\frac{2\pi}{a}}\,\exp\left(\frac{b^2}{2a} - \frac{c}{2}\right).$$

Thus, the distribution remains Gaussian after integrating out the nuisance parameters. Even in this simplified case, the algebra is tedious.

The full calculation is even more tedious and is perhaps better left to a computer algebra program, which confirms

$$p(G|\hat{u},\hat{y}) \propto \exp\left(-\frac{|\hat{y}-G\hat{u}|^2}{2\left(\hat{\sigma}_y^2 + |G|^2\hat{\sigma}_u^2\right)}\right).$$

Intuitively, the altered denominator reflects two sources of uncertainty: the usual one for y, which is $\hat{\sigma}_y$ and the effect of a shift in u of order $\hat{\sigma}_u$ that then further shifts y by the local slope G.

Another way of writing the χ^2 statistic that is also more intuitive is

$$\chi^2 = \frac{|\hat{y}-G\,\hat{u}|^2}{|\delta\hat{y}-G\,\delta\hat{u}|^2},$$

where

$$|\delta\hat{y} - G\,\delta\hat{u}|^2 = \hat{\sigma}_y^2 + |G|^2\hat{\sigma}_u^2 - 2\mathrm{Re}\left(\hat{\sigma}_{yu}^2 G^*\right) .$$

Here, the variation $\delta\hat{y}$ leads to $|\delta\hat{y}|^2 = \hat{\sigma}_y^2$, etc. See Problem 6.9, as well.

6.16 Time-domain identification. We go through the example presented in Section 6.3.2. Consider the first-order system $y_{k+1} = ay_k + bu_k + v_k$, with $\langle v_k\, v_\ell\rangle = v^2\,\delta_{k\ell}$.

a. Starting from the Bayesian and maximum-likelihood ideas formulated in Section A.8.2, show that the best estimate for a and b is the one that minimizes $\chi^2 = \frac{1}{v^2}\sum_k (y_{k+1} + ay_k - bu_k)^2$.
b. Minimize χ^2 to show that the best estimates for a and b are given by Eq. (6.18).
c. To show that the parameter estimates are biased but *consistent*, plot the relative bias $|\hat{a} - a|/a$, against the number of data pairs N . Confirm that the bias scales as N^{-1}, rather than the $N^{-1/2}$ scaling that is characteristic of stochastic errors.

Solution.

a. Although Eq. (6.17) resembles the structure of an ordinary linear least-squares problem (Section A.8.2), the values y_{k+1} are determined not simply by u_{k+1}, as they would ordinarily, but by y_k and u_k. Some software routines for least-squares fits can handle such an equation by using a vector-matrix form for χ^2 and by fitting "all at once" rather than "point by point." Here, we can easily solve for the best estimates of a and b directly. The arguments closely follow the standard least-squares arguments discussed in Section A.8.2. To make the notation more intuitive, we use y for the set of $\{y_k\}$, etc. From Bayes' theorem, we have

$$\begin{aligned} p(a,b|u,y) &= \int \mathrm{d}v\; p(a,b|u,y,v)p(v) \\ &\propto \int \mathrm{d}v\; p(y|a,b,u,v)\cancel{p(a,b)}\,p(v) \\ &= \left(\frac{1}{\sqrt{2\pi v^2}}\right)^N \prod_{k=1}^{N}\int \mathrm{d}v_k\,\delta\left(y_{k+1} + ay_k - bu_k - v_k\right)\exp\left(-v_k^2/2v^2\right) \\ &= \frac{1}{(2\pi v^2)^{N/2}}\exp\left(-\tfrac{1}{2}\chi^2\right) , \end{aligned}$$

with

$$\chi^2 = \frac{1}{v^2}\sum_{k=1}^{N}(y_{k+1} + ay_k - bu_k)^2 .$$

b. Taking $\partial_a\chi^2 = \partial_b\chi^2 = 0$ gives the desired equations. For example,

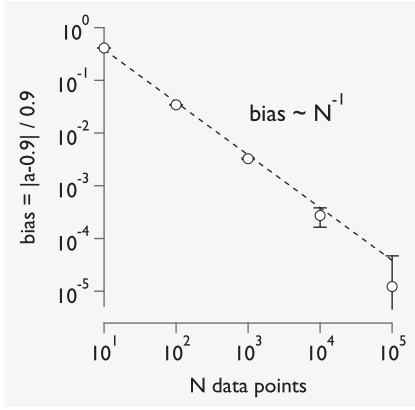
$$\frac{\partial\chi^2}{\partial a} = \frac{2}{v^2}\sum_{k=1}^{N}(y_{k+1} + ay_k - bu_k)\;y_k = 0 ,$$

so that

$$\sum_k y_{k+1}\, y_k + a \sum_k y_k^2 - b \sum_k u_k\, y_k = 0\,.$$

We then assemble the two equations into a single matrix equation for the parameter vector estimate $(\hat{a}\ \ \hat{b})^\mathsf{T}$.

c. We find the graph below, where the dotted line has slope −1.



6.17 AIC and cross-validation. Consider N measurements of $y = \theta^* x + \eta$, with independent scalar variable x, observed variable y, and parameter θ^*. The measurement noise $\eta \sim \mathcal{N}(0, \sigma^2)$, and the log likelihood is $L(\theta) = -\frac{1}{2\sigma^2} \sum_{j=1}^N (y_j - \theta x_j)^2$.

a. Show that the maximum-likelihood (ML) estimate is $\hat{\theta} = \sum_{j=1}^N (x_j\, y_j) / \sum_{j=1}^N (x_j^2)$.
b. Now consider the same data set, but without point i. Show that, to $O(1/N)$, the ML estimate of θ is $\hat{\theta}_{-i}$, where $\hat{\theta}_{-i} = \hat{\theta} - \frac{x_i}{\sum x_j^2}(y_i - \hat{\theta} x_i)$.
c. In one-point cross-validation, we calculate the likelihood of a missing point using $\hat{\theta}_{-i}$ and then average over all points. Define $A \equiv -\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - \hat{\theta}_{-i} x_i)^2$ as an assessment value and show that $A = L(\hat{\theta}) - 1 + O(1/N)$.

The case with K parameters proceeds similarly and leads to $A = L(\hat{\theta}) - K + O(1/N)$.

Solution.

a. We have

$$\frac{\partial L}{\partial \theta} = +\frac{1}{\sigma^2} \sum_j \left(y_j - \theta x_j\right) x_j = 0\,.$$

Thus,

$$\sum_j \left(x_j y_j\right) - \theta \sum_j \left(x_j^2\right) = 0\,.$$

Solving for $\theta = \hat{\theta}$ gives the desired expression.

b. If we eliminate the point i, the ML estimate is

$$\begin{aligned}\hat{\theta}_{-i} &= \frac{\sum_{j=1}^{N} x_j y_j - x_i y_i}{\sum_{j=1}^{N} x_j^2 - x_i^2} = \frac{\hat{\theta} - \frac{x_i y_i}{\sum x_j^2}}{1 - \frac{x_i^2}{\sum x_j^2}} \\ &\approx \hat{\theta} + \frac{\hat{\theta} x_i^2 - x_i y_i}{\sum x_j^2} \\ &= \hat{\theta} - \frac{x_i}{\sum x_j^2}\left(y_i - \hat{\theta} x_i\right) .\end{aligned}$$

c. We first note that the correction to $\hat{\theta}$ in the expression for $\hat{\theta}_{-i}$ is $O(1/N)$ and that subsequent terms increase in order by $1/N$. As a result, the assessment score A,

$$\begin{aligned}A &= -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \hat{\theta}_{-i} x_i\right)^2 \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left[(y_i - \hat{\theta} x_i) + \frac{x_i^2}{\sum x_j^2}(y_i - \hat{\theta} x_i)\right]^2 \\ &\approx -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left[(y_i - \hat{\theta} x_i)^2 + \frac{2x_i^2}{\sum x_j^2}(y_i - \hat{\theta} x_i)^2 + O(1/N)\right] \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \hat{\theta} x_i)^2 - \frac{1}{\sigma^2}\sum_{i=1}^{N}\frac{(y_i - \hat{\theta} x_i)^2 x_i^2}{\sum x_j^2} + O(1/N) \\ &\to L(\hat{\theta}) - \frac{1}{\sigma^2}\frac{\sum_i \eta_i^2 x_i^2}{\sum_j x_j^2} + O(1/N) \\ &\to L(\hat{\theta}) - \frac{1}{\sigma^2}\frac{N\sigma^2\langle x^2\rangle}{N\langle x^2\rangle} + O(1/N) \\ &= L(\hat{\theta}) - 1 + O(1/N) .\end{aligned}$$

Note that as $N \to \infty$, $\hat{\theta} \to \theta^*$, so that $(y_i - \hat{\theta} x_i) \to (y_i - \theta^* x_i) = \eta_i$. In this step, we assume that the model structure, θx, includes the true model, $\theta^* x$. We can then write $\sum_i \eta_i^2 x_i^2 \to N\sigma^2\langle x^2\rangle$, as well as $\sum x_j^2 \to N\langle x^2\rangle$. The angle brackets $\langle\cdot\rangle$ are averages over the distribution of x, which can be arbitrary.

6.18 $\langle\chi^2\rangle$ **for an orthonormal basis and model mismatch**. Consider a model function with an orthonormal basis set $\{e_k\}$, with $y^* = \sum_{k=1}^{\infty}(y^* \cdot e_k)\, e_k$ and $e_k \cdot e_\ell \equiv \frac{1}{N}\sum_{i=1}^{N} e_k(x_i)\, e_\ell(x_i) = \delta_{k\ell}$. For N points and K parameters, let $y = y^* + \xi$, with $\langle\xi^2\rangle = \sigma^2$, and show that $\langle\chi^2\rangle = \langle\frac{N}{\sigma^2}\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2\rangle = (N-K) + \frac{N}{\sigma^2}\sum_{\ell=K+1}^{\infty}(y^* \cdot e_\ell)\, e_\ell$. Hint: Subtract and add the true vector $\boldsymbol{y}^*$. The N in the definition of χ^2 is traditional.

Solution.

Following the hint, we write

$$\begin{aligned}\chi^2 &= \frac{N}{\sigma^2}||\boldsymbol{y}-\boldsymbol{y}^*+\boldsymbol{y}^*-\hat{\boldsymbol{y}}||^2\\ &= \frac{N}{\sigma^2}\left[||\boldsymbol{y}-\boldsymbol{y}^*||^2+2(\boldsymbol{y}-\boldsymbol{y}^*)\cdot(\boldsymbol{y}^*-\hat{\boldsymbol{y}})+||\boldsymbol{y}^*-\hat{\boldsymbol{y}}||^2\right]\\ &= \frac{N}{\sigma^2}\left[||\boldsymbol{\xi}||^2+2\boldsymbol{\xi}\cdot(\boldsymbol{y}^*-\hat{\boldsymbol{y}})+||\boldsymbol{y}^*-\hat{\boldsymbol{y}}||^2\right].\end{aligned}$$

The deviation between the true values $\boldsymbol{y}^*$ and the estimate $\hat{\boldsymbol{y}}$ is

$$\begin{aligned}\boldsymbol{y}^*-\hat{\boldsymbol{y}} &= \sum_{k=1}^{\infty}(\boldsymbol{y}^*\cdot e_k)\,e_k-\sum_{k=1}^{K}[(\boldsymbol{y}^*+\boldsymbol{\xi})\cdot e_k]e_k\\ &= \sum_{\ell=K+1}^{\infty}(\boldsymbol{y}^*\cdot e_\ell)\,e_\ell-\sum_{k=1}^{K}(\boldsymbol{\xi}\cdot e_k)\,e_k\,,\end{aligned}$$

so that

$$N||\boldsymbol{y}^*-\hat{\boldsymbol{y}}|| = \sum_{\ell=K+1}^{\infty}(\boldsymbol{y}^*\cdot e_\ell)^2+\sum_{k=1}^{K}(\boldsymbol{\xi}\cdot e_k)^2\,.$$

The mean value of χ^2 is given by

$$\begin{aligned}\left\langle\chi^2\right\rangle &= \frac{N}{\sigma^2}\left[\sigma^2+2\cancelto{0}{\left\langle\boldsymbol{\xi}\cdot\sum_{\ell=K+1}^{\infty}(\boldsymbol{y}^*\cdot e_\ell)\,e_\ell\right\rangle}-2\left\langle\boldsymbol{\xi}\cdot\sum_{k=1}^{K}(\boldsymbol{\xi}\cdot e_k)\,e_k\right\rangle\right.\\ &\quad\left.+\sum_{\ell=K+1}^{\infty}(\boldsymbol{y}^*\cdot e_\ell)^2+\left\langle\sum_{k=1}^{K}(\boldsymbol{\xi}\cdot e_k)^2\right\rangle\right].\\ &= \frac{N}{\sigma^2}\left[\sigma^2+0-\frac{2}{N}K\sigma^2+\sum_{\ell=K+1}^{\infty}(\boldsymbol{y}^*\cdot e_\ell)^2+\frac{K}{N}\sigma^2\right]\\ &= (N-K)+\frac{N}{\sigma^2}\sum_{\ell=K+1}^{\infty}(\boldsymbol{y}^*\cdot e_\ell)^2\,.\end{aligned}$$

We found the first term when we fit to a correct model (Section A.8.5). We term it a "stochastic" contribution since it is the average of a random variable. The second term is due to model mismatch and is deterministic. We have also used

$$\boldsymbol{\xi}\cdot e_k = \frac{1}{N}\sum_{i=1}^{N}\xi_i\,e_k(x_i)\,,$$

so that

$$\begin{aligned}\left\langle\boldsymbol{\xi}\cdot\sum_{k=1}^{K}(\boldsymbol{\xi}\cdot e_k)\,e_k\right\rangle &= \left\langle\sum_{k=1}^{K}\left(\frac{1}{N}\sum_{i=1}^{N}\xi_i e_k(x_i)\right)\left(\frac{1}{N}\sum_{j=1}^{N}\xi_j e_k(x_j)\right)\right\rangle\\ &= \frac{1}{N^2}\sum_{k=1}^{K}\sum_{i,j=1}^{N}e_k(x_i)\,e_k(x_j)\langle\xi_i\,\xi_j\rangle\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sigma^2}{N^2} \sum_{k=1}^{K} \sum_{i,j=1}^{N} e_k(x_i)\, e_k(x_j) \delta_{ij} \\
&= \frac{\sigma^2}{N^2} \sum_{k=1}^{K} \sum_{i=1}^{N} [e_k(x_i)]^2 \\
&= \frac{\sigma^2}{N} \sum_{k=1}^{K} (1) \\
&= \frac{K}{N} \sigma^2 \,.
\end{aligned}
$$

This solution is thanks to Antoine Baker.

6.19 AIC vs. BIC example. We work through the details of Example 6.8.

a. Write code to reproduce the plots in Example 6.8. First, simulate the data set itself from the true function $f(t) = \frac{\pi^2}{8}(1 - 2t)$, $0 < t < 1$, adding Gaussian noise of $\sigma = 0.01$. Then calculate the first 500 Fourier coefficients and the corresponding χ^2, AIC, and BIC statistics for each order.
b. Calculate the signal-to-noise ratio (SNR) of Example 6.8. Show that for large K, we have $p(\text{SNR}) = \frac{1}{\sqrt{2\pi \text{SNR}}} \mathrm{e}^{-\text{SNR}/2}$, where $\text{SNR} = \theta_j^2/(\sigma^2/2N)$. The extra factor of 2 comes from the normalization of the basis vectors.
c. By approximating a sum by an integral, show that an asymptotic, large N, analytic approximation for the model-mismatch term, accurate enough for $N > 2$, is given by $\chi^2_{\text{mm}} = \frac{1}{96\sigma^2(K+1)^3}$. Add the result to the stochastic contribution, $1 - K/N$, to generate the solid curves in the bottom three plots in the example.

Solution.

a. See the margin plots in Example 6.8.
b. The projections of noise onto noise are Gaussian random variables $y \sim \mathcal{N}(0, 1)$. The SNR then is distributed as y^2. From Example A.17, we indeed have

$$
p(s) = \frac{1}{\sqrt{2\pi s}} \mathrm{e}^{-s/2} \qquad s > 0 \,.
$$

We can verify that $\int_0^\infty \mathrm{d}s\, p(s) = \int_0^\infty \mathrm{d}s\, s\, p(s) = 1$ and $\int_0^\infty \mathrm{d}s\, s^2\, p(s) = 3$. The variance is then $\langle s^2 \rangle - \langle s \rangle^2 = 3 - 1 = 2$, implying a standard deviation of $\sqrt{2}$. The wide spread of the values seen in the bottom plot in the margin figure of Example 6.8d for Fourier coefficients of order greater than 100 (say) is compatible with this law, as a comparison of the normalized coefficient histogram with the pdf readily shows.

c. To get the analytic approximation, we note that

$$\sum_{j=K+1}^{\infty} \frac{1}{(2j-1)^4} \approx \int_{K+1}^{\infty} \frac{\mathrm{d}x}{(2x)^4} = \frac{1}{2^4\, 3(K+1)^3}$$
$$= \frac{1}{48(K+1)^3}\,.$$

Then there is an extra factor of $\frac{1}{2}$ because basis elements are not unit amplitude: $\int_{-1}^{1} \mathrm{d}t \cos^2 j\pi t = \frac{1}{2}$.

6.20 Balanced coordinates. Show that, for a system $G = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}\}$, that we can choose a coordinate transformation $\boldsymbol{T}$, with $\boldsymbol{x} = \boldsymbol{T}\boldsymbol{x}'$ such that $\boldsymbol{P}' = \boldsymbol{Q}' = \boldsymbol{\Sigma}$. Here, the Gramians are over $t = (0, \infty)$ and $\boldsymbol{\Sigma}$ is the diagonal matrix of Hankel singular values.

a. For the coordinate transformation $\boldsymbol{x} = \boldsymbol{T}\boldsymbol{x}'$, show that $\boldsymbol{A}' = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}$, $\boldsymbol{B}' = \boldsymbol{T}^{-1}\boldsymbol{B}$, $\boldsymbol{C}' = \boldsymbol{C}\boldsymbol{T}$. Then show that $\boldsymbol{P}' = \boldsymbol{T}^{-1}(\boldsymbol{P}\boldsymbol{T}^{-1})^{\mathsf{T}}$ and $\boldsymbol{Q}' = \boldsymbol{T}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{T}$.

b. Decompose $\boldsymbol{P} = \boldsymbol{R}\boldsymbol{R}^{\mathsf{T}}$ using Cholesky decomposition (Section A.1.5), and write $\boldsymbol{R}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{R} = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^{\mathsf{T}}$, and $\boldsymbol{T} = \boldsymbol{R}\boldsymbol{U}\boldsymbol{\Sigma}^{-1/2}$. Show that $\boldsymbol{P}' = \boldsymbol{Q}' = \boldsymbol{\Sigma}$.

c. Consider the example from Figure 6.7: $\boldsymbol{A} = \left(\begin{smallmatrix} -1 & 0 \\ 0 & -2 \end{smallmatrix}\right)$, $\boldsymbol{B} = \left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right)$, and $\boldsymbol{C} = (\,1\;2\,)$.

i. Find the infinite-time Gramians $\boldsymbol{P}$, $\boldsymbol{Q}$. Construct a balanced representation.

ii. Find symbolically or numerically $\boldsymbol{R}$, $\boldsymbol{U}$, $\boldsymbol{T}$, $\boldsymbol{\Sigma}$ and also $\boldsymbol{A}'$, $\boldsymbol{B}'$, and $\boldsymbol{C}'$. You will want to use a computer-algebra program for this part and the next.

iii. In the new coordinate system, verify that $\boldsymbol{P}' = \boldsymbol{Q}' = \boldsymbol{\Sigma}$, with the diagonal elements of $\boldsymbol{\Sigma}$ being $\sigma_{\pm} = \frac{1}{2} \pm \frac{\sqrt{2}}{3}$. (If you do not have access to symbolic-manipulation software, do this part numerically.)

Solution.

a. Substituting gives

$$\boldsymbol{T}\boldsymbol{x}' = \boldsymbol{A}\boldsymbol{T}\boldsymbol{x}' + \boldsymbol{B}u$$
$$\boldsymbol{x}' = \underbrace{\boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}}_{\boldsymbol{A}'}\boldsymbol{x} + \underbrace{\boldsymbol{T}^{-1}\boldsymbol{B}}_{\boldsymbol{B}'}u\,,$$

so that $\boldsymbol{A}' = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}$, $\boldsymbol{B}' = \boldsymbol{T}^{-1}\boldsymbol{B}$, and $\boldsymbol{C}' = \boldsymbol{C}\boldsymbol{T}$.

To find out how the controllability Gramian $\boldsymbol{P} \equiv \boldsymbol{P}(\infty)$ transforms, we write

$$\begin{aligned}
\boldsymbol{P}' &= \int_0^{\infty} \mathrm{d}t\, \mathrm{e}^{\boldsymbol{A}'t}\, \boldsymbol{B}'\boldsymbol{B}'^{\mathsf{T}}\, \mathrm{e}^{\boldsymbol{A}^{\mathsf{T}}t} \\
&= \boldsymbol{T}^{-1} \int_0^{\infty} \mathrm{d}t\, \mathrm{e}^{\boldsymbol{A}t}\, \boldsymbol{T}\boldsymbol{T}^{-1}\boldsymbol{B}\boldsymbol{B}^{\mathsf{T}}(\boldsymbol{T}^{-1})^{\mathsf{T}}\boldsymbol{T}^{\mathsf{T}}\, \mathrm{e}^{\boldsymbol{A}t}(\boldsymbol{T}^{-1})^{\mathsf{T}}\, dt \\
&= \boldsymbol{T}^{-1} \int_0^{\infty} \mathrm{d}t\, \mathrm{e}^{\boldsymbol{A}t}\, \boldsymbol{B}\boldsymbol{B}^{\mathsf{T}}\, \mathrm{e}^{\boldsymbol{A}t}\, (\boldsymbol{T}^{-1})^{\mathsf{T}} \\
&= \boldsymbol{T}^{-1}(\boldsymbol{P}\boldsymbol{T}^{-1})^{\mathsf{T}}\,.
\end{aligned}$$

Similarly, for the observability matrix,

$$\begin{aligned}\boldsymbol{Q}' &= \int_0^\infty \mathrm{d}t\, \mathrm{e}^{\boldsymbol{A}'^{\mathsf{T}}t}\boldsymbol{C}'^{\mathsf{T}}\boldsymbol{C}'\,\mathrm{e}^{\boldsymbol{A}'t}\\ &= \int_0^\infty \mathrm{d}t\, \boldsymbol{T}^{\mathsf{T}}\,\mathrm{e}^{\boldsymbol{A}^{\mathsf{T}}t}(\boldsymbol{T}^{-1})^{\mathsf{T}}\boldsymbol{T}^{\mathsf{T}}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{T}\boldsymbol{T}^{-1}\,\mathrm{e}^{\boldsymbol{A}t}\,\boldsymbol{T}\\ &= \boldsymbol{T}^{\mathsf{T}}\int_0^\infty \mathrm{d}t\,\mathrm{e}^{\boldsymbol{A}^{\mathsf{T}}t}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{C}\,\mathrm{e}^{\boldsymbol{A}t}\,\boldsymbol{T}\\ &= \boldsymbol{T}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{T}\,.\end{aligned}$$

b. First, we write

$$\boldsymbol{T}^{-1} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{U}^{-1}\boldsymbol{R}^{-1} \qquad \boldsymbol{T}^{\mathsf{T}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{U}^{T}\boldsymbol{R}^{T} \qquad (\boldsymbol{T}^{-1})^{\mathsf{T}} = (\boldsymbol{R}^{-1})^{\mathsf{T}}(\boldsymbol{U}^{-1})^{\mathsf{T}}\boldsymbol{\Sigma}^{1/2}\,.$$

For the controllability Gramian, we write

$$\begin{aligned}\boldsymbol{P}' &= \boldsymbol{T}^{-1}(\boldsymbol{P}\boldsymbol{T}^{-1})^{\mathsf{T}}\\ &= (\boldsymbol{\Sigma}^{1/2}\boldsymbol{U}^{-1}\boldsymbol{R}^{-1})(\boldsymbol{R}\boldsymbol{R}^{\mathsf{T}})[(\boldsymbol{R}^{-1})^{\mathsf{T}}(\boldsymbol{U}^{-1})^{\mathsf{T}}\boldsymbol{\Sigma}^{1/2}]\\ &= \boldsymbol{\Sigma}\,,\end{aligned}$$

noting that $\boldsymbol{U}$ is unitary, so that $\boldsymbol{U}^{-1} = \boldsymbol{U}^{\mathsf{T}}$. For the observability Gramian $\boldsymbol{Q} = \boldsymbol{R}^{-T}\boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^{\mathsf{T}}\boldsymbol{R}^{-1}$, we write

$$\begin{aligned}\boldsymbol{Q}' &= \boldsymbol{T}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{T}\\ &= (\boldsymbol{\Sigma}^{-1/2}\boldsymbol{U}^{T}\boldsymbol{R}^{T})((\boldsymbol{R}^{-1})^{\mathsf{T}}\boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^{\mathsf{T}}\boldsymbol{R}^{-1})(\boldsymbol{R}\boldsymbol{U}\boldsymbol{\Sigma}^{-1/2})\\ &= \boldsymbol{\Sigma}\,,\end{aligned}$$

c. i. The Gramians are easily evaluated exactly:

$$\boldsymbol{P} = \begin{pmatrix}\frac{1}{2} & \frac{1}{3}\\ \frac{1}{3} & \frac{1}{4}\end{pmatrix} \qquad \boldsymbol{Q} = \begin{pmatrix}\frac{1}{2} & \frac{2}{3}\\ \frac{2}{3} & 1\end{pmatrix}.$$

ii. With Mathematica, the rest of the matrices are

$$\boldsymbol{R} = \begin{pmatrix}\frac{1}{\sqrt{2}} & 0\\ \frac{\sqrt{2}}{3} & \frac{1}{6}\end{pmatrix} \approx \begin{pmatrix}0.707 & 0\\ 0.471 & 0.167\end{pmatrix} \qquad \boldsymbol{R}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{R} = \begin{pmatrix}\frac{11}{12} & \frac{\sqrt{2}}{9}\\ \frac{\sqrt{2}}{9} & \frac{1}{36}\end{pmatrix} \approx \begin{pmatrix}0.917 & 0.157\\ 0.157 & 0.028\end{pmatrix}.$$

The singular value decomposition of $\boldsymbol{R}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{R}$ is, with $\sigma_\pm^2 = \frac{17}{36} \pm \frac{\sqrt{2}}{3}$,

$$\begin{aligned}\boldsymbol{R}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{R} &= \begin{pmatrix}\sqrt{\sigma_+} & -\sqrt{\sigma_-}\\ \sqrt{\sigma_-} & \sqrt{\sigma_+}\end{pmatrix} \begin{pmatrix}\sigma_+^2 & 0\\ 0 & \sigma_-^2\end{pmatrix} \begin{pmatrix}\sqrt{\sigma_+} & \sqrt{\sigma_-}\\ -\sqrt{\sigma_-} & \sqrt{\sigma_+}\end{pmatrix}\\ &= \underbrace{\begin{pmatrix}0.986 & -0.169\\ 0.169 & 0.986\end{pmatrix}}_{\boldsymbol{U}} \underbrace{\begin{pmatrix}0.944 & 0\\ 0 & 0.00082\end{pmatrix}}_{\boldsymbol{\Sigma}^2} \underbrace{\begin{pmatrix}0.986 & 0.169\\ -0.169 & 0.986\end{pmatrix}}_{\boldsymbol{U}^{\mathsf{T}}}\end{aligned}$$

The transformation matrix $\boldsymbol{T}$ is then

$$\boldsymbol{T} = \tfrac{1}{2}\begin{pmatrix}\sqrt{2} & -\sqrt{2}\\ 1 & 1\end{pmatrix} \approx \begin{pmatrix}0.707 & -0.707\\ 0.5 & 0.5\end{pmatrix}.$$

In balanced coordinates, the system matrices are

$$\boldsymbol{A}' = \begin{pmatrix} \frac{-3}{2} & \frac{-1}{2} \\ \frac{-1}{2} & \frac{-3}{2} \end{pmatrix}, \qquad \boldsymbol{B}' = \begin{pmatrix} 1 + \frac{1}{\sqrt{2}} \\ 1 - \frac{1}{\sqrt{2}} \end{pmatrix}, \qquad \boldsymbol{C}' = \begin{pmatrix} 1 + \frac{1}{\sqrt{2}} & 1 - \frac{1}{\sqrt{2}} \end{pmatrix}.$$

Note that the eigenvalues of $\boldsymbol{A}'$ remain $(-1, -2)$, as all we have done is make a similarity transformation.

iii. Finally, it is straightforward, either symbolically or numerically, to show that in the new system $\{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}\}$, we have

$$\boldsymbol{P}' = \boldsymbol{Q}' = \boldsymbol{\Sigma} = \begin{pmatrix} \frac{1}{2} + \frac{\sqrt{2}}{3} & 0 \\ 0 & \frac{1}{2} - \frac{\sqrt{2}}{3} \end{pmatrix} \approx \begin{pmatrix} 0.971 & 0 \\ 0 & 0.029 \end{pmatrix}.$$

PART II

ADVANCED IDEAS

Optimal Control 7

Problems

7.1 **One-dimensional optimization, without shortcuts**. Redo the example in Section 7.1 without assuming that $u(t) = -Kx(t)$. Substitute the equation of motion $u = \dot{x} + ax$ into $L = \frac{1}{2}(x^2 + Ru^2)$, and find $x(t)$ directly. Confirm the assumed form of $u(t)$.

Solution. The cost function is

$$\begin{aligned} J &= \tfrac{1}{2}\int_0^\infty \mathrm{d}t\left(x^2(t) + Ru^2(t)\right) \\ &= \tfrac{1}{2}\int_0^\infty \mathrm{d}t\left(x^2(t) + R(\dot{x} + ax)^2(t)\right) \\ &\equiv \int_0^\infty \mathrm{d}t\, L(x, \dot{x})\,. \end{aligned}$$

The Euler-Lagrange equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{x}}\right) = \frac{\partial L}{\partial x}\,.$$

is then

$$R(\ddot{x} + \cancel{a\dot{x}}) = x + R(\cancel{\dot{x}} + ax)(a)\,,$$

which gives

$$\ddot{x} = \left(a^2 + \tfrac{1}{R}\right)x\,.$$

Since we want $x(t)$ finite as $t \to \infty$, we take only the stable solution. Thus, using the linearity of the equations of motion to scale the initial condition $x(0) = x_0$ to one, we find

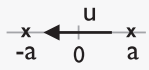
$$x(t) = \mathrm{e}^{-\sqrt{a^2+\frac{1}{R}}\,t}\,,$$

which implies

$$\begin{aligned} u(t) &= \dot{x} + ax \\ &= -\sqrt{a^2 + \tfrac{1}{R}}\ \mathrm{e}^{-\sqrt{a^2+\frac{1}{R}}\,t} + a\ \mathrm{e}^{-\sqrt{a^2+\frac{1}{R}}\,t} \end{aligned}$$

$$= -\left(\sqrt{a^2 + \tfrac{1}{R}} - a\right) x(t)$$
$$\equiv -K^* x(t) .$$

Thus, we recover both the form of the feedback law (negative proportional control) and the value K^* that optimizes the control. Compare Eq. (7.3).



7.2 **Unstable system**. Repeat the example in Sec. 7.1 for an unstable system, $\dot{x} = +ax + u$, with $a > 0$. Compare with the stable case. Discuss the cheap ($R \to 0$) and expensive ($R \to \infty$) control limits. Show that expensive control leads to a gain that replaces the unstable eigenvalue a with its stable "mirror image" at $-a$. This is a general result.

Solution.

The problem is the same as the example in Sec. 7.1, except that $a \to -a$. Thus,

$$K^* = \sqrt{a^2 + 1/R} + a .$$

As with the original example, the equation for K^* clearly has two regimes, depending on R:

- $Ra^2 \ll 1$, *cheap control*: Then $K \sim 1/\sqrt{R}$, independent of the sign of a. This makes sense: when control is cheap, we apply so much gain that it does not matter what the original dynamics looks like. Differences are "washed out."
- $Ra^2 \gg 1$, *expensive control*: Then $K \to 2a$ and $u = -2a\,x$. The open-loop system $\dot{x} = +ax$ is thus replaced by the closed-loop system $\dot{x} = -ax$. This clearly stabilizes the system. Because control "effort" $\sim u^2(t)$, we can see that $K = 2a$ is also the minimum-control-effort solution, in the sense that $K(R) > 2a$ for finite R. Intuitively, even though a K in the range $(a, 2a)$ also stabilizes the system and with less gain, the resulting excursions are large enough that the net penalty in J resulting from larger values of u is greater than with the optimal gain $K = 2a$.

  Contrast this result with that of the stable case, where $K \to 0$ and $u \to 0$. That is, when $R \to \infty$, the gain $K \to 0$. If control is too costly, do nothing. But this strategy works only when the underlying system is stable and can relax on its own. If not, we need to apply a minimum gain, no matter how much control costs.

  The conclusion that the minimum-control-effort solution amounts to a control that replaces poles in the right-hand s-plane with their "mirror images" in the left-hand plane turns out to be a general one for linear systems. (More precisely, the recipe is to replace a pole $p' + ip''$ with $-p' + ip''$, for each pole with $p' = \mathrm{Re}\ p > 0$.)

7.3 **One-dimensional control, with initial and final state conditions**. Solve the coupled linear equations for $x(t)$ and $\lambda(t)$ in Example 7.1. Show, in particular, that $x(t) = x_\tau\left(\frac{\sinh \sqrt{2}t}{\sinh \sqrt{2}\tau}\right)$ and $\lambda(t) = -(1 + \sqrt{2}\coth\sqrt{2}t)\,x(t)$, implying that

$u(t) = -\lambda(t) = +K(t)\,x(t)$ can be expressed as a positive feedback with time-dependent gain. Explain the behavior of λ and K for $\tau \to 0$ and $\tau \to \infty$.

Solution.

The equations of motion are

$$\dot{x} = -x + u\,, \qquad \dot{\lambda} = +\lambda - x\,, \qquad u = -\lambda\,.$$

with boundary conditions $x(0) = 0$ and $x(\tau) = x_\tau$. The systematic way to solve such an equation is to eliminate u in favor of λ and express the two coupled equations in vector-matrix form:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x \\ \lambda \end{pmatrix} = \underbrace{\begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix}}_{\boldsymbol{A}}\begin{pmatrix} x \\ \lambda \end{pmatrix},$$

where the matrix $\boldsymbol{A}$ has eigenvalues $\pm\sqrt{2}$ and eigenvectors $\binom{1-\sqrt{2}}{1}$ for $+\sqrt{2}$ and $\binom{1+\sqrt{2}}{1}$ for $-\sqrt{2}$. Thus, the solution for $x(t)$ is of the form

$$x(t) = \alpha\left(1 - \sqrt{2}\right)\mathrm{e}^{\sqrt{2}t} + \beta\left(1 + \sqrt{2}\right)\mathrm{e}^{-\sqrt{2}t}\,,$$

where the constants α and β are fixed by the boundary conditions at $t = 0$ and $t = \tau$:

$$x(0) = 0 = \alpha\left(1 - \sqrt{2}\right) + \beta\left(1 + \sqrt{2}\right) \quad \Longrightarrow \quad \beta = \alpha\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right),$$

so that

$$\begin{aligned} x(t) &= \alpha\left[\left(1 - \sqrt{2}\right)\mathrm{e}^{\sqrt{2}t} + \left(\sqrt{2} - 1\right)\mathrm{e}^{-\sqrt{2}t}\right] \\ &= \alpha\left(1 - \sqrt{2}\right)2\sinh\sqrt{2}t\,. \end{aligned}$$

Thus,

$$x(\tau) = x_\tau = \alpha\left(1 - \sqrt{2}\right)2\sinh\sqrt{2}\tau \quad \Longrightarrow \quad \alpha = \frac{x_\tau}{\left(1 - \sqrt{2}\right)2\sinh\sqrt{2}\tau}\,.$$

Finally,

$$x(t) = x_\tau\left(\frac{\sinh\sqrt{2}t}{\sinh\sqrt{2}\tau}\right).$$

To find the adjoint, we use the other part of the matrix solution:

$$\begin{aligned} \lambda(t) &= \alpha\ \mathrm{e}^{\sqrt{2}t} + \beta\ \mathrm{e}^{-\sqrt{2}t} \\ &= \alpha\left(\mathrm{e}^{\sqrt{2}t} + \left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right)\mathrm{e}^{-\sqrt{2}t}\right) \\ &= \left(\frac{\alpha}{\sqrt{2}+1}\right)\left[\left(\sqrt{2}+1\right)\mathrm{e}^{\sqrt{2}t} + \left(\sqrt{2}-1\right)\mathrm{e}^{-\sqrt{2}t}\right] \end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{\alpha}{\sqrt{2}+1}\right)\left(2\sqrt{2}\cosh\sqrt{2}t + 2\sinh\sqrt{2}t\right) \\
&= -x_\tau\left(\frac{\sqrt{2}\cosh\sqrt{2}t + \sinh\sqrt{2}t}{\sinh\sqrt{2}\tau}\right) \\
&= -x(t)\left(\frac{\sqrt{2}\cosh\sqrt{2}t + \sinh\sqrt{2}t}{\sinh\sqrt{2}t}\right) \\
&= -x(t)\left(1 + \sqrt{2}\coth\sqrt{2}t\right) .
\end{aligned}
$$

Since $u = -\lambda$, this has the form of a *positive* feedback term, with gain

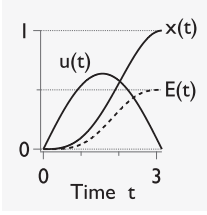
$$K(t) = \left(1 + \sqrt{2}\coth\sqrt{2}t\right) ,$$

which tends to $1 + \sqrt{2}$ for large t (and large τ). The gain diverges at short times $t \to 0$, which makes sense: we need a high gain to "get the system started."

For $\tau \to 0$, $\lambda(t) \sim x_\tau/\tau$, which diverges. This makes sense: moving x from 0 to x_τ over a shorter cycle should require more "work."

7.4 **Move a harmonic oscillator**. For an undamped, simple harmonic oscillator, let the goal be to move from rest states $x = 0$ to $x = 1$ using the least control effort:

$$\ddot{x} + x = u\,, \quad x(0) = \dot{x}(0) = 0\,, \quad x(\tau) = 1\,,\ \dot{x}(\tau) = 0\,, \qquad J = \int_0^\tau \mathrm{d}t\left(\tfrac{1}{2}u^2(t)\right) .$$



a. Solve the problem analytically, and find expressions for $x(t)$ and $u(t)$. Find also $J(\tau)$ and show that $J_{\text{short}}(\tau) \sim 12/\tau^3$ and $J_{\text{long}}(\tau) \sim 2/\tau$.

b. Plot $x(t)$ and $u(t)$ for $\tau = 1$, 2, π, 2π, 3π, and 10π and discuss. A sample plot for $\tau = \pi$ is shown at left along with the energy $E(t) = \frac{1}{2}(x^2 + \dot{x}^2)$. What if $\tau < \pi$?

c. Plot $J(\tau)$, with its short- and long-protocol limits, and discuss.

Solution.

a. This is a modified LQR problem, with variable boundary conditions. Let us put the problem in LQR form. The two-dimensional state vector is

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \equiv \begin{pmatrix} x \\ \dot{x} \end{pmatrix} .$$

From Eq. (7.13), the equations of motion of this SISO system are

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}}_{\boldsymbol{A}}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{\boldsymbol{B}} u\,, \qquad y = \underbrace{\begin{pmatrix} 1 & 0 \end{pmatrix}}_{\boldsymbol{C}}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} ,$$

with boundary conditions

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}_{t=0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} , \qquad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}_{t=\tau} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} .$$

The adjoint equations are

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\lambda_1\\ \lambda_2\end{pmatrix} = \underbrace{\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}}_{-\boldsymbol{A}^{\mathsf{T}}}\begin{pmatrix}\lambda_1\\ \lambda_2\end{pmatrix}.$$

As in Problem 7.3, there are no boundary conditions on $\boldsymbol{\lambda}$, since they are imposed on $\boldsymbol{x}$ at the beginning and end of the protocol. The equation for the control u is,

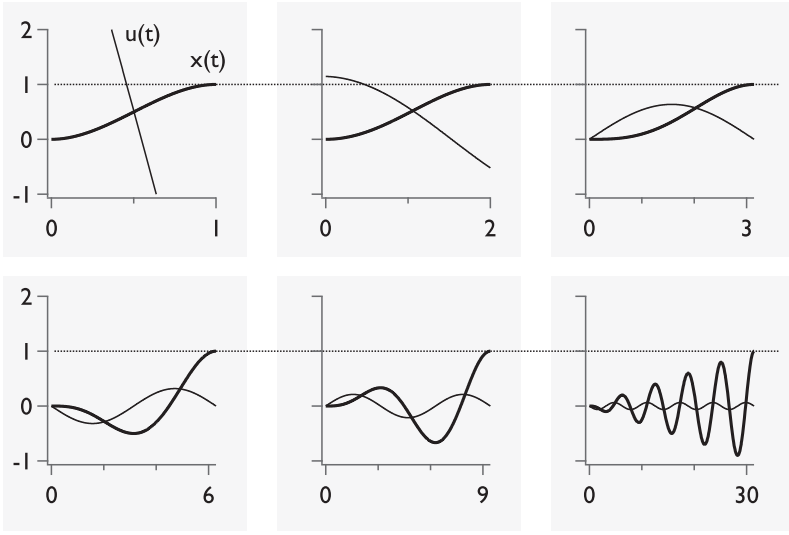
$$u = -\underbrace{\begin{pmatrix}0 & 1\end{pmatrix}}_{\boldsymbol{B}^{\mathsf{T}}}\begin{pmatrix}\lambda_1\\ \lambda_2\end{pmatrix} = -\lambda_2$$

To solve these equations, we can immediately see that $\lambda(t) = a\cos t + b\sin t$. We could then substitute this equation into the equation for $x(t)$ and solve the forced harmonic oscillator equation. The forcing is resonant, leading to secular terms of the form $t\cos t$ and $t\sin t$. The final step would be to impose the four boundary conditions on $x(t)$ and $\dot{x}(t)$ at $t = 0$ and τ. The lazier way is to use Mathematica. We find

$$x(t) = \frac{2[t\cos t(\sin\tau + \tau\cos\tau) - \sin t(-t\tau\sin\tau + \sin\tau + \tau\cos\tau)]}{2\tau^2 - 1 + \cos 2\tau}$$

$$u(t) = \frac{4[\tau\cos\tau\sin t + \sin\tau(\sin t - \tau\cos t)]}{2\tau^2 - 1 + \cos 2\tau}.$$

b. Here are the plots for x and u:



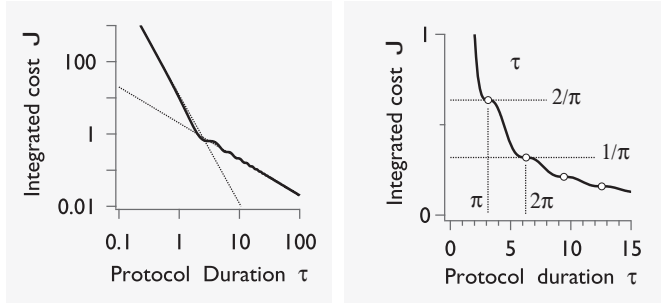
A few observations: For small τ, the movement is direct. For larger τ, the optimal movement (to reduce the cost J) is to use smaller movements and resonance, to gradually pump energy into the oscillator. Of course, unlike the pendulum example studied elsewhere, a non-zero $u = 1$ would be required to maintain the oscillator position at $x = 1$ for $t > \tau$. Another interesting

observation is that the energy, for $\tau < \pi$, goes above the value associated with the final state (0.5). For longer protocols, the energy monotonically increases, approaching a linear ramp for $\tau \to \infty$.

c. The integrated cost is given by

$$J(\tau) = \frac{2(2\tau + \sin 2\tau)}{2\tau^2 - 1 + \cos 2\tau} .$$

By taking limits for small and large τ, it is easy to find the advertised power laws. Note that the denominator for small τ is $\frac{2}{3}\tau^4 + O(\tau^6)$.



The log-log plot shows well the two asymptotic limits. The linear plot shows well how for $\tau \approx \pi$, the system has stationary points for $\tau_n = n\pi$, where $J'[\tau_n] = 0$. At those special values (half-integral numbers of the natural oscillation period), the cost is $J(\tau_n) = 2/(n\pi)$.

7.5 Pendulum swing up. Fill in the missing steps from Example 7.2. In particular:

a. Identify $\boldsymbol{x}$, $\boldsymbol{u}$, and $\boldsymbol{\lambda}$, along with the functions L and $\boldsymbol{f}$.
b. Compute the various derivative terms: $\partial_{\boldsymbol{x}} L$, $\partial_{\boldsymbol{u}} L$ and $\partial_{\boldsymbol{x}} \boldsymbol{f}$, $\partial_{\boldsymbol{u}} \boldsymbol{f}$.
c. Write the Euler-Lagrange equations as two sets of equations for the two-vectors $\boldsymbol{x}$ and $\boldsymbol{\lambda}$. Then rewrite as two coupled second-order equations for $\theta(t)$ and $\lambda(t)$.
d. Derive a single fourth-order, nonlinear differential equation for $\theta(t)$.
e. In Figure 7.2, why is $u(\tau) < 0$? Hint: relate the pendulum energy E to u.

Solution.

a. We have

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \theta \\ \dot{\theta} \end{pmatrix}, \qquad \boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} \lambda \\ \dot{\lambda} \end{pmatrix}, \qquad \boldsymbol{u} = u(t) ,$$

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}) = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = \begin{pmatrix} x_2 \\ -\sin x_1 + u \end{pmatrix}, \qquad L = \tfrac{1}{2} u^2(t) .$$

b. We have

$$\partial_{\boldsymbol{x}} L = \begin{pmatrix} 0 & 0 \end{pmatrix}, \quad \partial_u L = u , \qquad \partial_{\boldsymbol{x}} \boldsymbol{f} = \begin{pmatrix} 0 & 1 \\ -\cos x_1 & 0 \end{pmatrix}, \quad \partial_u \boldsymbol{f} = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

c. The "official" version of the equations of motion are given by applying the Euler-Lagrange equations, (7.8) to the above quantities. We have,

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_2 \\ -\sin x_1 + u \end{pmatrix},$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} \lambda_1 & \lambda_2 \end{pmatrix} = -\begin{pmatrix} \lambda_1 & \lambda_2 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -\cos x_1 & 0 \end{pmatrix} = \begin{pmatrix} \lambda_2 \cos x_1 & -\lambda_1 \end{pmatrix}.$$

The adjoint equation can also be written as its transpose,

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} \lambda_2 \cos x_1 \\ -\lambda_1 \end{pmatrix}.$$

and has boundary conditions at the final time $t = \tau$,

$$\begin{pmatrix} \lambda_1(\tau) \\ \lambda_2(\tau) \end{pmatrix} = \begin{pmatrix} (x_1(\tau) - \pi) \\ x_2(\tau) \end{pmatrix}.$$

Finally, the algebraic equation for $u(t)$ is

$$0 = \begin{pmatrix} \lambda_1 & \lambda_2 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} + u\,,$$

$$u(t) = -\lambda_2(t)\,.$$

Substituting $x_1 = \theta$, $x_2 = \dot{\theta}$, $\lambda_2 = \lambda$, $\lambda_1 = -\dot{\lambda}$ then leads directly to Eq. (7.11).

d. Differentiating twice $\ddot{\theta} + \sin\theta = u$ gives,

$$\theta^{(iv)} + \cos\theta\,\ddot{\theta} - \sin\theta\,\dot{\theta}^2 = \ddot{u}\,.$$

Substituting u into the adjoint equation for λ gives $\ddot{u} + u\cos\theta = 0$, so that

$$\theta^{(iv)} + \cos\theta\,\ddot{\theta} - \sin\theta\,\dot{\theta}^2 = -u\cos\theta\,,$$

and

$$\theta^{(iv)} + \cos\theta\,(2\ddot{\theta} + \sin\theta) - \sin\theta\,\dot{\theta}^2 = 0\,.$$

The boundary conditions at $t = 0$ are

$$\theta(0) = \dot{\theta}(0) = 0\,, \qquad \theta(\tau) = \pi\,, \dot{\theta}(\tau) = 0\,.$$

Thus, $\theta(t)$ obeys a single nonlinear fourth-order equation, with two boundary conditions at $t = 0$ and two at $t = \tau$. It is easy to verify that the numerical solutions of this single equation are equivalent to those found in the system of state-adjoint equations given in the first part of this problem.

e. In all the protocols explored, $u(t)$ is negative at the end. To understand why, let us consider the pendulum energy,

$$E = \tfrac{1}{2}\dot{\theta}^2 + (1 - \cos\theta)\,.$$

Differentiating with respect to time gives

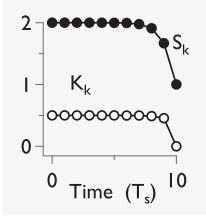
$$\dot{E} = \dot{\theta}\ddot{\theta} + \sin\theta\,\dot{\theta} = \dot{\theta}(\ddot{\theta} + \sin\theta) = \dot{\theta}\,u\,.$$

Assume that $\dot{\theta} > 0$ and that the pendulum is approaching the upright position $\theta = \pi$ in the counterclockwise direction. This is the situation shown in all three cases of Figure 7.2. Then $u > 0$ implies that $\dot{E} > 0$.

The second part of the argument is to note that the upright stationary position corresponds to $E = 2$ whereas the initial down state has $E = 0$. Thus, E needs to increase to swing up the pendulum. One possibility would be to set $u = 0$ once the pendulum energy has increased to $E = 2$ and let the pendulum "coast" to the top. But such a protocol would need an infinite amount of time for the pendulum to reach $\theta = \pi$ and $\dot{\theta} = 0$, whereas the boundary condition requires a finite time τ. Thus, the protocol needs to first give the pendulum $E > 2$ and then reverse to actively break as the pendulum approaches $\theta = \pi$. This requires $u < 0$ at the end.

We could also swing the pendulum up the other way, so that $\dot{\theta} < 0$ as it approaches π in a clockwise direction. This would reverse the signs of u in the above argument.

7.6 **Discrete, one-dimensional dynamics**. Consider applying a controllable force to a free, overdamped particle. The forward Euler method for the continuous dynamics $\dot{x} = u$ gives $x_{k+1} = x_k + T_s u_k$. Let the cost function be $J = \frac{1}{2}\sum_{k=0}^{N}\left(x_k^2 + Ru_k^2\right)$.



a. Form the one-dimensional augmented cost function J' that respects the constraint that x_k and u_k must obey the equations of motion.
b. Optimizing J' directly, show that $K_{N-1} = \frac{T_s}{R+T_s^2}$. Show, too, that this expression agrees with the result derived from general formula, Eq. (7.23).
c. Show that, in steady state, $S = \frac{1}{2} + \sqrt{\frac{1}{4} + \frac{R}{T_s^2}}$ and $K = \frac{ST_s}{R+ST_s^2}$.
d. For $T_s = 1$ and $R = 2$, plot S_k and K_k (see left).
e. For finite T_s, take the limit $R \to 0$. Why must $R > 0$ in the continuous case?

Solution.

a. The augmented cost function is formed by adding the constraint and its associated set of Lagrange multipliers. From Eq. (7.19), we have

$$J' = \frac{1}{2}\sum_{k=0}^{N}\left(x_k^2 + Ru_k^2\right) + \lambda_{k+1}\left(-x_{k+1} + x_k + T_s u_k\right) .$$

b.

$$x_{k+1} = x_k + T_s u_k\,, \quad \text{and} \quad J = \frac{1}{2}\sum_{k=0}^{N}\left(x_k^2 + Ru_k^2\right) ,$$

where we choose for cost function the discrete analog of Eq. (7.1). A direct way to solve the optimization problem is to note that if we start at time step $N-1$, then

$$J_{(N-1)\to N} = \frac{1}{2}\left[x_{N-1}^2 + Ru_{N-1}^2 + x_N^2\right] = \frac{1}{2}\left[x_{N-1}^2 + Ru_{N-1}^2 + (x_{N-1} + T_s u_{N-1})^2\right]$$

We set $u_N = 0$ because it affects only x_{N+1}, which is not part of J. Then we pick u_{N-1} to minimize $J_{(N-1)\to N}$:

$$\frac{\partial J_{(N-1)\to N}}{\partial u_{N-1}} = Ru_{N-1} + (x_{N-1} + T_s u_{N-1})(T_s) = 0\,,$$

and

$$u_{N-1} = -\frac{T_s}{R + T_s^2}\, x_{N-1}\,, \quad \Rightarrow \quad K_{N-1} = \frac{T_s}{R + T_s^2}\,. \tag{7.756}$$

c. The steady-state discrete algebraic Riccati equation is

$$\boldsymbol{S} = \boldsymbol{A}^{\mathsf{T}}\left[\boldsymbol{S} - \boldsymbol{S}\boldsymbol{B}(\boldsymbol{R} + \boldsymbol{B}^{\mathsf{T}}\boldsymbol{S}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{S}\right]\boldsymbol{A} + \boldsymbol{Q}\,,$$

which, for $A = Q = 1$ and $B = T_s$, gives

$$S = (1)\left(S - \frac{S^2 T_s^2}{R + S T_s^2}\right)(1) + 1 \quad \Longrightarrow \quad \frac{S^2 T_s^2}{R + S T_s^2} = 1\,,$$

which implies

$$S^2 - S - \frac{R}{T_s^2} = 0 \quad \Longrightarrow \quad S = \frac{1}{2} + \sqrt{\frac{1}{4} + \frac{R}{T_s^2}}\,.$$

Then,

$$\boldsymbol{K} = (\boldsymbol{R} + \boldsymbol{B}^{\mathsf{T}}\boldsymbol{S}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{S}\boldsymbol{A} \quad \Longrightarrow \quad \frac{T_s S}{R + T_s^2 S}$$

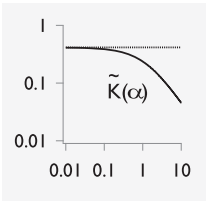
d. See computer program.

e. The limit $R \to 0$ gives $S = 1$ and $K = \frac{1}{T_s}$. Substituting into the equation of motion, we have

$$\begin{aligned} x_{k+1} &= x_k + T_s u_k\,, \qquad u_k = -\frac{1}{T_s} x_k \\ &= x_k - T_s\left(\frac{1}{T_s} x_k\right) = 0\,. \end{aligned}$$

This optimal controller is just the *deadbeat control* algorithm that we saw in Section 5.4.2. Notice that if $T_s \to 0$, then $K \to \infty$. Thus, from a mathematical point of view, we cannot set $R = 0$ for a continuous controller because the feedback gain becomes infinite. Physically, this result just restates the observation we made in Chapter 5: that deadbeat control really reflects a limitation of discrete control, in that no matter how much control effort is available, the sampling time T_s sets an upper limit to the gain. If you go beyond that gain, then you have oscillatory behavior (with a higher cost function). By contrast, in the continuous case, there is no mathematical limit to the maximum gain. Of course, an infinite feedback gain implies an infinite control signal, which is physically impossible.

7.7 **Discounted LQR**. The cost function $J = \frac{1}{2}\int_0^\infty \mathrm{d}t\, \mathrm{e}^{-2\alpha t}\left(\boldsymbol{x}^\mathsf{T}\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{u}^\mathsf{T}\boldsymbol{R}\boldsymbol{u}\right)$ is particularly popular in economics. The parameter $\alpha > 0$ *discounts*, or reduces the influence of future costs exponentially, on a time scale $(2\alpha)^{-1}$. This type of cost function has a steady-state solution, even though it is effectively a finite-horizon control problem.



a. By defining new variables $\tilde{\boldsymbol{x}} = \mathrm{e}^{-\alpha t}\boldsymbol{x}$ and $\tilde{\boldsymbol{u}} = \mathrm{e}^{-\alpha t}\boldsymbol{u}$, show that the problem reduces to solving a time-independent LQR problem with modified dynamics $\tilde{\boldsymbol{A}}$.
b. Show that the optimal control of the discounted problem has the form $\boldsymbol{u} = -\tilde{\boldsymbol{K}}\boldsymbol{x}$, and find $\tilde{\boldsymbol{K}}$ in terms of the solution to a steady-state Riccati equation.
c. Find $\tilde{K}(\alpha)$ for the one-dimensional problem of Section 7.1. You should get the plot at left, for $a = R = 1$. Intuitively, why does $\tilde{K}$ decrease with α?

Solution.

a. Defining $\tilde{\boldsymbol{x}} = \mathrm{e}^{-\alpha t}\boldsymbol{x}$ and $\tilde{\boldsymbol{u}} = \mathrm{e}^{-\alpha t}\boldsymbol{u}$, the new cost function is

$$\tilde{J} = \frac{1}{2}\int_0^\infty \mathrm{d}t\left(\tilde{\boldsymbol{x}}^\mathsf{T}\boldsymbol{Q}\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{u}}^\mathsf{T}\boldsymbol{R}\tilde{\boldsymbol{u}}\right),$$

which is time invariant in terms of the new states and controls. We need to work out the dynamics of the new state:

$$\begin{aligned}\dot{\tilde{\boldsymbol{x}}} &= -\alpha\, \mathrm{e}^{-\alpha t}\boldsymbol{x} + \mathrm{e}^{-\alpha t}\dot{\boldsymbol{x}}\\ &= -\alpha\tilde{\boldsymbol{x}} + \mathrm{e}^{-\alpha t}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u})\\ &= -\alpha\tilde{\boldsymbol{x}} + \boldsymbol{A}\tilde{\boldsymbol{x}} + \boldsymbol{B}\tilde{\boldsymbol{u}}\\ &= (\boldsymbol{A} - \alpha\boldsymbol{I})\,\tilde{\boldsymbol{x}} + \boldsymbol{B}\tilde{\boldsymbol{u}}\,.\end{aligned}$$

Thus, if we define $\tilde{\boldsymbol{A}} \equiv \boldsymbol{A} - \alpha\boldsymbol{I}$, we have completely reduced our problem to a stationary LQR problem in terms of $\tilde{\boldsymbol{x}}$, $\tilde{\boldsymbol{u}}$, and $\tilde{\boldsymbol{A}}$.

b. The solution to the standard LQR stationary problem is derived in the main text. In terms of our transformed variables, it is

$$\tilde{\boldsymbol{u}} = -\boldsymbol{R}^{-1}\boldsymbol{B}^\mathsf{T}\tilde{\boldsymbol{S}}\,\tilde{\boldsymbol{x}} \equiv -\tilde{\boldsymbol{K}}\tilde{\boldsymbol{x}}\,.$$

where

$$\dot{\tilde{\boldsymbol{S}}} = -\boldsymbol{Q} - \tilde{\boldsymbol{A}}^\mathsf{T}\tilde{\boldsymbol{S}} - \tilde{\boldsymbol{S}}\tilde{\boldsymbol{A}} + \tilde{\boldsymbol{S}}\boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^\mathsf{T}\tilde{\boldsymbol{S}}\,.$$

Now transform back to the original state and control variables:

$$\boldsymbol{u} = -\boldsymbol{R}^{-1}\boldsymbol{B}^\mathsf{T}\tilde{\boldsymbol{S}}\,\boldsymbol{x} \equiv -\tilde{\boldsymbol{K}}\boldsymbol{x}\,.$$

Thus, the solution is time-independent, with $\tilde{\boldsymbol{K}}$ expressed in terms of $\tilde{\boldsymbol{S}}$, which differs from the original problem in using the modified dynamics $\tilde{\boldsymbol{A}}$. Notice, of course, that when $\alpha \to 0$, we recover the original LQR steady-state solution.

c. Let us apply the general solution to the one-dimensional problem,

$$J = \tfrac{1}{2}\int_0^\infty \mathrm{d}t\, \mathrm{e}^{-2\alpha t}\left(x^2(t) + Ru^2(t)\right),$$
$$\dot{x} = -ax + u(t), \qquad x(0) = x_0, \qquad a > 0.$$

The transformed dynamics have $a \to \tilde{a} = (a + \alpha)$, so that the Riccati equation becomes

$$\dot{\tilde{S}} = 0 = -1 + 2\tilde{a}\tilde{S} + \frac{\tilde{S}^2}{R},$$

with solution

$$\tilde{S} = -\tilde{a}R \pm \sqrt{\tilde{a}^2R^2 + R}$$

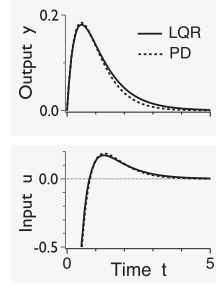
and feedback gain

$$\tilde{K}(\alpha) = R^{-1}(1)\tilde{S} = -(a + \alpha) + \sqrt{(a + \alpha)^2 + 1/R}.$$

The optimal gain $\tilde{K}$ decreases with α because the discounting implies dynamics equivalent to a stationary problem that is more stable than the original dynamics (increase of α in the decay rate). Thus, less feedback is needed to stabilize the equivalent dynamics. In the numerical example with $a = R = 1$, the optimal gain of the undiscounted problem is $\tilde{K}(0) = \sqrt{2} - 1 \approx 0.41$.

If the initial system were unstable, the reduced gain might not be enough to stabilize the closed-loop dynamics. This makes sense: if you truly care only about the finite-time behavior [out to a time $(2\alpha)^{-1}$], then you can accept a possibly unstable solution. Still, you probably would want to limit your value of α so that such instability does not occur.

7.8 **Optimal control of an undamped harmonic oscillator**. In Chapter 4, we studied the PD strategy for regulating a harmonic oscillator against input disturbances. In Example 4.10, we found that for $G = \frac{1}{1+s^2}$ that the PD controller $K = k_1 + k_2 s$ gave good results for $k_1 = 3$ and $k_2 = 4$. Here, design a similar controller using optimal control. Fix the weights $\boldsymbol{Q}$ to be the 2×2 identity matrix and vary R.



a. Find, numerically or algebraically the "LQR" gains $k_1(R)$ and $k_2(R)$.
b. By plotting the system output $y(t)$ and controlled input $u(t)$, show that $R \approx 0.08$ gives a response similar to the PD controller. (Plot should resemble one at right.)
c. Show that the only value of R giving critical damping is $R = 1/8$.
d. Why is the LQR controller, in general, not critically damped?

Solution.

See the Mathematica file on the book website.

a. To solve the LQR gains for $\boldsymbol{K} = (k_1\, k_2)$, we can use built-in routines in control software such as Mathematica or Matlab. These programs will, however, typically give only numerical solutions, for a specified value of R. To

solve symbolically, we use Mathematica to derive (and solve explicitly) the algebraic Riccati equation,

$$\boldsymbol{Q} = -\boldsymbol{A}^{\mathsf{T}}\boldsymbol{S} - \boldsymbol{S}\boldsymbol{A} + \boldsymbol{S}\boldsymbol{B}R^{-1}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{S}\,,$$

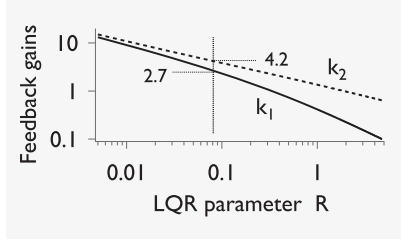
for the 2×2 matrix $\boldsymbol{S}$. Here,

$$\boldsymbol{A} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \boldsymbol{Q} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{and } R = (1)\,.$$

The solution for $\boldsymbol{S}$ is complicated. As discussed in the text, there are multiple solutions, but only for one is $\boldsymbol{S}$ positive definite. Selecting that one and calculating $\boldsymbol{K} = R^{-1}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{S}$, we find

$$\boldsymbol{K} = \begin{pmatrix} k_1 & k_2 \end{pmatrix}, \qquad k_1 = -1 + \sqrt{1 + \frac{1}{R}}\,, \quad k_2 = \frac{\sqrt{R - 2R^2 + 2\sqrt{R^3(1+R)}}}{R}\,..$$

These k-dependent gains are plotted here below.



b. At $R = 0.08$, we have $k_1 \approx 2.7$ and $k_2 \approx 4.2$, which are close to, but different from the (3,4) solution used for the PD controller giving critical damping (pole at $s = -2$).

c. To have critical damping of the closed-loop response, the transfer function should be of the form $T(s) = \frac{1}{(s+a)^2} = \frac{1}{s^2+2as+a^2}$. On the other hand, a controller $\boldsymbol{K} = (k_1\, k_2)$ leads to a closed-loop transfer function with denominator

$$s^2 + k_2 s + (1 + k_1) \leftrightarrow s^2 + 2as + a^2\,.$$

Matching terms leads to $k_2^2/4 = 1 + k_1$. Inserting the expressions for R given above leads to a single solution, $R = 1/8$, which corresponds to a decay rate $a = \sqrt{3}$.

d. Why is the LQR controller not critically damped (except at $R = 1/8$)? Well, why should it be? The point is that critical damping for a closed-loop response can come from a cost function that seeks to minimize the decay time of the response-amplitude envelope. The LQR controller discussed here minimizes a different cost. Two different cost functions in general will lead to two different controllers. As discussed in several places in this chapter, there is nothing inherently "good" about optimal control. Rather, you replace a direct design of the controller with an indirect one that translates the cost function into a controller.

7.9 **Minimum-effort control**. Consider our canonical SISO linear system, $\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}u$. Assume that we want to move the state $\boldsymbol{x}(0) = \boldsymbol{0}$ to $\boldsymbol{x}(\tau) = \boldsymbol{x}_\tau$. We do not care what the intermediate path $\boldsymbol{x}(t)$ is; rather, our goal is to choose the function $u(t)$ so as to minimize the required *control effort* $\mathcal{E} \equiv \int_0^\tau \mathrm{d}t\, u^2(t)$ (often an energy-like quantity).

a. Show that the optimal input is given by $u = \boldsymbol{B}^\mathsf{T} \mathrm{e}^{\boldsymbol{A}^\mathsf{T}(\tau-t)}\ \boldsymbol{P}^{-1}(\tau)\,\boldsymbol{x}_\tau$, where the $n \times n$ dimensional *controllability Gramian* matrix is $\boldsymbol{P}(\tau) = \int_0^\tau \mathrm{d}t\, \mathrm{e}^{\boldsymbol{A}t}\,\boldsymbol{B}\,\boldsymbol{B}^\mathsf{T}\,\mathrm{e}^{\boldsymbol{A}^\mathsf{T}t}$. Hint: This problem is just LQR with a boundary condition on $\boldsymbol{x}(\tau)$.
b. Deduce the normalized minimum control effort, $\mathcal{E}_\mathrm{n} = \frac{\boldsymbol{x}_\tau^\mathsf{T}\boldsymbol{P}(\tau)^{-1}\boldsymbol{x}_\tau}{\boldsymbol{x}_\tau^\mathsf{T}\boldsymbol{x}_\tau} = \hat{\boldsymbol{n}}^\mathsf{T}\,\boldsymbol{P}(\tau)^{-1}\,\hat{\boldsymbol{n}}$, assuming that the target $\boldsymbol{x}_\tau = \hat{\boldsymbol{n}}$ is a vector on the unit n-sphere.
c. Show that if $\boldsymbol{x}(0) = \boldsymbol{x}_0 \neq \boldsymbol{0}$, then $\boldsymbol{x}_\tau \to \Delta\boldsymbol{x} \equiv \boldsymbol{x}_\tau - \mathrm{e}^{\boldsymbol{A}\tau}\,\boldsymbol{x}_0$. Please interpret.
d. For $\dot{x} = -x + u$, with $x(0) = 0$ and $x(\tau) = x_\tau$, find the input $u(t)$ with minimum effort and the corresponding state $x(t)$. Show that the corresponding normalized minimum effort is $\mathcal{E}_\mathrm{n} = 2(1 - \mathrm{e}^{-2\tau})^{-1}$. Physically, the system corresponds to moving a Brownian particle in a harmonic potential, and we ask for the minimum effort to push a particle "up the potential." Discuss the limits $\tau \to \infty$ and $\tau \to 0$.

Solution.

a. In the general solution, we have $\boldsymbol{Q} = \boldsymbol{0}$ and $\boldsymbol{R} = 1$. Also, we have to impose the boundary condition on $\boldsymbol{x}$ at *both* $t = 0$ and $t = \tau$. As a result, we do not impose a boundary condition on $\boldsymbol{\lambda}$ at $t = \tau$. Thus, from Eq. (7.13), we have

$$\dot{\boldsymbol{\lambda}} = -\boldsymbol{A}^\mathsf{T}\boldsymbol{\lambda}\,, \qquad \Longrightarrow \qquad \boldsymbol{\lambda}(t) = \mathrm{e}^{\boldsymbol{A}^\mathsf{T}(\tau-t)}\,\boldsymbol{\lambda}_\tau\,,$$

where the integration constant is such that $\boldsymbol{\lambda}(\tau) \equiv \boldsymbol{\lambda}_\tau$. Then,

$$\boldsymbol{u} = -\boldsymbol{B}^\mathsf{T}\boldsymbol{\lambda} = -\boldsymbol{B}^\mathsf{T}\,\mathrm{e}^{\boldsymbol{A}^\mathsf{T}(\tau-t)}\,\boldsymbol{\lambda}_\tau\,.$$

Inserting into the solution for $\boldsymbol{x}(t)$ gives

$$\begin{aligned}\boldsymbol{x}_\tau &= \int_0^\tau \mathrm{d}t\, \mathrm{e}^{\boldsymbol{A}(\tau-t)}\,\boldsymbol{B}\,u(t)\\ &= -\int_0^\tau \mathrm{d}t\, \mathrm{e}^{\boldsymbol{A}(\tau-t)}\,\boldsymbol{B}\,\boldsymbol{B}^\mathsf{T}\,\mathrm{e}^{\boldsymbol{A}^\mathsf{T}(\tau-t)}\,\boldsymbol{\lambda}_\tau \equiv -\boldsymbol{P}(\tau)\boldsymbol{\lambda}_\tau\,,\end{aligned}$$

where we note that

$$\int_0^\tau \mathrm{d}t\, \mathrm{e}^{\boldsymbol{A}(\tau-t)}\,\boldsymbol{B}\,\boldsymbol{B}^\mathsf{T}\,\mathrm{e}^{\boldsymbol{A}^\mathsf{T}(\tau-t)} = \int_0^\tau \mathrm{d}s\, \mathrm{e}^{\boldsymbol{A}s}\,\boldsymbol{B}\,\boldsymbol{B}^\mathsf{T}\,\mathrm{e}^{\boldsymbol{A}^\mathsf{T}s} = \boldsymbol{P}(\tau)\,,$$

as can be seen by substituting $s = \tau - t$. Thus, $\boldsymbol{\lambda}_\tau = -\boldsymbol{P}^{-1}(\tau)\boldsymbol{x}_\tau$, and

$$u = \boldsymbol{B}^\mathsf{T}\,\mathrm{e}^{\boldsymbol{A}^\mathsf{T}(\tau-t)}\ \boldsymbol{P}^{-1}(\tau)\,\boldsymbol{x}_\tau\,,$$

as claimed.

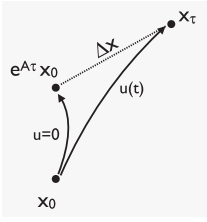
b. Since $\boldsymbol{P}$ is symmetric by construction, we can write the expression for the control effort as

$$\begin{aligned}\mathcal{E} &= \int_0^\tau \mathrm{d}t\, u^2(t) \\ &= \boldsymbol{x}_\tau^\mathsf{T} \boldsymbol{P}^{-1}(\tau) \underbrace{\int_0^\tau \mathrm{d}t\, \mathrm{e}^{\boldsymbol{A}(\tau-t)}\, \boldsymbol{B}\, \boldsymbol{B}^\mathsf{T}\, \mathrm{e}^{\boldsymbol{A}^\mathsf{T}(\tau-t)}}_{\boldsymbol{P}(\tau)}\, \boldsymbol{P}^{-1}(\tau)\, \boldsymbol{x}_\tau \\ &= \boldsymbol{x}_\tau^\mathsf{T}\, \boldsymbol{P}^{-1}(\tau)\, \boldsymbol{x}_\tau\,.\end{aligned}$$

Since the minimum control effort clearly scales with the initial condition $|\boldsymbol{x}_\tau|^2$, it often makes sense to normalize by this factor and define $\mathcal{E}_\mathrm{n} = \mathcal{E}/|\boldsymbol{x}_\tau|^2$.

c. For the case where $\boldsymbol{x}(0) = \boldsymbol{x}_0 \neq \boldsymbol{0}$, we recall that the solution for $\boldsymbol{x}_\tau$ that takes account for the transient produced by the initial condition:

$$\boldsymbol{x}_\tau = \mathrm{e}^{\boldsymbol{A}\tau}\, \boldsymbol{x}_0 + \int_0^\tau \mathrm{d}t\, \mathrm{e}^{\boldsymbol{A}(\tau-t)}\, \boldsymbol{B}\, u(t)\,.$$



Rewrite this equation as

$$\boldsymbol{\Delta x} \equiv \boldsymbol{x}_\tau - \mathrm{e}^{\boldsymbol{A}\tau}\, \boldsymbol{x}_0 = \int_0^\tau \mathrm{d}t\, \mathrm{e}^{\boldsymbol{A}(\tau-t)}\, \boldsymbol{B}\, u(t)\,,$$

with $\boldsymbol{\Delta x}$ the difference between the final state under control, $\boldsymbol{x}_\tau$, and the final state of the system in the absence of control, $\mathrm{e}^{\boldsymbol{A}\tau}\, \boldsymbol{x}_0$, as illustrated at left. Then $\boldsymbol{\lambda}_\tau = -\boldsymbol{P}^{-1}(\tau)\boldsymbol{\Delta x}$, where

$$\boldsymbol{\Delta x} \equiv \boldsymbol{x}_\tau - \mathrm{e}^{\boldsymbol{A}\tau}\, \boldsymbol{x}_0\,,$$

and the optimal control is

$$u = \boldsymbol{B}^\mathsf{T}\, \mathrm{e}^{\boldsymbol{A}^\mathsf{T}(\tau-t)}\, \boldsymbol{P}^{-1}(\tau)\, \boldsymbol{\Delta x}\,, \qquad \mathcal{E} = \boldsymbol{\Delta x}^\mathsf{T}\, \boldsymbol{P}^{-1}(\tau)\, \boldsymbol{\Delta x}\,,$$

which makes sense: it is moving the system from its "natural," uncontrolled state that requires control effort. Similarly, the normalized control effort is

$$\mathcal{E}_\mathrm{n} = \frac{\boldsymbol{\Delta x}^\mathsf{T}\, \boldsymbol{P}^{-1}(\tau)\, \boldsymbol{\Delta x}}{\boldsymbol{\Delta x}^\mathsf{T}\, \boldsymbol{\Delta x}}\,.$$

d. With $\boldsymbol{A} = -1$ and $\boldsymbol{B} = 1$, we have

$$P(\tau) = \int_0^\tau \mathrm{d}t\, \mathrm{e}^{-t}(1)(1)\, \mathrm{e}^{-t} = \int_0^\tau \mathrm{d}t\, \mathrm{e}^{-2t} = \tfrac{1}{2}\left(1 - \mathrm{e}^{-2\tau}\right)\,.$$

Then,

$$u(t) = (1)\, \mathrm{e}^{-(\tau-t)} \left(\frac{2}{1-\mathrm{e}^{-2\tau}}\right) x_\tau = \frac{2\,\mathrm{e}^{-\tau}}{1-\mathrm{e}^{-2\tau}}\, x_\tau\, \mathrm{e}^{t} = x_\tau \left(\frac{2}{\mathrm{e}^{\tau}-\mathrm{e}^{-\tau}}\right) \mathrm{e}^{t}\,.$$

We find the state $x(t)$ for $0 < t < \tau$ by

$$\begin{aligned} x(t) &= \int_0^t \mathrm{d}t' \, \mathrm{e}^{-(t-t')}(1) \underbrace{x_\tau \left(\frac{2}{\mathrm{e}^\tau - \mathrm{e}^{-\tau}} \right) \mathrm{e}^{t'}}_{u(t')} \\ &= x_\tau \left(\frac{2}{\mathrm{e}^\tau - \mathrm{e}^{-\tau}} \right) \mathrm{e}^{-t} \int_0^t \mathrm{d}t' \, \mathrm{e}^{2t'} \\ &= x_\tau \left(\frac{\cancel{2}}{\mathrm{e}^\tau - \mathrm{e}^{-\tau}} \right) \mathrm{e}^{-t} \left(\frac{\mathrm{e}^{2t} - 1}{\cancel{2}} \right) \\ &= x_\tau \left(\frac{\mathrm{e}^t - \mathrm{e}^{-t}}{\mathrm{e}^\tau - \mathrm{e}^{-\tau}} \right) . \end{aligned}$$

The normalized control effort is then

$$\begin{aligned} \mathcal{E}_\mathrm{n} &= \frac{1}{x_\tau^2} \int_0^\tau \mathrm{d}t \, u^2(t) = \frac{1}{x_\tau^2} \left[\frac{4}{(\mathrm{e}^\tau - \mathrm{e}^{-\tau})^2} \right] x_\tau^2 \int_0^\tau \mathrm{d}t \, (\mathrm{e}^t)^2 \\ &= \frac{2\left(\mathrm{e}^{2\tau} - 1\right)}{(\mathrm{e}^\tau - \mathrm{e}^{-\tau})^2} \\ &= \frac{2}{1 - \mathrm{e}^{-2\tau}} . \end{aligned}$$

In the limit $\tau \gg 1$, we have $\mathcal{E}_\mathrm{n} \to 2$. For $\tau \ll 1$, we have $\mathcal{E}_\mathrm{n} \to \tau^{-1}$. The divergence in τ is expected: moving quickly requires a lot of control effort.

7.10 Minimum-energy control. In Problem 7.9, we discussed the minimum-effort control; however, the relation of "effort" to thermodynamic work is not entirely obvious. In this problem, we explore operations that minimize the heat dissipated into the surrounding fluid bath. Consider an overdamped particle in a harmonic potential, with equations of motion $\dot{x} = -x + u$, with $x(0) = 0$ and $x(\tau) = x_\tau$. For simplicity, ignore thermal fluctuations, which would add a stochastic term.

a. Define the heat dissipated into the bath as $Q = -\int_0^\tau \mathrm{d}t \, \dot{x}^2$, where the negative sign denotes that energy is lost by the particle and reappears as heat in the bath. Generalize the cost function in Eq. (7.12) to allow for a cross term proportional to $u(t)\,x(t)$; deduce the trajectory $x(t)$ and control $u(t)$ that minimizes Q.
b. Calculate Q_min for this minimum-dissipation trajectory. Discuss the limits of large and small τ (assuming x_τ to be fixed and finite).
c. The work done on the particle is $W = \int_0^\tau \mathrm{d}t \, u(t)\,\dot{x}(t)$, where $u(t)$ is interpreted as the applied "force." Calculate W_min, and interpret for large and small τ.
d. Verify that the first law of thermodynamics holds, in the form of $\Delta U = W + Q$.

Solution.

a. We quickly repeat the optimal-control derivation. The absolute value of the heat functional $Q[u(t)]$ is given by

$$|Q| = + \int_0^\tau \mathrm{d}t\, \dot{x}^2 = \int_0^\tau \mathrm{d}t\, (-x+u)^2 = \int_0^\tau \mathrm{d}t \left(x^2 - 2xu + u^2\right) .$$

We then define an "augmented cost function" that adds the equation of motion as a constraint:

$$J = \tfrac{1}{2} \int_0^\tau \mathrm{d}t \left(x^2 - 2xu + u^2\right) + \lambda(-\dot{x} - x + u)\, dt \equiv \int_0^\tau \mathrm{d}t\, L(t) ,$$

where the factor of $\frac{1}{2}$ amounts to a redefinition of λ that is done to conform to the notation in the text. We then apply the variational equations.

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{x}} = -\dot{\lambda} = \frac{\partial L}{\partial x} = x - u + \lambda \quad \Longrightarrow \quad \dot{\lambda} = \lambda - x + u$$
$$\frac{\partial L}{\partial u} = -x + u + \lambda = 0 \quad \Longrightarrow \quad \lambda = x - u .$$

Putting these equations together then gives $\dot{\lambda} = 0$, or $\lambda = \lambda_0$. From the equation of motion, $\dot{x} = -x + u = -\lambda_0$, so that $x(t) = -\lambda_0 t$. Imposing the boundary condition $x(\tau) = x_\tau$ then gives

$$x(t) = \frac{x_\tau}{\tau} t , \qquad \dot{x} = \frac{x_\tau}{\tau} , \qquad u(t) = \dot{x} + x = \frac{x_\tau}{\tau}(1+t) .$$

b. The minimum heat dissipation is

$$Q_{\min} = -\int_0^\tau \mathrm{d}t\, \dot{x}^2 = \left(\frac{x_\tau}{\tau}\right)^2 \tau = -\frac{x_\tau^2}{\tau} .$$

We note that $Q_{\min} \to 0$ as $\tau \to \infty$. If the movement is done adiabatically (infinitely slowly), the dissipated heat goes to 0. For $\tau \to 0$, the heat dissipation diverges, as expected.

c. The minimum work done is

$$W_{\min} = \int_0^\tau \mathrm{d}t\, ux = \left(\frac{x_\tau}{\tau}\right)^2 \int_0^\tau \mathrm{d}t\, (1+t)t , = \left(\frac{x_\tau}{\tau}\right)^2 \left(\tau + \tfrac{1}{2}\tau^2\right) = \tfrac{1}{2}x_\tau^2 + \frac{x_\tau^2}{\tau} .$$

For $\tau \to \infty$, we have $W_{\min} = \frac{1}{2}x_\tau^2$, which is just the change in potential energy. For $\tau \to 0$, the work diverges.

d. We verify that $W + Q = \frac{1}{2}x_\tau^2 = \Delta U$.

7.11 Soft end-time constraints. Consider one-dimensional motion of a Newtonian particle with $\ddot{x} = u$ and $x(0) = \dot{x}(0) = 0$. The goal is to move the particle close to $x(\tau) = x_\tau$, while minimizing the fuel cost. We impose no constraint on the velocity at $t = \tau$, and the dynamics-constrained cost functional is given by Eq. (7.6). The soft end-time constraint leads to the usual Euler–Lagrange equations (7.8), which hold for $t \in (0, \tau]$; however, the boundary condition at time τ becomes $\partial_{\boldsymbol{x}}\, \varphi + \partial_{\dot{\boldsymbol{x}}}\, L = 0$. If $\dot{\boldsymbol{x}}$ enters only in the constraint on the dynamical equation, the

boundary condition becomes $\boldsymbol{\lambda}^{\mathsf{T}}(\tau) = \partial_{\boldsymbol{x}}\varphi(\tau)$. We choose $L = \frac{1}{2}u^2$ (no x dependence) and $\varphi = \frac{1}{2}S(\bar{x}-1)^2$. Scaling x by x_τ, t by τ, and defining $\bar{x} = x(\tau)$, we have,

$$J' = \tfrac{1}{2}S(\bar{x}-1)^2 + \tfrac{1}{2}\int_0^1 \mathrm{d}t\, u^2(t) + \int_0^1 \mathrm{d}t\, \boldsymbol{\lambda}^{\mathsf{T}}(-\dot{\boldsymbol{x}} + \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u})\,.$$

where $(\bar{x}-1) = x(\tau) - x_\tau$ in unscaled units. The parameter S balances accuracy of the end state against control effort ("fuel consumption").

a. Write the equations of motion in the standard form $\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}$.
b. Write the adjoint equations for $\boldsymbol{\lambda}(t)$, with boundary conditions at $t = \{0, 1\}$.
c. Find $x_1(t)$, $x_2(t)$, $\lambda_1(t)$, $\lambda_2(t)$, and $u(t)$, as well as a relation between $\bar{x}$ and S and an expression for $J^*(S)$. Show that $u(t) = \frac{3S}{S+3}(1-t)$.
d. Discuss the limits $S \to \infty$ and $S \to 0$.
e. For $S = 1$ and $S = \infty$, plot u, x, and $\dot{x}$ over the interval $t = (0, 1)$.

Solution.

a. The equations of motion are $x_1 = x$, $\dot{x}_1 = x_2$, $\dot{x}_2 = u$, which, in matrix form, give

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}}_{\boldsymbol{A}} \underbrace{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}}_{\boldsymbol{x}} + \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{\boldsymbol{B}} u\,.$$

b. The Euler-Lagrange equations are

$$\dot{\boldsymbol{\lambda}} = -\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{A}^{\mathsf{T}}\boldsymbol{\lambda}\,,$$

with $\boldsymbol{Q} = \boldsymbol{0}$, giving

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = -\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \qquad \Longrightarrow \qquad \dot{\lambda}_1 = 0\,, \quad \dot{\lambda}_2 = -\lambda_1\,.$$

The second equation, from minimizing with respect to u is $\boldsymbol{u} = -\boldsymbol{R}^{-1}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{\lambda}$. With $R = 1$, we have

$$u = -(1)^{-1}\begin{pmatrix} 0 & 1 \end{pmatrix}\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = -\lambda_2\,.$$

The boundary condition at $t = 0$ is on the state equation: $\boldsymbol{x}(t=0) = \boldsymbol{x}_0 = \binom{0}{0}$, since the particle is specified as being at the origin ($x_1 = 0$) and stationary ($x_2 = 0$) at $t = 0$.
The boundary condition at $t = 1$ is, with $\varphi(x) = \frac{1}{2}S(\bar{x}-1)^2$ and $\bar{x} = x_1(t=1)$. Taking the transpose (so that we write column vectors, instead of row vectors), we have

$$\begin{pmatrix} \lambda_1(1) \\ \lambda_2(1) \end{pmatrix} = \begin{pmatrix} S(\bar{x}-1) \\ 0 \end{pmatrix}.$$

c. Collecting the equations and boundary conditions, we have five equations for the five unknown quantities x_1, x_2, λ_1, λ_2, and u. The equations and boundary conditions are

$$\begin{aligned} \dot{x_1} &= x_2 & x_1(0) &= 0 \\ \dot{x_2} &= u & x_2(0) &= 0 \\ \dot{\lambda_1} &= 0 & \lambda_1(1) &= S(\bar{x} - 1) \\ \dot{\lambda_2} &= -\lambda_1 & \lambda_2(1) &= 0 \\ u &= -\lambda_2 \,. \end{aligned}$$

To solve this system of equations, we start by noting

$$\lambda_1(t) = S(\bar{x} - 1) \,.$$

Then

$$\begin{aligned} \lambda_2(t) &= S(\bar{x} - 1)(1 - t) \,, \\ u(t) &= S(1 - \bar{x})(1 - t) \,, \end{aligned}$$

Then, we can determine the state equations:

$$\begin{aligned} x_2(t) &= S(1 - \bar{x})\left(t - \tfrac{1}{2}t^2\right) \\ x_1(t) &= S(1 - \bar{x})\left(\tfrac{1}{2}t^2 - \tfrac{1}{6}t^3\right) \,. \end{aligned}$$

We still need to determine $\bar{x}$. Imposing $x_1(t = 1) = \bar{x}$, we have

$$\bar{x} = \frac{S}{S + 3} \,.$$

The final solution then is

$$x_1(t) = \left(\frac{3S}{S + 3}\right)\left(\tfrac{1}{2}t^2 - \tfrac{1}{6}t^3\right) \quad x_2(t) = \left(\frac{3S}{S + 3}\right)\left(t - \tfrac{1}{2}t^2\right) \qquad \bar{x} = \frac{S}{S + 3}$$

$$\lambda_1(t) = -\left(\frac{3S}{S + 3}\right) \qquad \lambda_2(t) = -\left(\frac{3S}{S + 3}\right)(1 - t) \quad u(t) = \left(\frac{3S}{S + 3}\right)(1 - t) \,.$$

The optimal cost is

$$J^* = \frac{1}{2}\left[S\left(\frac{9}{(S + 3)^2}\right) + \frac{9S^2}{(S + 3)^2}\underbrace{\int_0^1 \mathrm{d}t\,(1 - t)^2}_{1/3}\right] = \frac{1}{2}\left(\frac{3S}{S + 3}\right) \,.$$

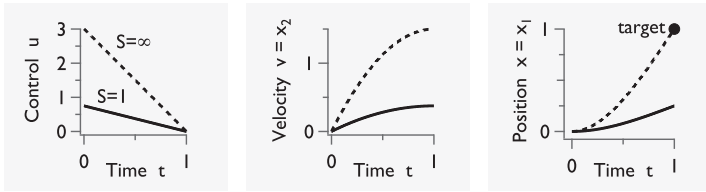
d. The key quantity is $\bar{x}$. We have

$$\begin{aligned} S &= 0 & \bar{x} &= 0 & J^* &= 0 \\ S &= 1 & \bar{x} &= \frac{3}{4} & J^* &= \frac{3}{8} \\ S &\to \infty & \bar{x} &\to 1 & J^* &= \frac{3}{2} \,. \end{aligned}$$

The limits are easy to understand. For large S, the solution tends to the target value. For small S, the best thing is to stay put and not do anything. For finite S, the best thing is to always fall short relative to the target ($\bar{x} < 1$) by a precise amount.

In the $S \to \infty$ limit, J^* does *not* go to ∞. All we are doing in this limit is imposing the constraint that we *must* hit the target. While this requires more fuel than falling short, it does not require an infinite amount, and thus, J^* is finite.

e. For the control, velocity, and position, we have



7.12 Sequential optimization and the Bellman equation. The "magic" of the Bellman equation arises because the optimization over N variables has a special "sequentially coupled" form. Consider optimizing $L(x_1, x_2, x_3) = f_0(x_0, x_1) + f_1(x_1, x_2) + f_2(x_2, x_3)$, where x_0 is given (the "initial condition" for a dynamical problem). Although one could solve the three coupled equations $\partial_{x_1} L = \partial_{x_2} L = \partial_{x_3} L = 0$, the special "intertwined" structure of the problem suggests a simpler, sequential solution.

a. Solve the equations sequentially, starting from $\partial_{x_3} f_2 = 0$.

b. For $f_0 = \frac{1}{2}(x_0 - x_1)^2 - x_1$, $f_1 = \frac{1}{2}(x_1 - x_2)^2 - x_2$, $f_2 = \frac{1}{2}(x_2 - x_3)^2 - x_3$, find the minimizing set $\{x_1^*, x_2^*, x_3^*\}$ by naive "global" optimization and by the easier "sequential" optimization. [See Rawlings et al. (2017), Section 1.3.2.]

Solution.

a. We first minimize $f_2(x_2, x_3)$ over x_3:

$$x_3^{**} = \arg\min_{x_3} \frac{\partial f_2}{\partial x_3} = 0\,, \qquad \text{and } f_2^*(x_2) \equiv f_2(x_2, x_3^{**})\,.$$

We then minimize $f_1(x_1, x_2) + f_2^*(x_2)$ over x_2:

$$x_2^{**} = \arg\min_{x_2} \frac{\partial (f_1 + f_2^*)}{\partial x_2} = 0\,, \qquad \text{and } f_1^*(x_1) \equiv f_1(x_1, x_2^{**}) + f_2^*(x_2^{**})\,.$$

Last, we minimize $f_0(x_0, x_1) + f_1^*(x_1)$ over x_1:

$$x_1^{**} = \arg\min_{x_1} \frac{\partial (f_0 + f_1^*)}{\partial x_1} = 0\,,$$

We have converted the solution of a set of three equations in three unknowns into three sequential optimization problems, each involving a single variable. With such a structure, we can clearly extend to N time steps, as we do in

the main text in the Bellman equation. We use the x_2^{**} notation because this intermediate minimizer is a function of x_1.

As an aside, the kind of simplification we see here is analogous to the distinction between general matrix inversion and the much-easier tridiagonal matrix inversion of linear algebra.

The next step is to generate explicit minimizers x_1^*, x_2^*, x_3^* by forward iteration:

$$\begin{aligned} x_1^* &= x_1^{**} \\ x_2^* &= x_2^{**}(x_1^*) \\ x_3^* &= x_3^{**}(x_2^*)\,. \end{aligned}$$

The first equation will be trivial because x_1^{**} only depends on x_0, which is considered known. Then we can propagate forward, substituting the optimal values as we go along. This will be clearer in the next section, in an explicit example.
Lastly, we can write the value of the overall function $L = f_0 + f_1 + f_2$:

$$L = f_0(x_0, x_1^*) + f_1(x_1^*, x_2^*) + f_2(x_2^*, x_3^*)\,.$$

b. We consider the example

$$L(x_1, x_2, x_3) = \underbrace{\tfrac{1}{2}(x_0 - x_1)^2 - x_1}_{f_0} + \underbrace{\tfrac{1}{2}(x_1 - x_2)^2 - x_2}_{f_1} + \underbrace{\tfrac{1}{2}(x_2 - x_3)^2 - x_3}_{f_2}\,.$$

The global optimization leads to a set of three coupled equations,

$$\frac{\partial L}{\partial x_1} = \frac{\partial L}{\partial x_2} = \frac{\partial L}{\partial x_3} = 0\,,$$

which leads to the matrix equations

$$\underbrace{\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}}_{\boldsymbol{A}} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_0 + 1 \\ 1 \\ 1 \end{pmatrix},$$

which implies that

$$\begin{pmatrix} x_1^* \\ x_2^* \\ x_3^* \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix}}_{\boldsymbol{A}^{-1}} \begin{pmatrix} x_0 + 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} x_0 + 3 \\ x_0 + 5 \\ x_0 + 6 \end{pmatrix}.$$

We can do this problem as a sequence of single-variable optimizations if we take advantage of its intertwined structure of variables. We first solve

$$\begin{aligned} x_3^{**} &= \arg\min_{x_3} \frac{\partial f_2}{\partial x_3} = 0 \\ &= x_2 + 1 \quad \Longrightarrow \quad f_2^*(x_2) = f_2(x_2, x_3^{**}) = -x_2 - \tfrac{1}{2}\,. \end{aligned}$$

We then solve

$$x_2^{**} = \arg\min_{x_2} \frac{\partial(f_1 + f_2^*)}{\partial x_2} = 0$$
$$= x_1 + 2\,,$$

which implies that

$$f_1^*(x_1) = f_1(x_1, x_2^{**}) + f_2^*(x_2^{**}) = -2x_1 - \tfrac{5}{2}\,.$$

For the last stage, we solve

$$x_1^{**} = \arg\min_{x_1} \frac{\partial(f_0 + f_1^*)}{\partial x_1} = 0$$
$$= x_0 + 3 \quad \Longrightarrow \quad f_0^*(x_0) = f_0(x_0, x_1^{**}) + f_1^*(x_1^{**}) = -3x_0 - 7\,.$$

We note that $f_0^*(x_0) = L(x_1^*, x_2^*, x_3^*)$.

Notice that we now have a *forward* recursion relation for the x_i^*:

$$x_1^* = x_1^{**}(x_0) = x_0 + 3$$
$$x_2^* = x_2^{**}(x_1^*) = x_1^* + 2 = x_0 + 5$$
$$x_3^* = x_3^{**}(x_2^*) = x_2^* + 1 = x_0 + 6\,,$$

which reproduces the global solution found above. In the above discussion, we have distinguished between x_2^{**}, the optimal value for an arbitrary x_1 from x_2^*, the optimal value given the optimal x_1. We start at x_1 because, different from the other equations, we assume that we know x_0 as an initial condition for the problem. Thus, the recursion relation goes back from the final stage in a first pass and then forward from the initial condition in a second stage.

7.13 HJB for LQR. Using Example 7.4 and the *ansatz* $J^* = \frac{1}{2}\boldsymbol{x}^\mathsf{T}\boldsymbol{S}\boldsymbol{x}$, derive the steady-state Linear Quadratic Regulator (Eq. 7.17) by starting from the Hamilton–Jacobi–Bellman equation (Eq. 7.31). The running cost is $L = \frac{1}{2}(\boldsymbol{x}^\mathsf{T}\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{u}^\mathsf{T}\boldsymbol{R}\boldsymbol{u})$.

Solution.

The time-independent HJB equation is

$$\inf_{\boldsymbol{u}} \left[L(\boldsymbol{x}, \boldsymbol{u}) + (\partial_{\boldsymbol{x}} J^*)\ \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u})\right] = 0\,.$$

From the equations of motion are $\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}$, we have $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}$, so that

$$\inf_{\boldsymbol{u}} \left[\tfrac{1}{2}(\boldsymbol{x}^\mathsf{T}\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{u}^\mathsf{T}\boldsymbol{R}\boldsymbol{u}) + (\partial_{\boldsymbol{x}} J^*)\ (\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u})\right] = 0\,.$$

The suggested *ansatz* $J^* = \frac{1}{2}\boldsymbol{x}^\mathsf{T}\boldsymbol{S}\boldsymbol{x}$ implies first of all that $\boldsymbol{S}$ is symmetric, since any antisymmetric component will contribute 0 to J^*. Thus,

$$(\partial_{\boldsymbol{x}} J^*) = \boldsymbol{x}^\mathsf{T}\boldsymbol{S}\,.$$

Hence,

$$\inf_{\boldsymbol{u}} \left[\tfrac{1}{2}(\boldsymbol{x}^\mathsf{T}\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{u}^\mathsf{T}\boldsymbol{R}\boldsymbol{u}) + \left(\boldsymbol{x}^\mathsf{T}\boldsymbol{S}\right)(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}) \right] = 0\,.$$

Because $\boldsymbol{u}$ is unbounded, the "inf" is found by taking $\partial_{\boldsymbol{u}}$ and setting to zero:

$$\boldsymbol{u}^\mathsf{T}\boldsymbol{R} + \boldsymbol{x}^\mathsf{T}\boldsymbol{S}\,\boldsymbol{B} = \boldsymbol{0}^\mathsf{T}\,.$$

Taking a transpose and remembering that $\boldsymbol{R}$ and $\boldsymbol{S}$ are symmmetric gives,

$$\boldsymbol{u} = -(\boldsymbol{R}^{-1}\boldsymbol{B}^\mathsf{T}\boldsymbol{S})\boldsymbol{x}\,,$$

which is Equation (7.14). We then substitute $\boldsymbol{u}$ back into the HJB equation:

$$\begin{aligned} \tfrac{1}{2}\left[\boldsymbol{x}^\mathsf{T}\boldsymbol{Q}\boldsymbol{x} + (\boldsymbol{x}^\mathsf{T}\boldsymbol{S}\boldsymbol{B}\boldsymbol{R}^{-1})\boldsymbol{R}(\boldsymbol{R}^{-1}\boldsymbol{B}^\mathsf{T}\boldsymbol{S}\boldsymbol{x})\right] + \left(\boldsymbol{x}^\mathsf{T}\boldsymbol{S}\right)\left[\boldsymbol{A}\boldsymbol{x} - \boldsymbol{B}(\boldsymbol{R}^{-1}\boldsymbol{B}^\mathsf{T}\boldsymbol{S})\boldsymbol{x}\right] &= 0 \\ \boldsymbol{x}^\mathsf{T}\left[\tfrac{1}{2}\left(\boldsymbol{Q} + \boldsymbol{S}\boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^\mathsf{T}\boldsymbol{S}\right) + \tfrac{1}{2}\left(\boldsymbol{S}\boldsymbol{A} + \boldsymbol{A}^\mathsf{T}\boldsymbol{S}\right) - \boldsymbol{S}\boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^\mathsf{T}\boldsymbol{S}\right]\boldsymbol{x} &= 0\,, \end{aligned}$$

which implies that

$$\boldsymbol{Q} + \boldsymbol{S}\boldsymbol{A} + \boldsymbol{A}^\mathsf{T}\boldsymbol{S} - \boldsymbol{S}\boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^\mathsf{T} = \boldsymbol{0}\,,$$

which is Eq. (7.17). Note the decomposition $\boldsymbol{S}\boldsymbol{A} \to \frac{1}{2}(\boldsymbol{S}\boldsymbol{A} + \boldsymbol{A}^\mathsf{T}\boldsymbol{S})$. This follows because the condition $\boldsymbol{x}^\mathsf{T}[\cdots]\boldsymbol{x} = 0$ implies that the *symmetric* part of the bracketed terms $[\cdots] = 0$. The linear combination isolates the symmetric part of $\boldsymbol{S}\boldsymbol{A}$. There is no constraint placed on antisymmetric terms.

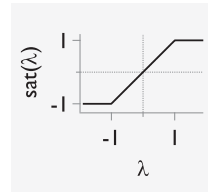
7.14 Anti-windup control. In Section 7.5.1, we showed that we could improve a PI controller's performance by ensuring that the integrator does not update if the controller value would exceed its physical limits. Here, the goal is to reproduce Figure 7.5.

a. Start with parts (a) and (b). Although you could use the step response function of standard control packages, write a simple forward Euler routine to integrate the equations, $\dot{x} = u$, $u(t) = K_0\left(e + \int_0^t \mathrm{d}t'\, e(t')\right)$ directly. Here, $e(t) = x_r - x(t)$, $x_r = 5$, and $K_0 = 1$. Recall that for forward Euler, $\dot{x} \approx \frac{1}{T_\mathrm{s}}(x_{k+1} - x_k)$, with T_s the time step. Find a time step T_s that is small enough that numerical accuracy is good.
b. To reproduce (c), impose saturation, $|u| \le 1$. In your code, distinguish the signal $v(t)$ that the controller would send to the system from $u(t)$, the signal actually sent.
c. To reproduce (d), add anti-windup control: whenever $|v(t)| > 1$, freeze the integral value by disabling the update.

Solution.

Code using the standard numerical ODE routine of *Mathematica*, `NDSolveValue`, which allows constraints, is given on the book website. Figure 7.5 was actually produced with simpler code based on a first-order Euler discretization. The latter is closer to a digital implementation of a PID loop. With a short-enough time step, there is little difference between the two solutions.

7.15 One-dimensional constrained optimization. Finish the analysis in Example 7.5.



a. Show that the Pontrayagin minimum principle implies that $u(t) = -\text{sat}[\lambda(t)]$, where the saturation function sat limits $\lambda(t)$ to ± 1 (see right).
b. By minimizing the Hamiltonian $H(x, \lambda, u)$, show that the the crossover between constrained and unconstrained dynamics occurs at $\tau = \ln[(1+x_0)/(2+\sqrt{2})]$. Or, make a less-rigorous argument by assuming that $u(t)$ is continuous at $t = \tau$.
c. Generate the plot shown in the example, with $x_0 = 5$ and $\tau \approx 0.56$.

Solution.

a. The control Hamiltonian is $H = \frac{1}{2}(x^2 + u^2) + \lambda(-x + u)$. The PMP asks us to choose, at each instant in time, $u(t)$ to minimize $H(x, \lambda, u)$, while respecting the constraint $|u| \leq 1$. There are two cases:
   - The constraint is not active. Then we can take $\partial_u H = u + \lambda = 0$, which implies $u(t) = -\lambda(t)$.
   - The constraint is active. Then $u = \pm 1$. Since $u^2 = +1$, the term plays no role in deciding which u to choose. Rather, we minimize the term $+\lambda u$, which implies taking $u = -\text{sign}(\lambda)$.

   Putting these conditions together, we have

$$u(t) = -\text{sat}[\lambda(t)] = \begin{cases} -1, & \lambda \geq 1 \\ -\lambda, & 0 < \lambda < 1 \\ +1, & \lambda \leq -1 \,. \end{cases}$$

b. First, the less-rigorous argument: The two solutions are easy to calculate. For $t < \tau$, we expect the constrained solution, where $u = -1$. Then $\dot{x} = -x - 1$, with $x(0) = x_0$ implies that $x(t) = -1 + (x_0 + 1)\,\mathrm{e}^{-t}$ and, hence at the crossover, $x(\tau) = -1 + (x_0 + 1)\,\mathrm{e}^{-\tau}$.
   The other solution has $u = -K^* x$, with $K^* = \sqrt{2} - 1$. This goes from $t = \tau$, where $x = x(\tau)$ (given above). Using the results from Section 7.1 implies $x(t) = x_\tau\,\mathrm{e}^{-\sqrt{2}(t-\tau)}$ and $u(\tau) = -K^* x(\tau)$.
   At the crossover, we can equate the two expressions for $u(\tau)$, giving

$$u(\tau) = -1 = -K^* x(\tau) = -\left(\sqrt{2} - 1\right)\left[-1 + (x_0 + 1)\,\mathrm{e}^{-\tau}\right] .$$

   Solving for τ gives the requested crossover time.

   The above argument assumed a lot about the structure of the solution. A more fundamental way to approach the problem is to evaluate the Hamiltonian along the solution. We know that it is constant. For the unconstrained solution, as $t \to \infty$, the functions x, λ, and u all tend to zero, implying that $H = \frac{1}{2}(x^2 + u^2) + \lambda(-x + u) \to 0$, too. Thus, our task is equivalent to evaluating the Hamiltonian for the first part of the solution. Using *Mathematica* to evaluate $\lambda(t)$ and then H, we find that

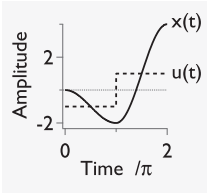
$$H = 1 - \left(2 - \sqrt{2}\right)(1 + x_0)\,\mathrm{e}^{-\tau} + \left(\tfrac{3}{2} - \sqrt{2}\right)(1 + x_0)^2\,\mathrm{e}^{-2\tau} .$$

Substituting the expression for the crossover time or solving the quadratic equation for $(1 + x_0)\,e^{-\tau}$ gives,

$$e^{-\tau} = \frac{2 + \sqrt{2}}{1 + x_0}, \quad \Longrightarrow \quad H = 1 - 2 + \underbrace{\left(\tfrac{3}{2} - \sqrt{2}\right)\left(2 + \sqrt{2}\right)^2}_{1} = 0\,.$$

Thus, H has a minimum $H = 0$ for the crossover τ given above. Choosing this value of τ gives $H = 0$ for the entire solution, as expected.

c. See book website for code.



7.16 Bang-bang control of a harmonic oscillator. Consider an undamped harmonic oscillator, $\ddot{x} + x = u$, where the piecewise-continuous forcing $u(t)$ is restricted to the range $|u(t)| \le 1$. Starting from initial conditions $x(0) = \dot{x}(0) = 0$, find the piecewise-continuous control that maximizes $x(\tau)$. Encode this goal in J by setting the penalty $L(\boldsymbol{x}, u, t) = 0$ for $0 \le t < \tau$ and the end-time penalty $\varphi[x(\tau)] = -x(\tau)$.

a. Solve for $u^*(t)$ using the Hamilton–Jacobi–Bellman equation.
b. Solve again using the Pontryagin Minimum Principle.
c. Show that if $\tau = 2\pi$, then $x(\tau) = 4$. Plot $u^*(t)$ and $x^*(t)$ for $0 \le t \le 2\pi$.
d. If you "knew" that $u(t)$ switched from -1 to $+1$ at an unknown time τ_0, then you could compute $J(\tau_0, \tau) \equiv -x(\tau)$ and minimize J directly. Do this analytically for this problem and confirm the result of part (c). [See Kappen (2011).]

Solution.

a. We rewrite the equation of motion in vector form, defining $x_1(t) = x(t)$ and $x_2(t) = \dot{x}(t)$:

$$\dot{\boldsymbol{x}} = \begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \boldsymbol{f}(\boldsymbol{x}, u) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u = \begin{pmatrix} x_2 \\ -x_1 + u \end{pmatrix}$$

Then the HJB equation, Eq. (7.31), for the value function $J(\boldsymbol{x}, u, t)$ is

$$\partial_t J + \min_{u \in [-1,1]} \left[x_2 \partial_{x_1} J - x_1 \partial_{x_2} J + u \partial_{x_2} J \right] = 0\,.$$

We have to minimize over u at each point in time. Because the HJB is linear in u, the optimal $u(t)$ is ± 1 and is given by

$$u(t) = -\text{sign}\left(\partial_{x_2} J\right)\,.$$

Since $|x| = \text{sign}(x)\, x$, the HJB is now

$$\partial_t J + x_2 \partial_{x_1} J - x_1 \partial_{x_2} J - |\partial_{x_2} J| = 0\,,$$

with boundary condition $J(\tau) = -x_1(\tau)$. We can solve this partial differential equation using the *ansatz* that J is linear in $\boldsymbol{x}$:

$$J(\boldsymbol{x}, t) = f_0(t) + x_1 f_1(t) + x_2 f_2(t)\,.$$

Substituting this form of J into the HJB gives $\dot{f}_1 = -f_2$, $\dot{f}_2 = f_1$, and $\dot{f}_0 = |f_2|$, with boundary conditions $f_0(\tau) = f_2(\tau) = 0$ and $f_1(\tau) = -1$. The solution is

$$f_1(t) = -\cos(t-\tau)\,, \qquad f_2(t) = \sin(t-\tau)\,.$$

The optimal control is then

$$u^*(t) = -\text{sign}[\sin(t-\tau)]$$

Given this control, we can integrate the equations of motion forward from time $t = 0$, to find $x(t)$. There is no need to solve for $f_0(t)$, unless we want to evaluate the cost-to-go itself. The solution for $x(t)$ is complicated algebraically to express for general τ but simple once τ is fixed. We explore a specific protocol duration below.

b. The control Hamiltonian $H(\boldsymbol{x}, \boldsymbol{\lambda}, u)$ is given by

$$H = 0 + \lambda_1 x_2 + \lambda_2(-x_1 + u)\,.$$

Minimizing over u at each t, we have

$$u^* = -\text{sign}(\lambda_2)\,.$$

Then Hamiltonian is then

$$\mathcal{H} = \lambda_1 x_2 - \lambda_2 x_1 - |\lambda_2|\,,$$

and Hamilton's equations give

$$\dot{x}_1 = \partial_{\lambda_1}\mathcal{H} = x_2\,, \qquad \dot{x}_2 = \partial_{\lambda_2}\mathcal{H} = -x_1 - \text{sign}(\lambda_2)$$
$$\dot{\lambda}_1 = -\partial_{x_1}\mathcal{H} = \lambda_2\,, \qquad \dot{\lambda}_2 = -\partial_{x_2}\mathcal{H} = -\lambda_1\,,$$

with boundary conditions

$$\boldsymbol{x}(0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \boldsymbol{\lambda}(\tau) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

The solution for $\boldsymbol{\lambda}(t)$ is

$$\lambda_1 = \cos(t-\tau)\,, \qquad \lambda_2 = \sin(t-\tau)\,,$$

and

$$u^*(t) = -\text{sign}(\lambda_2) = -\text{sign}[\sin(t-\tau)]\,.$$

We thus find the same $u^*(t)$ and hence same $x^*(t)$ as before.

c. With $\tau = 2\pi$, we can evaluate the control explicitly:

$$u^*(t) = \begin{cases} -1 & 0 \le t < \pi \\ +1 & \pi \le t < 2\pi \end{cases}.$$

and, hence,

$$x(t) = \begin{cases} -1 + \cos(t) & 0 < t \le \pi\,, \\ 1 + 3\cos(t) & \pi < t \le 2\pi\,, \end{cases}$$

which is gives $x(2\pi) = 4$. This is the largest possible end-time amplitude at $\tau = 2\pi$, given that $u(t)$ is restricted to be in the range of $[-1, 1]$. Intuitively, we push down until the spring stretches to its maximum negative value, and then we jerk up to get the biggest possible amplitude.

d. As discussed briefly in the main text, the Pontryagin minimum principle predicts that there should be $n - 1$ sign changes for an n-th order linear system, which implies that there should be a single switch at some unknown time τ_0 during our protocol. Intuitively, the first period should have $u = -1$ and the second $u = +1$, so we can calculate directly the solutions using these values of u. That is, we solve

$$\begin{aligned} \ddot{x} + x &= -1 \qquad x(0) = \dot{x}(0) = 0\,, \qquad 0 < t < \tau_0 \\ \ddot{x} + x &= +1 \qquad x = x(\tau_0) = \dot{x} = \dot{x}(\tau_0)\,, \quad \tau_0 < t < \tau\,, \end{aligned}$$

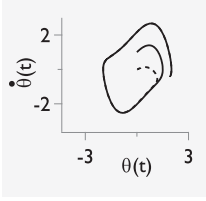
which leads to

$$x(t) = \begin{cases} -1 + \cos t & 0 < t < \tau_0 \\ 1 + \cos t - 2\cos(t - \tau_0) & \tau_0 < t < \tau\,. \end{cases}$$

In the spirit of simplicity (for this part of the problem), we just want to maximize

$$\begin{aligned} x(\tau = 2\pi) &= 1 + \cos 2\pi - 2\cos(2\pi - \tau_0) \\ &= 2(1 - \cos\tau_0)\,, \end{aligned}$$

by varying τ_0. Clearly, this is maximized for $\tau_0 = \pi$, which makes $x(2\pi) = 4$.



7.17 Observer for van der Pol oscillator. Design an observer for the van der Pol equation, $\dot{x}_1 = x_2$, $\dot{x}_2 = \epsilon(1 - x_1^2)x_2 - x_1$, with output $y = x_1$.

a. Derive a time-dependent, linear equation for the error $\boldsymbol{e} = \boldsymbol{x} - \hat{\boldsymbol{x}}$.
b. Find gains $\boldsymbol{L}$ that give critically damped observer dynamics.
c. Plot $\dot{\theta}$ and $\dot{\hat{\theta}}$, θ and $\hat{\theta}$ versus time and each other for $\epsilon = 1$ and $\ell_1 = 2$. See left.

Solution.

a. In state-space form, the van der Pol dynamics are

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= \epsilon(1 - x_1^2)x_2 - x_1\,. \end{aligned}$$

The observer dynamics are

$$\begin{aligned} \frac{\mathrm{d}\hat{x}_1}{\mathrm{d}t} &= \hat{x}_2 + \ell_1(x_1 - \hat{x}_1) \\ \frac{\mathrm{d}\hat{x}_1}{\mathrm{d}t} &= \epsilon(1 - \hat{x}_1^2)\hat{x}_2 - \hat{x}_1 + \ell_2(x_1 - \hat{x}_1)\,. \end{aligned}$$

The error $\boldsymbol{e} = \boldsymbol{x} - \hat{\boldsymbol{x}}$ then has dynamics $\dot{\boldsymbol{e}} = \boldsymbol{A}\boldsymbol{e}$, with

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ -1 - 2\epsilon\hat{x}_1\hat{x}_2 & \epsilon(1-\hat{x}_1^2) \end{pmatrix},$$

and $\boldsymbol{A}' = \boldsymbol{A} - \boldsymbol{L}\boldsymbol{C}$ is

$$\boldsymbol{A}' = \begin{pmatrix} -\ell_1 & 1 \\ -1 - 2\epsilon\hat{x}_1\hat{x}_2 - \ell_2 & \epsilon(1-x_1^2) \end{pmatrix}.$$

Since $\hat{x}_1$ and $\hat{x}_2$ depend on time, so does $\boldsymbol{A}'$.

b. The characteristic equation is

$$(\lambda + \ell_1)\left(\lambda - \epsilon(1-\hat{x}_1^2)\right) + 1 + \ell_2 + 2\epsilon\hat{x}_1\hat{x}_2 = 0\,.$$

With help from a computer-algebra program, we find the eigenvalues

$$\lambda = \frac{1}{2}\left(1 - \hat{x}_1^2 - \ell_1 \pm \sqrt{-3 + 2\ell_1 + \ell_1^2 - 4\ell_2 - 2\hat{x}_1^2 - 2\ell_1\hat{x}_1^2 + \hat{x}_1^4 - 8\hat{x}_1\hat{x}_2}\right).$$

The condition for critical damping is that the square root vanish:

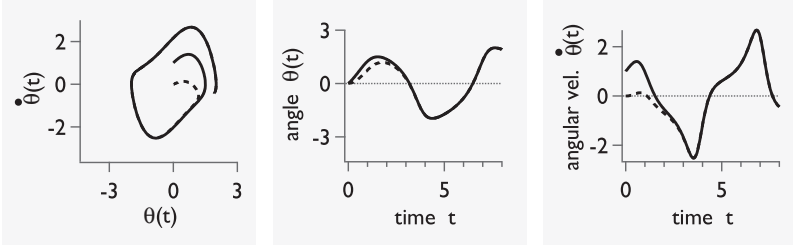
$$-3 + 2\ell_1 + \ell_1^2 - 4\ell_2 - 2\hat{x}_1^2 - 2\ell_1\hat{x}_1^2 + \hat{x}_1^4 - 8\hat{x}_1\hat{x}_2 = 0\,,$$

which implies

$$\ell_2 = \tfrac{1}{4}\left(-3 + 2\ell_1 + \ell_1^2 - 2\hat{x}_1^2 - 2\ell_1\hat{x}_1^2 + \hat{x}_1^4 - 8\hat{x}_1\hat{x}_2\right).$$

c. The plots are for $\epsilon = 1$, $x_1 = 0$, $x_2 = 1$, and $\ell_1 = 2$ and are given below. For critical damping, the formula from part (b) implies

$$\ell_2 = \tfrac{1}{4}\left(5 - 6\hat{x}_1^2 + \hat{x}_1^4 - 8\hat{x}_1\hat{x}_2\right).$$



7.18 Pendulum swing up: numerics. A simple numerical method to solve the pendulum swing-up boundary-value problem is to discretize time functions.

a. Rewrite the equations of motion given in Example 7.2 to eliminate $u(t)$. Write them as a four-component vector $\dot{\boldsymbol{z}} = \theta, \dot{\theta}, \lambda, \dot{\lambda}$ obeying $\dot{\boldsymbol{z}} = \boldsymbol{h}(\boldsymbol{z})$.

b. Use a standard numerical routine to solve the pendulum swing-up boundary-value problem for a given value of τ. Confirm the plots shown in Figure 7.2.

c. Define n time intervals $\Delta t = \tau/n$ and denote $\boldsymbol{z}_k$ the values of the four components at time $t = k\,\Delta t$, with $k \in \{0, \ldots, n\}$. From the trapezoidal rule of integration,

$$\boldsymbol{z}_{k+1} = \boldsymbol{z}_k + \tfrac{1}{2}\Delta t\left(\boldsymbol{h}(\boldsymbol{z}_k) + \boldsymbol{h}(\boldsymbol{z}_{k+1})\right), k = 0, \ldots n-1\,,$$

and adding the four boundary conditions, write down a coupled set of $4(n+1)$ nonlinear algebraic equations for the $4(n+1)$ variables z_k. Express your equations in the form $\boldsymbol{h}(z_k) = \mathbf{0}$. For coding, write them out explicitly, too.

d. Solve these equations using a standard numerical root finder. You can try Newton's method, which requires calculating the Jacobian matrix for $\boldsymbol{h}(z_k)$ (tricky). The secant method, which approximates the Jacobian using finite differences, is simpler and also works. Confirm Figures 7.8 and 7.2.

Solution.

a. The original equations for the pendulum state and adjoint are given by

$$\ddot{\theta}(t) + \sin\theta(t) = u(t)\,, \qquad \theta(0) = \dot{\theta}(0) = 0\,, \qquad \theta(\tau) = \pi\,, \dot{\theta}(\tau) = 0\,.$$

and

$$\ddot{\lambda}(t) + \lambda(t)\,\cos\theta(t) = 0\,, \qquad u(t) = -\lambda(t)\,.$$

If we eliminate $u(t)$ in favor of $\boldsymbol{\theta}(t)$ and $\boldsymbol{\lambda}(t)$ alone and convert to first-order equations, the equations of motion are

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\theta\\ \dot{\theta}\\ \lambda\\ \dot{\lambda}\end{pmatrix} = \begin{pmatrix}\dot{\theta}\\ -\sin\theta - \lambda\\ \dot{\lambda}\\ -\lambda\cos\theta\end{pmatrix}, \qquad \theta(0) = \dot{\theta}(0) = 0\,, \quad \theta(\tau) = \pi\,,\ \dot{\theta}(\tau) = 0\,.$$

With $z = (\theta, \dot{\theta}, \lambda, \dot{\lambda})^{\mathsf{T}}$, the four equations of motion can be written as $\dot{z} = \boldsymbol{h}(z)$.

b. See code on book website, based on *Mathematica*'s `NDSolveValue`.

c. Written explicitly, the $4(n+1)$ equations are

$$\begin{aligned}
\theta_0 &= 0\\
\dot{\theta}_0 &= 0\\
&\vdots\\
\theta_{k+1} - \theta_k - \tfrac{1}{2}\Delta t\,(\dot{\theta}_k + \dot{\theta}_{k+1}) &= 0\\
\dot{\theta}_{k+1} - \dot{\theta}_k - \tfrac{1}{2}\Delta t\,(-\sin\theta_k - \sin\theta_{k+1} - \lambda_k - \lambda_{k+1}) &= 0\\
\lambda_{k+1} - \lambda_k - \tfrac{1}{2}\Delta t\,(\dot{\lambda}_k + \dot{\lambda}_{k+1}) &= 0\\
\dot{\lambda}_{k+1} - \dot{\lambda}_k - \tfrac{1}{2}\Delta t\,(-\lambda_k\cos\theta_k - \lambda_{k+1}\cos\theta_{k+1}) &= 0\\
&\vdots\\
\theta_n - \pi &= 0\\
\dot{\theta}_n &= 0\,,
\end{aligned}$$

where k ranges from 0 to $n-1$.

d. The challenge in coding is to formulate the list of nonlinear equations for an arbitrary number $n+1$ grid points (or n intervals). One way to do this is to define a band-diagonal matrix (plus a couple of odd elements for the

boundary conditions) that, when multiplying the vector z_k gives the appropriate equations. Alternatively, *Mathematica* has built-in tools that make the task even easier.

See the book website for code.

7.19 Pendulum swing up: adding feedback. Add linear feedback to the swing-up-and-balance protocol. First, calculate the nominal optimal control $u_{\text{ff}}(t)$ and $\theta_{\text{ff}}(t)$ (Problem 7.18). To find a linear feedback law for small deviations, assume a cost function where the weight Q on each state deviation matches the weight R on control effort.

a. Calculate the linear feedback $\boldsymbol{K} = (k_1 \;\; k_2)$ gains three ways:
   i. Use the time-independent LQR gains for the upright, balanced state, $k_1 = k_2 = 1 + \sqrt{2}$. Standard LQR routines will give this result numerically. Find it analytically by solving the algebraic Riccati equations. (Cf. Problem 7.8.)
   ii. Find $k_1[\theta_{\text{ff}}(t)]$ and $k_2[\theta_{\text{ff}}(t)]$ assuming the quasistationary approximation.
   iii. Find the optimal $k_1(t)$ and $k_2(t)$ by solving the time-dependent Riccati equation assuming the dynamical matrix $\boldsymbol{A}(t) = \boldsymbol{A}[\theta_{\text{ff}}(t)]$.

   Plot and discuss $k_1(t)$ and $k_2(t)$ for the three cases.
b. Add feedback to your numerical code and produce plots resembling those in Example 7.7. Recall that a "kick" to $\dot{\theta}(t)$ imposes a slope discontinuity on $\theta(t)$. Show that all three feedback schemes give very nearly the same response.

Solution.

a. Here are three ways of computing the linear feedback gains $\boldsymbol{K} = (k_1 \;\; k_2)$:
   i. This is a standard steady-state LQR problem, with

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ -\cos\theta & 0 \end{pmatrix}, \qquad \boldsymbol{B} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad \boldsymbol{Q} = \begin{pmatrix} Q & 0 \\ 0 & Q \end{pmatrix}, \qquad \boldsymbol{R} = R,$$

   with $\theta = \pi$ and $Q = R$. (From rescaling J, it is clear that only the ratio Q/R matters).

   ii. To find the gains as a function of θ analytically, we formulate the steady-state Riccati equations:

$$-\boldsymbol{Q} - \boldsymbol{A}^{\mathsf{T}}\boldsymbol{S} - \boldsymbol{S}\boldsymbol{A} + \boldsymbol{S}\boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{S} = \boldsymbol{0}, \qquad \text{with } \boldsymbol{S} = \begin{pmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{pmatrix}.$$

   This leads to

$$\begin{pmatrix} Q - \frac{s_{12}^2}{R} - 2s_{12}\cos\theta & s_{11} - \frac{s_{12}s_{22}}{R} - s_{22}\cos\theta \\ s_{11} - \frac{s_{12}s_{22}}{R} - s_{22}\cos\theta & Q + 2s_{12} - \frac{s_{22}^2}{R} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Solving these algebraic equations then gives

$$s_{12} = -R\cos\theta + \sqrt{QR + R^2\cos^2\theta}$$
$$s_{22} = \sqrt{QR - 2R^2\cos\theta + 2R\sqrt{R(Q + R\cos\theta^2)}}\,.$$

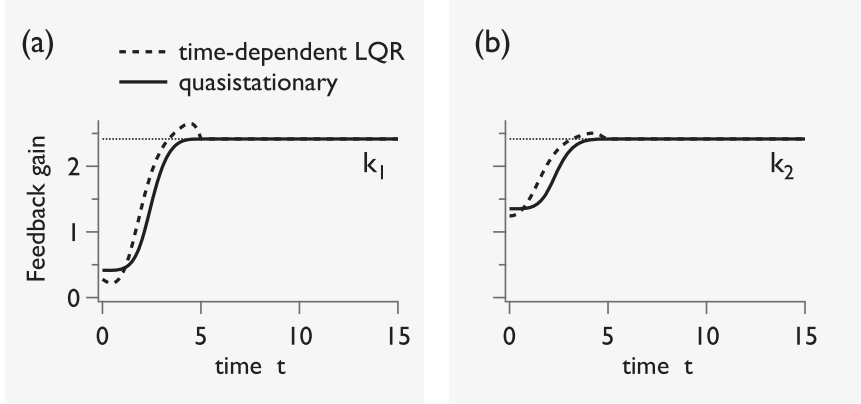
The gain is given by

$$\begin{aligned}\boldsymbol{K} = R^{-1}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{S} &= \begin{pmatrix}\frac{s_{12}}{R} & \frac{s_{22}}{R}\end{pmatrix} \\ &= \left(-\cos\theta + \sqrt{Q/R + \cos^2\theta} \quad \sqrt{Q/R - 2\cos\theta + 2\sqrt{(Q/R + \cos\theta^2)}}\right) \\ &\to \left(-\cos\theta + \sqrt{1 + \cos^2\theta} \quad \sqrt{1 - 2\cos\theta + 2\sqrt{(1 + \cos\theta^2)}}\right).\end{aligned}$$

As expected, the gains depend only on Q/R. The last line evaluates the gains for $Q/R = 1$.

iii. For the fully time-dependent, the matrix $\boldsymbol{A}$ depends on time via its dependence on the linearization of the nominal trajectory, i.e., via $\theta_{\text{ff}}(t)$. We proceed as in (ii) but now solve (numerically) the three coupled nonlinear ODEs arising from the matrix Riccati equation. Notice that because we integrate backwards from the end of the balance protocol, the values (neglecting any final transient) are just the values found in (i). Here, they serve as final conditions that start the integration at time $t = \tau$. The equations are integrated from τ to 0.

Plotting all three gains for $Q/R = 1$ vs. time gives the figure below.



The dotted line is $1 + \sqrt{2}$, the steady-state gain in (a). The solid lines show the quasistationary solutions, while the dashed lines show the full LQR solutions. The main observation is that the gain is smaller when the pendulum is down than when it is up. This makes sense: we always need more gain to stabilize an unstable system. We also see that there is more feedback gain on velocity perturbations near $\theta = 0$, while the penalties are larger on both (but equal) near $\theta = \pi$. This observation seems less intuitive.

b. See website for code. The precise perturbation was left unspecified, as there many types one can impose. If we take the time-dependent LQR in (iii) as our reference, the constant-gain approximation in (i) differs by about 10% whereas the approximation in (ii) differs by about 1%. We can see this in the plots in (a): there is a factor of two for small times between (i) and (iii) whereas (ii) and (iii) differ 20%. The differences are smaller at larger θ. On the other hand, we likely care more about the behavior near $\theta \approx \pi$, where all three schemes give nearly the same feedback gains.

In practice, even the simplest scheme in (i) would likely be good enough. An issue with scheme (iii) is that if the feedforward is recomputed, we would also have to recompute the feedback gains. In (ii), we would usually have to solve the gains numerically, too.

8 Stochastic Systems

Problems

8.1 **Kalman filter for prediction observer**. In the text, we have formulated Kalman filters in terms of the current observer, which uses observations up to y_{k+1} in the estimate $\hat{\boldsymbol{x}}_{k+1}$. The prediction observer uses only up to y_k and is appropriate for cases where sensor and computational delays are on the order of T_s (Section 5.4.2).

a. Redo the 1d-Kalman filter calculations for the prediction observer, defining

$$\begin{aligned} e_k &= x_k - \hat{x}_k & x_{k+1} &= x_k + \nu_k & y_k &= x_k + \xi_k \\ P_k &= \langle e_k^2 \rangle & \hat{x}_{k+1} &= \hat{x}_k + L(y_k - \hat{y}_k) & \hat{y}_k &= \hat{x}_k \,. \end{aligned}$$

Calculate the recurrence relations for the optimal L_k^* and P_k^*, as well as the steady state P^* and L^*. You should find that $P^* = L^*\xi^2 + \nu^2$ (cf. the previous result, $P^* = L^*\xi^2$). Comment on L^* and P^* in the limits $\alpha \ll 1$ and $\gg 1$, where $\alpha \equiv \nu^2/\xi^2$.

b. Generalize to an n-dimensional MIMO system, by defining

$$\begin{aligned} \boldsymbol{e}_k &= \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k & \boldsymbol{x}_{k+1} &= \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{\nu}_k & \boldsymbol{y}_k &= \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{\xi}_k \\ \boldsymbol{P}_k &= \langle \boldsymbol{e}_k \boldsymbol{e}_k^{\mathsf{T}} \rangle & \hat{\boldsymbol{x}}_{k+1} &= \boldsymbol{A}\hat{\boldsymbol{x}}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{L}(\boldsymbol{y}_k - \hat{\boldsymbol{y}}_k) & \hat{\boldsymbol{y}}_k &= \boldsymbol{C}\hat{\boldsymbol{x}}_k \,, \end{aligned}$$

Show that the recurrence relations for the time-dependent Kalman filter become

$$\begin{aligned} &\boldsymbol{P}_{k+1}^y = \boldsymbol{C}\boldsymbol{P}_k\boldsymbol{C}^{\mathsf{T}} + \boldsymbol{Q}_\xi \,, \quad \boldsymbol{P}_{k+1}^{xy} = \boldsymbol{P}_k\boldsymbol{C}^{\mathsf{T}} \,, \qquad \boldsymbol{L}_{k+1}^* = \boldsymbol{A}\boldsymbol{P}_{k+1}^{xy}\left(\boldsymbol{P}_{k+1}^y\right)^{-1} \,, \\ &\boldsymbol{P}_{k+1}^* = \boldsymbol{A}\boldsymbol{P}_k^*\boldsymbol{A}^{\mathsf{T}} + \boldsymbol{Q}_\nu - \boldsymbol{L}_{k+1}^*\boldsymbol{P}_k^y\boldsymbol{L}_{k+1}^{\mathsf{T}} \,. \end{aligned}$$

Show that, in steady state, $\boldsymbol{P}^*$ obeys an algebraic Riccati equation that maps precisely onto Eq. (7.24) from the discussion on LQR optimal control in Chapter 7. Comment on the different values of $\boldsymbol{L}^*$ for the two forms of Kalman filter.

Solution.

a. The calculations in the text are slightly modified. We have

$$\begin{aligned} e_{k+1} &= x_{k+1} - [\hat{x}_k + L(y_k - \hat{y}_k)] \\ &= x_k + \nu_k - [\hat{x}_k + L(x_k + \xi_k - \hat{x}_k)] \end{aligned}$$

$$= e_k + \nu_k - L(e_k + \xi_k)$$
$$= (1 - L)e_k + \nu_k - L\xi_k \,.$$

This is the error at time $k+1$ in terms of the error at time k and the noise and observer gains. The variance is then

$$P_{k+1} = \langle e_{k+1}^2 \rangle = (1 - L)^2 P_k + \nu^2 + L^2\xi^2 \,,$$

and

$$\frac{\mathrm{d}P_{k+1}}{\mathrm{d}L} = -2(1 - L)P_k + 2L\xi^2 = 0$$
$$\implies L_{k+1}^* = \frac{P_k}{P_k + \xi^2} \,.$$

Substituting L_{k+1}^* to find the optimal P_{k+1}, we have

$$P_{k+1}^* = \frac{\xi^4 P_k^*}{(P_k^* + \xi^2)^2} + \frac{P_k^{*2}\xi^2}{(P_k^* + \xi^2)^2} + \nu^2$$
$$= \xi^2 L_{k+1}^* + \nu^2 \,.$$

The steady-state equations are

$$L^* = \frac{P^*}{P^* + \xi^2} \,, \quad P^* = L\xi^2 + \nu^2 \,.$$

Eliminating P^*, we find that L^* satisfies the same quadratic equation as before,

$$L^{*2} + \alpha L^* - \alpha = 0 \,, \quad \alpha \equiv \frac{\nu^2}{\xi^2} \,, \qquad \implies \qquad L^* = -\frac{1}{2}\left(-\alpha + \sqrt{\alpha^2 + 4\alpha}\right) \,.$$

Thus, the Kalman observer gain remains the same, but the variance is larger by ν^2. This makes sense: if we base our estimate at time $k+1$ on y_k, then the system is unobserved for a time T_s, and P^* is accordingly larger.
Now let's look at the limits $\alpha \ll 1$ and $\gg 1$.

$$\alpha \ll 1 : \qquad L^* \approx \sqrt{\alpha} \qquad P^* \approx \xi\nu$$
$$\alpha \gg 1 : \qquad L^* \approx 1 \qquad P^* \approx \xi^2 + \nu^2 \approx \nu^2$$

In the case of small observation noise, we find $P^* \approx \nu^2$ rather than the relation $P \approx \xi^2$ that we found for the current observer. Again, the reason is that the system increases its variance by ξ^2 during the time interval between the last measurement and the estimate. In this limit, this extra variance dominates over the observation variance ξ^2. But in the interesting limit, $\nu \ll \xi$, we find the same result as before, to lowest order.

b. The error in the estimate at time $k+1$ is

$$\boldsymbol{e}_{k+1} = \boldsymbol{x}_{k+1} - \hat{\boldsymbol{x}}_{k+1}$$
$$= [\boldsymbol{A}\boldsymbol{x}_k + \cancel{\boldsymbol{B}\boldsymbol{u}_k} + \boldsymbol{\nu}_k] - [\boldsymbol{A}\hat{\boldsymbol{x}}_k + \cancel{\boldsymbol{B}\boldsymbol{u}_k} + \boldsymbol{L}(\boldsymbol{y}_k - \hat{\boldsymbol{y}}_k)$$
$$= \boldsymbol{A}\boldsymbol{e}_k + \boldsymbol{\nu}_k - \boldsymbol{L}\boldsymbol{\varepsilon}_k \,.$$

The variance $\boldsymbol{P}_{k+1} = \langle \boldsymbol{e}_{k+1}\, \boldsymbol{e}_{k+1}^{\mathsf{T}} \rangle$ is

$$\boldsymbol{P}_{k+1} = \boldsymbol{A}\boldsymbol{P}_k\boldsymbol{A}^{\mathsf{T}} + \boldsymbol{Q}_\nu - \boldsymbol{A}\boldsymbol{P}_k^{xy}\boldsymbol{L}^{\mathsf{T}} - \boldsymbol{L}\left(\boldsymbol{P}_k^{xy}\right)^{\mathsf{T}}\boldsymbol{A}^{\mathsf{T}} + \boldsymbol{L}\boldsymbol{P}_k^{y}\boldsymbol{L}^{\mathsf{T}},$$

where $\boldsymbol{P}_{k+1}^{y}$ is the covariance of the innovations,

$$\begin{aligned}\boldsymbol{P}_k^{y} &\equiv \left\langle \boldsymbol{\varepsilon}_k\, \boldsymbol{\varepsilon}_k^{\mathsf{T}} \right\rangle \\ &= \left\langle (\boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{\xi}_k - \boldsymbol{C}\hat{\boldsymbol{x}}_k)\, (\boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{\xi}_k - \boldsymbol{C}\hat{\boldsymbol{x}}_k)^{\mathsf{T}} \right\rangle \\ &= \left\langle (\boldsymbol{C}\boldsymbol{e}_k + \boldsymbol{\xi}_k)\, (\boldsymbol{C}\boldsymbol{e}_k + \boldsymbol{\xi}_k)^{\mathsf{T}} \right\rangle \\ &= \boldsymbol{C}\boldsymbol{P}_k\boldsymbol{C}^{\mathsf{T}} + \boldsymbol{Q}_\xi,\end{aligned}$$

and $\boldsymbol{P}_k^{xy}$ is the covariance between the predicted state and the predicted observation,

$$\boldsymbol{P}_k^{xy} \equiv \left\langle \boldsymbol{e}_k\, \boldsymbol{\varepsilon}_k^{\mathsf{T}} \right\rangle = \left\langle \boldsymbol{e}_k\, (\boldsymbol{C}\boldsymbol{e}_k + \boldsymbol{\xi}_k)^{\mathsf{T}} \right\rangle = \boldsymbol{P}_k\boldsymbol{C}^{\mathsf{T}}.$$

Differentiating $\mathrm{Tr}\, \boldsymbol{P}_{k+1}$ with respect to $\boldsymbol{L}$ gives

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{L}}\left(\mathrm{Tr}\, \boldsymbol{P}_{k+1}\right) = -2\left(\boldsymbol{P}_k^{xy}\right)^{\mathsf{T}}\boldsymbol{A}^{\mathsf{T}} + 2\boldsymbol{P}_k^{y}\boldsymbol{L}^{\mathsf{T}} = \boldsymbol{0},$$

so that

$$\boldsymbol{L}_{k+1}^{*} = \boldsymbol{A}\boldsymbol{P}_k^{xy}\left(\boldsymbol{P}_k^{y}\right)^{-1}.$$

Substituting $\boldsymbol{L}_{k+1}^{*}$ to find the minimum value of $\boldsymbol{P}_{k+1}^{*}$ gives

$$\begin{aligned}\boldsymbol{P}_{k+1}^{*} &= \boldsymbol{A}\boldsymbol{P}_k^{*}\boldsymbol{A}^{\mathsf{T}} + \boldsymbol{Q}_\nu - \boldsymbol{A}\boldsymbol{P}_k^{xy}\left(\boldsymbol{P}_k^{y}\right)^{-1}\left(\boldsymbol{P}_k^{xy}\right)^{\mathsf{T}}\boldsymbol{A}^{\mathsf{T}} \\ &= \underbrace{\boldsymbol{A}\boldsymbol{P}_k^{*}\boldsymbol{A}^{\mathsf{T}}}_{\text{dynamics}} + \underbrace{\boldsymbol{Q}_\nu}_{\text{disturbances}} - \underbrace{\boldsymbol{L}_{k+1}^{*}\,\boldsymbol{P}_{k+1}^{y}\boldsymbol{L}_{k+1}^{*\mathsf{T}}}_{\text{observations}},\end{aligned}$$

which is the same expression derived for the current observer (Eq. 8.42)—with a slightly lower optimal gain $\boldsymbol{L}_{k+1}^{*}$.

As before, the Hessian matrix (second derivative of $\mathrm{Tr}\, \boldsymbol{P}_{k+1}$ with respect to $\boldsymbol{L}$) is $2\boldsymbol{P}_k^{y}$, which is positive definite. Thus, the optimal $\boldsymbol{L}^{*}$ ensures that we have minimized the covariances.

The steady-state equations are simply given by dropping the temporal indices. For the covariance matrix, we have a discrete algebraic Riccati equation,

$$\boldsymbol{P} = \boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^{\mathsf{T}} + \boldsymbol{Q}_\nu - \boldsymbol{A}\boldsymbol{P}\boldsymbol{C}^{\mathsf{T}}\left(\boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^{\mathsf{T}} + \boldsymbol{Q}_\xi\right)^{-1}\boldsymbol{C}\boldsymbol{P}\boldsymbol{A}^{\mathsf{T}}.$$

Having solved for $\boldsymbol{P}^{*}$, we can calculate the steady-state gain,

$$\begin{aligned}\boldsymbol{P}^{y} &= \boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^{\mathsf{T}} + \boldsymbol{Q}_\xi \\ \boldsymbol{P}^{xy} &= \boldsymbol{P}\boldsymbol{C}^{\mathsf{T}} \\ \boldsymbol{L}^{*} &= \boldsymbol{A}\boldsymbol{P}^{xy}\left(\boldsymbol{P}^{y}\right)^{-1} = \boldsymbol{A}\boldsymbol{P}^{*}\boldsymbol{C}^{\mathsf{T}}\left(\boldsymbol{C}\boldsymbol{P}^{*}\boldsymbol{C}^{\mathsf{T}} + \boldsymbol{Q}_\xi\right)^{-1}.\end{aligned}$$

Compared to the current-observer form of the Kalman filter, the expression for $\boldsymbol{L}^{*}$ has an extra factor of $\boldsymbol{A}$ in front. With stable discrete dynamics, applying $\boldsymbol{A}$ reduces the "size" of a state vector over a time step T_s, implying that the gain of the prediction observer is lower than that of the current observer.

This makes sense: because it uses older information, the prediction observer puts less weight on observations relative to predictions than does the current observer.

For LQR optimal control, the state and costate decoupled via a matrix $\boldsymbol{S}$, which, in steady state, obeys

$$\boldsymbol{S} = \boldsymbol{A}^\mathsf{T}\left[\boldsymbol{S} - \boldsymbol{S}\boldsymbol{B}(\boldsymbol{R} + \boldsymbol{B}^\mathsf{T}\boldsymbol{S}\boldsymbol{B})^{-1}\boldsymbol{B}^\mathsf{T}\boldsymbol{S}\right]\boldsymbol{A} + \boldsymbol{Q}$$
$$\boldsymbol{K} = (\boldsymbol{R} + \boldsymbol{B}^\mathsf{T}\boldsymbol{S}\boldsymbol{B})^{-1}\boldsymbol{B}^\mathsf{T}\boldsymbol{S}\boldsymbol{A}\,.$$

Comparing with our result for $\boldsymbol{P}$, above, we see that we have

$$\boldsymbol{S} \to \boldsymbol{P}\,, \quad \boldsymbol{A} \to \boldsymbol{A}^\mathsf{T}\,, \quad \boldsymbol{B} \to \boldsymbol{C}^\mathsf{T}\,, \quad \boldsymbol{Q} \to \boldsymbol{Q}_v\,, \quad \boldsymbol{R} \to \boldsymbol{Q}_\xi\,.$$

Note that $\boldsymbol{S}$, $\boldsymbol{P}$, $\boldsymbol{Q}$, $\boldsymbol{Q}_v$, $\boldsymbol{Q}_\xi$, and $\boldsymbol{R}$ are all symmetric matrices.

8.2 **Estimating unstable dynamics**. Consider 1d *deterministic* dynamics with noisy observations. Let $x_{k+1} = ax_k$, with $y_k = x_k + \xi_k$ and x_0 unknown. The noise $\xi_k \sim \mathcal{N}(0, \xi^2)$. When $a > 1$, the dynamics are unstable. Using the prediction observer from Problem 8.1, show that the steady-state variance of the optimal estimate is $P^* = (a^2 - 1)\xi^2$ for $|a| > 1$, and 0 otherwise. Interpret the two cases.

Solution.

From the steady-state discrete algebraic Riccati equation for the variance of the prediction observer, we have

$$P = a^2 P - \frac{a^2 P^2}{P + \xi^2}\,.$$

An immediate solution is $P^* = 0$. For $P^* \neq 0$, we can divide by P and find the other root: $P^* = (a^2 - 1)\xi^2$.

We can interpret these two solutions qualitatively, as follows:

a. $|a| \leq 1$: In this case, the dynamics is stable and the actual state converges to 0, independent of the unknown initial condition or marginal, in which case state is constant (but unknown). For the marginal ($a = 1$) case, we are just estimating an unknown value on the basis of N measurements. If the initial variance is σ^2, then the variance after N measurements is σ^2/N, which goes to 0 as $N \to \infty$. Thus, in this case, we have that the steady-state variance $P^* = 0$. For $|a| < 1$, we can improve the estimate to be *better* than the marginal case, because we know more and more about the position (it has to be closer and closer to 0). Thus, we effectively average N variables with decreasing variance, meaning that P^* converges to 0 even faster.

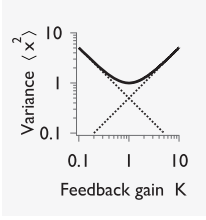
b. $|a| > 1$: In this case, the steady-state variance is finite. The reason is that there is a balance between the gain of information by having a new observation and the loss of information that occurs because the value of x_k is "blowing up" as time goes on: Indeed, $x_k = a^k x_0$. Notice that the two solutions are continuous at $|a| = 1$. In Chapter 15, we will develop the theme that unstable dynamics increase the uncertainty in the behavior of a dynamical system.

8.3 Diffusion with continuous measurements. There are subtleties:

a. Start by formulating the steady-state Kalman–Bucy filter for one-dimensional diffusion. The equations of motion are $\gamma\dot{x} = \nu_c$, with $\langle \nu_c(t)\,\nu_c(t')\rangle = 2D\gamma^2\,\delta(t-t')$. The measurement relation is $y = x + \xi_c$, with $\langle \xi_c(t)\,\xi_c(t')\rangle = \xi_c^2\,\delta(t-t')$. Find the optimal Kalman gain, and show that the variance is $P = \sqrt{2D}\,\xi_c$.

b. Contrast the above results with those found for the discrete case: for signal-to-noise ratio $\alpha = \nu^2/\xi^2$ (for power), the limit $\alpha \gg 1$ implies $L \to 1$ and $P \approx \xi^2$. By contrast, the limit $\alpha \ll 1$ implies $L \to \sqrt{\alpha}$ and $P \approx \nu\xi$. Reconcile these different behaviors. Hint: Connect the discrete quantities ν^2 and ξ^2 with our continuous versions, $\nu_c^2 = 2D$ and ξ_c^2. Notice that the units of ν_c and ξ_c are different. Write α for the discrete case in terms of continuous quantities and the time step T_s.

c. Laplace transform the equations of motion to show that the Kalman filter acts as a first-order, low-pass filter between the observations, $y(t)$, and the estimate, $\hat{x}(t)$. What is the cutoff frequency? Argue (justify) that the filter "trusts the measurements" $y(t)$ at frequencies where the signal dominates over the noise. At higher frequencies, noise is important, and the filter attenuates the measurements.



d. Add negative feedback, $u(t) = -K\,y(t)$, to stabilize the diffusing particle near the origin. Show that $\langle x^2\rangle = (K^2\xi_c^2 + \nu_c^2)/2K$. See at left for $\xi_c = \nu_c = 1$. Interpret.

Solution.

a. For the continuous case, we have $A_c = 0$, $B_c = C = 1$. Then

$$\dot{P} = \cancelto{0}{A_c^2 P} + \nu_c^2 - P^2/\xi_c^2 \qquad \to P = \nu_c\,\xi_c = \sqrt{2D}\xi_c$$
$$L = P/\xi_c^2 = \sqrt{2D}/\xi_c\,.$$

There is only one solution. (The negative solution is not physical, as the variance must be positive.) Check the units: D has units ℓ^2/t, while ξ_c^2 has units of ℓ^2/t. P thus has units of ℓ^2. L has units of $1/t$, or frequency (see below for an interpretation).

b. We write the discrete case in terms of continuous terms. Thus, $A = e^{A_c T_s} = 1$, $B_c = 1 \cdot T_s$, $C = 1$, $\nu^2 = \frac{\nu_c^2}{T_s}$, $\xi^2 = \frac{\xi_c^2}{T_s}$. We then have, for the discrete P_d and L_d,

$$P'_d = P_d + T_s^2\left(\frac{\nu_c^2}{T_s}\right) = P_d + \nu_c^2 T_s$$

$$L_d = \frac{P'_d}{P'_d + \frac{\xi_c^2}{T_s}} = \frac{P_d + \nu_c^2 T_s}{P_d + \nu_c^2 T_s + \frac{\xi_c^2}{T_s}} = \frac{L_d\frac{\xi_c^2}{T_s} + \nu_c^2 T_s}{L_d\frac{\xi_c^2}{T_s} + \nu_c^2 T_s + \frac{\nu_c^2}{T_s}} = \frac{L_d + \alpha}{L_d + \alpha + 1}$$

$$\alpha = \frac{\nu^2}{\xi^2} = \frac{\nu_c^2 T_s}{\xi_c^2/T_s} = \frac{\nu_c^2 T_s^2}{\xi_c^2}\,.$$

Thus, when we take the limit $T_s \to 0$, we automatically have $\alpha \to 0$. In this limit, we showed that $L_d \to \sqrt{\alpha}$ and

$$P_d \to \left(\frac{\xi_c^2}{T_s}\right) L_d = \left(\frac{\xi_c^2}{T_s}\right)\left(\frac{\nu_c T_s}{\xi_c}\right) = \xi_c \nu_c ,$$

which is independent of T_s. Thus, as the measurements are made with shorter and shorter T_s, the value of α approaches zero, and we are always in the "useful" limit for a Kalman filter.

c. We have

$$\begin{aligned} \frac{d\hat{x}}{dt} &= \cancelto{0}{A_c \hat{x}} + L(y - \hat{x}) \\ s\hat{X}(s) &= 0 + L[Y(s) - \hat{X}(s)] \\ \hat{X}(s) &= \left(\frac{L}{s+L}\right) Y(s) , \end{aligned}$$

so that the transfer function is the first-order low-pass filter

$$G_{yx} = \left(\frac{1}{1 + s/L}\right) ,$$

with cutoff frequency $\omega_0 = L = \sqrt{2D}/\xi_c$.

d. The equation of motion is

$$\dot{x} = -Kx - K\xi + \nu ,$$

where K has units of inverse time. One way to find the variance is to calculate the autocorrelation function, $\langle x(t)\, x(0)\rangle$ and take $t \to 0$. We drop the $_c$ subscripts for convenience. Then,

$$x(t) = e^{-Kt} \int_{-\infty}^{t} dt'\, e^{+Kt'} [-K\xi(t') + \nu(t')] ,$$

and

$$x(t)\, x(0) = e^{-Kt} \int_{-\infty}^{t} dt' \int_{-\infty}^{0} dt''\, e^{+K(t'+t'')} [-K\xi(t') + \nu(t')]\, [-K\xi(t'') + \nu(t'')] .$$

Taking the ensemble average and using the independence of the two noise sources, we have

$$\begin{aligned} \langle x(t)\, x(0)\rangle &= e^{-Kt} \int_{-\infty}^{t} dt' \int_{-\infty}^{0} dt''\, e^{+K(t'+t'')} \left(K^2\xi^2 + \nu^2\right) \delta(t' - t'') \\ &= \left(K^2\xi^2 + \nu^2\right) e^{-Kt} \int_{-\infty}^{0} dt''\, e^{2Kt''} \\ &= \left(\frac{K^2\xi^2 + \nu^2}{2K}\right) e^{-Kt} . \end{aligned}$$

Taking $t \to 0$ gives

$$\langle x^2 \rangle = \frac{K^2\xi^2 + \nu^2}{2K}$$

The $\nu^2/2K$ term represents the benefits of feedback, which reduces the variance. Since the relaxation time is K^{-1}, the variance from thermal fluctuations is reduced by the same factor.

The $K\xi^2/2$ term represents the degradation in performance because measurement noise is injected. The raw injection increases the variance by K^2, which is partially compensated for by the speed up in dynamical response.

8.4 Discretizion of a continuous stochastic system. We discretize a harmonic oscillator driven by thermal noise, following Nørrelykke and Flyvbjerg (2011).

a. Integrate the linear, time-invariant system $\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{\nu}$, with $\langle \boldsymbol{\nu}(t) \rangle = 0$ and $\langle \boldsymbol{\nu}(t)\, \boldsymbol{\nu}(t')^\mathsf{T} \rangle = \delta(t-t')$ over a time T_s to find discrete dynamics $\boldsymbol{x}_{k+1} = \boldsymbol{A}_\mathrm{d}\boldsymbol{x}_k + \boldsymbol{\nu}_k$, with $\boldsymbol{A}_\mathrm{d} = \mathrm{e}^{\boldsymbol{A}T_\mathrm{s}}$ and $\boldsymbol{\nu}_k = \int_0^{T_\mathrm{s}} \mathrm{d}t'\, \boldsymbol{\nu}(t')\, \mathrm{e}^{\boldsymbol{A}(T_\mathrm{s}-t')}\, \boldsymbol{B}$ a Gaussian random vector of mean $\boldsymbol{0}$ and covariance $\langle \boldsymbol{\nu}_k\, \boldsymbol{\nu}_\ell^\mathsf{T} \rangle = \delta_{k\ell} \int_0^{T_\mathrm{s}} \mathrm{d}t'\, \mathrm{e}^{\boldsymbol{A}(T_\mathrm{s}-t')}\, \boldsymbol{B}\boldsymbol{B}^\mathsf{T}\, \mathrm{e}^{\boldsymbol{A}^\mathsf{T}(T_\mathrm{s}-t')}$.

b. For the noisy critically damped harmonic oscillator, $\ddot{x} + 2\dot{x} + x = \sqrt{8D}\,\nu(t)$, show that $\boldsymbol{A} = \left(\begin{smallmatrix} 0 & 1 \\ -1 & -2 \end{smallmatrix}\right)$ and $\boldsymbol{A}_\mathrm{d} = \mathrm{e}^{-T_\mathrm{s}} \left(\begin{smallmatrix} 1+T_\mathrm{s} & T_\mathrm{s} \\ -T_\mathrm{s} & 1-T_\mathrm{s} \end{smallmatrix}\right)$.

c. Show that the covariance matrix $\langle \boldsymbol{\nu}_n\, \boldsymbol{\nu}_n^\mathsf{T} \rangle = \left(\begin{smallmatrix} \sigma_{xx}^2 & \sigma_{xv}^2 \\ \sigma_{xv}^2 & \sigma_{vv}^2 \end{smallmatrix}\right)$, with $\sigma_{xx}^2 = 2D[1 - \mathrm{e}^{-2T_\mathrm{s}}(1 + 2T_\mathrm{s} + 2T_\mathrm{s}^2)]$, $\sigma_{xv}^2 = 4D\,\mathrm{e}^{-2T_\mathrm{s}}\, T_\mathrm{s}^2$, and $\sigma_{vv}^2 = 2D[1 - \mathrm{e}^{-2T_\mathrm{s}}(1 - 2T_\mathrm{s} + 2T_\mathrm{s}^2)]$.

Notice that although the original physical system has only a single noise source (thermal fluctuations) that drives only the velocity, the sampled system is driven by *two* uncorrelated noise sources. The sources then become correlated by the input coupling, leading to a structure for the discrete equations that is quite different from that of the original continuous system. In the limit $T_\mathrm{s} \to 0$, we see that $\sigma_{vv}^2 = 8DT_\mathrm{s} + O(T_\mathrm{s}^2)$, while σ_{xx}^2 and σ_{xv}^2 are higher order in T_s. We then recover the continuum situation.

Solution.

a. The form for $\boldsymbol{A}_\mathrm{d}$ was previously derived as our exact solution for the initial-value problem (with $\boldsymbol{x}_n$ being the initial condition). For the random variable, the mean is zero because the integrals just multiply and add terms to an underlying Gaussian variable of zero mean. The covariance is

$$\langle \boldsymbol{\nu}_k\, \boldsymbol{\nu}_\ell^\mathsf{T} \rangle = \iint_0^{T_\mathrm{s}} \mathrm{d}t'\, \mathrm{d}t''\, \mathrm{e}^{\boldsymbol{A}(T_\mathrm{s}-t')}\, \boldsymbol{B}\, \langle \boldsymbol{\nu}(t)\, \boldsymbol{\nu}^\mathsf{T}(t') \rangle\, \boldsymbol{B}^\mathsf{T}\, \mathrm{e}^{\boldsymbol{A}^\mathsf{T}(T_\mathrm{s}-t'')}\,,$$

which gives directly the desired covariance after integrating over the delta function.

b. To evaluate the matrix exponential manually, we follow Example A.4 but take into account that the eigenvalues $(-1, -1)$ are degenerate. We have

$$\mathrm{e}^{\boldsymbol{A}T_\mathrm{s}} = \alpha_0 \mathbb{I} + \alpha_1 \boldsymbol{A} \qquad \Longrightarrow \qquad \mathrm{e}^{-T_\mathrm{s}} = \alpha_0 - T_\mathrm{s}\alpha_1\,.$$

The derivative with respect to the eigenvalue λ gives $\alpha_1 = \mathrm{e}^{-T_\mathrm{s}}$, so that $\alpha_0 = \mathrm{e}^{-T_\mathrm{s}}(1 + T_\mathrm{s})$. We then have

$$\exp\left\{\left[\begin{pmatrix} 0 & 1 \\ -1 & -2 \end{pmatrix} T_\mathrm{s}\right]\right\} = \mathrm{e}^{-T_\mathrm{s}}(1 + T_\mathrm{s})\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + T_\mathrm{s}\,\mathrm{e}^{-T_\mathrm{s}}\begin{pmatrix} 0 & 1 \\ -1 & -2 \end{pmatrix}$$
$$= \mathrm{e}^{-T_\mathrm{s}}\begin{pmatrix} 1 + T_\mathrm{s} & T_\mathrm{s} \\ -T_\mathrm{s} & 1 - T_\mathrm{s} \end{pmatrix}.$$

c. For the covariance matrix, $\boldsymbol{B} = \binom{0}{1}$ and the random increments $\boldsymbol{\nu}_n$ are given by

$$\begin{pmatrix} \nu_x \\ \nu_v \end{pmatrix}_k = \begin{pmatrix} \int_0^{T_\mathrm{s}} \mathrm{d}t\, \mathrm{e}^{T_\mathrm{s}-t}(T_\mathrm{s} - t)\,\eta(t) \\ \int_0^{T_\mathrm{s}} \mathrm{d}t\, \mathrm{e}^{T_\mathrm{s}-t}(1 - T_\mathrm{s} + t)\,\eta(t) \end{pmatrix}$$

Including the $\sqrt{8D}$ amplitude, the elements of the covariance matrix are

$$\sigma_{xx}^2 = 8D\int_0^{T_\mathrm{s}} \mathrm{d}t\, \mathrm{e}^{-2(T_\mathrm{s}-t)}(T_\mathrm{s} - t)^2 = 2D\left[1 - \mathrm{e}^{-2T_\mathrm{s}}(1 + 2T_\mathrm{s} + 2T_\mathrm{s}^2)\right]$$
$$\sigma_{xv}^2 = 8D\int_0^{T_\mathrm{s}} \mathrm{d}t\, \mathrm{e}^{-2(T_\mathrm{s}-t)}(1 - T_\mathrm{s} + t)(T_\mathrm{s} - t) = 4D\,\mathrm{e}^{-2T_\mathrm{s}}\,T_\mathrm{s}^2$$
$$\sigma_{vv}^2 = 8D\int_0^{T_\mathrm{s}} \mathrm{d}t\, \mathrm{e}^{-2(T_\mathrm{s}-t)}(1 - T_\mathrm{s} + t)^2 = 2D\left[1 - \mathrm{e}^{-2T_\mathrm{s}}(1 - 2T_\mathrm{s} + 2T_\mathrm{s}^2)\right].$$

8.5 One-step LQG. Consider one-dimensional, deterministic dynamics $x_1 = x_0 + u_0$, with cost function $J_\mathrm{det} = \frac{1}{2}(Qx_1^2 + Ru_0^2)$. Only a single control u_0 is applied.

a. Find the optimal value of u_0, and show it has the form $u_0^* = -K^* x_0$, a linear feedback. Show that $K^* = Q/(R + Q)$, and evaluate the optimal cost $J_\mathrm{det}^*(K^*)$.
b. Add a stochastic disturbance: $x_1 = x_0 + u_0 + \nu_0$, with $\nu_0 \sim \mathcal{N}(0, \nu^2)$. Show that u_0^* is unaltered and $\langle J^* \rangle = J_\mathrm{det}^* + \frac{1}{2}Q\nu^2$ (certainty equivalence principle).
c. Consider a noisy observation $y_0 = x_0 + \xi_0$. Why is the optimal control gain still K^* (separation principle)? Why is $u_0 = -Ky_0$ not optimal?
d. In formulating the cost function, you might have expected to see a total cost $J = \frac{1}{2}\sum_k \left(Qx_k^2 + Ru_k^2\right)$, with $k \in \{0, 1\}$. Why ignore the x_0 and u_1 terms?

Solution.

a. Substituting for x_1 gives a cost

$$J(u_0) = \tfrac{1}{2}Q(x_0 + u_0)^2 + \tfrac{1}{2}Ru_0^2.$$

Let us simplify the notation by dropping the 0 subscripts:

$$J(u) = \tfrac{1}{2}Q(x + u)^2 + \tfrac{1}{2}Ru^2.$$

Setting $\partial_u J = 0$ to find the minimum of $J(u)$ gives

$$\partial_u J = Q(x + u) + Ru = 0,$$

which implies,

$$u^* = -\left(\frac{Q}{R+Q}\right)x \equiv -K^* x\,, \quad J^* \equiv J^*_{\text{det}} = \tfrac{1}{2}x^2\left(\frac{QR}{R+Q}\right).$$

b. With noisy dynamics, the cost function becomes

$$\begin{aligned} J(u) &= \tfrac{1}{2}Q\,(x+u+v_0)^2 + \tfrac{1}{2}Ru^2 \\ &= \tfrac{1}{2}Q\,(x+u)^2 + \tfrac{1}{2}Ru^2 + Q\,(x+u)\,v_0 + \tfrac{1}{2}Qv_0^2 \\ &= J_{\text{det}}(u) + Q\,(x+u)\,v_0 + \tfrac{1}{2}Qv_0^2\,. \end{aligned}$$

Taking an expectation and remembering that $\langle v_0\rangle = 0$ and $\langle v_0^2\rangle = v^2$, we have

$$\langle J\rangle(u) = J_{\text{det}}(u) + \tfrac{1}{2}Qv^2\,,$$

where J_{det} is the cost function for the deterministic problem (part a). The additional cost is independent of u, implying that the optimal control u^* remains the same. This makes sense, because we do not know anything about the stochastic disturbance to bias our choice of control (i.e., it can push the system to greater or smaller values of x).

c. Now we consider noisy observations. The general strategy is to use a Kalman filter to find the best estimate of the state at time k, which we denote $\hat{x}_k$. This estimate includes the observation $y_k = x_k + \xi_k$ and requires finding the Kalman gain L_k. However, the first observation y leads to a correction in $\hat{x}_1$, not in $\hat{x} = x$, which is taken as given in the problem. The control u continues to be based on x only and will continue to have the same optimal (linear feedback) form, with the same optimal gain K^*. This is the separation principle.

Notice that we should *not* use feedback based on the noisy observation, $u = -Ky$, since the optimal gain, $K^{**} = Q/(R+Q+\xi^2)$, is then lower than $K^* = Q/(R+Q)$. Essentially, in the game of "trust the model" vs. "trust the observation" the initial condition is to trust the model, since the observation can only be noisier. However, over time, both contributions become important and the full structure of the Kalman filter is needed.

d. For this part, we restore the time-step subscripts. We ignore the $\frac{1}{2}Qx_0^2$ cost because the control cannot alter the initial state. We ignore the $\frac{1}{2}Ru_1^2$ term because its optimal value is trivially $u_1 = 0$. There is no point in applying a control at time step 1 because its costs are billed as part of J, but its benefits are not realized (since the accounting stops at $k = 1$).

8.6 Variance of observer control for a 1d Brownian particle. Observer-based feedback can lead to a minimum-variance control strategy (Section 8.2.1):

a. Find the variance $\langle x^2\rangle$ and K^* for feedback based on perfect state information, naive observations, and observer. Find L^* for the observer case.

b. Write code to simulate all three cases; check the results from (a).

Solution.

a. We derive the three expressions for state variances.

i. If the observation is *perfect*, we use the actual state in the feedback and set $u_k = -Kx_k$, which implies

$$\begin{aligned} x_{k+1} &= (1-K)x_k + \nu_k\,, \\ \langle x_{k+1}^2\rangle &= (1-K)^2\langle x_k^2\rangle + \nu^2 \qquad \text{using } \langle x_k\nu_k\rangle = 0 \\ \langle x^2\rangle &= (1-K)^2\langle x^2\rangle + \nu^2 \qquad \text{using } \langle x_k^2\rangle = \langle x_{k+1}^2\rangle = \langle x^2\rangle \\ \langle x^2\rangle &= \frac{\nu^2}{1-(1-K)^2}\,. \end{aligned}$$

ii. If the observation is *naive*, We use the measurements and set $u_k = -Ky_k = -K(x_k + \xi_k)$. To evaluate the variance, we write

$$x_{k+1} = (1-K)x_k - K\xi_k + \nu_k\,,$$

$$\left\langle x^2\right\rangle = \frac{K^2\xi^2 + \nu^2}{1-(1-K)^2}\,, \tag{8.1}$$

where we have skipped steps that are analogous to the previous derivation.

iii. If we use an *observer*, the variance is

$$\begin{aligned} x_{k+1} &= x_k + u_k + \nu_k\,, \qquad u_k = -K\hat{x}_k\,, \qquad \hat{x}_k = x_k - e_k\,, \\ &= (1-K)x_k + Ke_k + \nu_k\,. \end{aligned}$$

where e_k is the estimation error at time k. Then

$$\langle x_{k+1}^2\rangle = (1-K)^2\langle x_k^2\rangle + K^2\langle e_k^2\rangle + \nu^2 + 2K(1-K)\langle e_k\, x_k\rangle\,,$$

using the relations

$$\langle x_k\,\nu_k\rangle = \langle e_k\,\nu_k\rangle = \langle x_k\,\xi_k\rangle = 0\,.$$

From Eq. (8.21), the steady-state estimation error $\langle e^2\rangle$ is given by

$$\langle e^2\rangle = \frac{(1-L)^2\nu^2 + L^2\xi^2}{1-(1-L)^2}\,.$$

The complication is that $\langle e_k\, x_k\rangle \neq 0$. To calculate this term, we recall from Eqs. (8.11) and (8.9) that

$$\begin{aligned} e_k &= (1-L)e_k^- - L\xi_k\,, \qquad e_k^- = e_{k-1} + \nu_{k-1} \\ &= (1-L)e_{k-1} + (1-L)\nu_{k-1} - L\xi_k\,. \end{aligned}$$

Then,

$$\langle e_k\, x_k\rangle = (1-L)\langle e_{k-1}x_k\rangle + (1-L)\langle \nu_{k-1}x_k\rangle - L\cancelto{0}{\langle x_k\xi_k\rangle}.$$

Noting that

$$\langle \nu_{k-1}x_k\rangle = \langle \nu_{k-1}\,(x_{k-1} + u_{k-1} + \nu_{k-1})\rangle = \nu^2\,,$$

and

$$\begin{aligned}\langle e_{k-1}x_k\rangle &= \langle e_{k-1}\Big(x_{k-1}+u_{k-1}+\cancelto{0}{y_{k-1}}\Big)\rangle\\ &= \langle e_{k-1}x_{k-1}\rangle - K\langle e_{k-1}\hat{x}_{k-1}\rangle\\ &= \langle e_{k-1}x_{k-1}\rangle - K\langle e_{k-1}(x_{k-1}-e_{k-1})\rangle\\ &= (1-K)\langle e_{k-1}x_{k-1}\rangle + K\langle e_{k-1}^2\rangle\,,\end{aligned}$$

we have

$$\langle e_k\, x_k\rangle = (1-L)(1-K)\langle e_{k-1}x_{k-1}\rangle + (1-L)K\langle e_{k-1}^2\rangle + (1-L)\nu^2\,.$$

Thus, in steady state, we have the two relations (since $\langle e^2\rangle$ is known),

$$\begin{aligned}\langle x^2\rangle &= (1-K)^2\langle x^2\rangle + K^2\langle e^2\rangle + \nu^2 + 2K(1-K)\langle e\, x\rangle\\ \langle e\, x\rangle &= (1-L)(1-K)\langle e\, x\rangle + (1-L)\Big(K\langle e^2\rangle + \nu^2\Big)\,.\end{aligned}$$

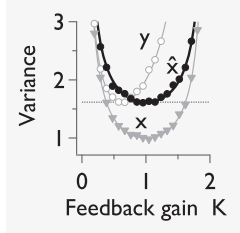
We can solve first for $\langle e\, x\rangle$ and then for $\langle x^2\rangle$, which gives a complicated function of K and L, in terms of ν^2 and ξ^2. Then, either numerically or using a symbolic-algebra program, we minimize with respect to K and L, to find $K^* = 1$ and $L^* = \frac{1}{2}(\sqrt{5}-1) \approx 0.62$. The minimum cost is just $J = \langle x^2\rangle$, since $R = 0$.

As a side note, one can repeat the calculation adding a control-effort cost, $R > 0$. The cost function is then expressed as

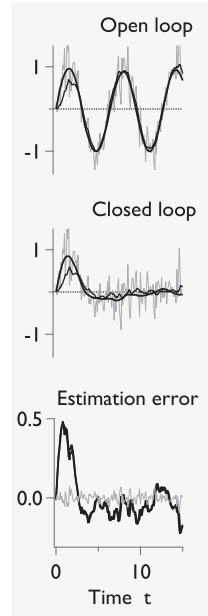
$$\begin{aligned}J &= \langle x^2\rangle + R\langle u^2\rangle\\ &= \langle x^2\rangle + RK^2\langle \hat{x}^2\rangle\\ &= \langle x^2\rangle + RK^2(\langle x^2\rangle + \langle e^2\rangle - 2\langle e\, x\rangle)\\ &= \tfrac{1}{2}\Big(\sqrt{5} + \sqrt{1+4R}\Big)\,.\end{aligned}$$

In the explicit expression, $\nu^2 = \xi^2 = 1$ and $L = L^* = \frac{1}{2}(\sqrt{5}-1)$. For $R = 0$, the minimum average cost per time step is $J = \frac{1}{2}(\sqrt{5}+1) \approx 1.62$. For $R = 1$, it is $J = \sqrt{5} \approx 2.24$. For $R \gg 1$, we have the asymptotic scaling $J \sim \sqrt{R}$.

b. See code on the book website. Below is a representative plot, using 10^4 time steps per simulation (i.e., per marker). You should find good agreement with the predicted curves.



8.7 LQG for undamped, noisy oscillator. Let $\ddot{x} + x = u(t) + \nu(t)$, with $y(t) = x(t) + \xi(t)$.



a. What is the state-space representation of the continuous system?
b. For sampling at $T_s = 0.1$, what is the ZOH discrete state-space representation?
c. Assuming that the standard deviations $\nu = \xi = 0.3$ for the process and measurement noise (when sampled at intervals T_s) and assuming state-space weights of $\boldsymbol{Q} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ and input weighting $r = 0.1$, derive the LQG controller. As a check, you should find an optimal observer (Kalman) gain of $\boldsymbol{L}^{\mathsf{T}} \approx (0.09, 0.03)$ and optimal control gain of $\boldsymbol{K} \approx (1.7, 3.3)$. (Give some more digits, please!)
d. Using the above observer and controller gains and adding process and measurement noise, plot the disturbance response (right, top graph, for $x(0) = 0$, $\dot{x}(0) = 1$).
e. Simulate the controller and plot the disturbance response (right, middle graph). Show that the difference between the position and its optimal (Kalman) estimate is the same for both the closed- and open-loop cases (dark lines in the bottom plot at right). Their difference is nearly zero, as shown by the gray trace.
f. Show that, after a transient, the standard deviation of both state and estimate are well below that of the measurement errors.

Solution.

a. The undamped oscillator obeys $\ddot{x} + x = u(t) + \nu(t)$, with $y(t) = x(t) + \xi(t)$, which corresponds to a two-input, one-output state-space system with

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}, \quad \boldsymbol{C} = \begin{pmatrix} 1 & 0 \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} 0 \end{pmatrix}.$$

b. If we sample at $T_s = 0.1$, the ZOH discrete state-space representation is, to three figures,

$$\boldsymbol{A} = \begin{pmatrix} 0.995 & 0.0998 \\ -0.0998 & 0.995 \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} 0.00500 & 0.00500 \\ 0.0998 & 0.0998 \end{pmatrix}, \quad \boldsymbol{C} = \begin{pmatrix} 1 & 0 \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} 0 \end{pmatrix}.$$

c. For process-noise covariance $\boldsymbol{Q} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ and input weighting $r = 0.1$, the LQG controller has optimal feedback gain

$$\boldsymbol{K} = \begin{pmatrix} 1.741 & 3.261 \end{pmatrix}$$

and optimal observer gain

$$\boldsymbol{L} = \begin{pmatrix} 0.0905 \\ 0.0306 \end{pmatrix}$$

d. See plot in text. Note that we use the same noise realization for open and closed loop, to facilitate comparison between the two cases.
e. See plot in text.

f. The standard deviations of x and $\hat{x}$ are, after transients have died away, given by 0.11 and 0.08, respectively. Both are well below the standard deviation of measurement errors, $\xi = 0.3$. Thus, using the averaging property of the Kalman filter (accounting for the dynamics), we can have less error than the naive measurement error.

8.8 **Delayed choice**. Fill in the details from Example 8.5. Reproduce the plots in the example, using $R = 2D = x_0 = 1$.

Solution.

To summarize the problem from Example 8.5, the equations of motion are

$$\dot{x} = u + \nu\,, \qquad x(0) = 0\,, \qquad x(\tau) = \pm 1\,,$$

with control input $u(t)$ and Gaussian white noise, $\langle \nu(t)\,\nu(t')\rangle = 2D\,\delta\,(t - t')$. The goal is to minimize control effort, with running cost $L = \frac{1}{2}Ru^2$ and cost-to-go $J(x, u, t) = \int_t^\tau \mathrm{d}t'\, L[u(t')]$. The HJB equation is then

$$\partial_t J^*(x,t) + \inf_u \left[\tfrac{1}{2}Ru^2 + (\partial_x J^*)\; u + D\partial_{xx} J^*\right] = 0\,.$$

The first step is, at each time t, to minimize over u. Taking a derivative with respect to u in the HJB equation and remembering that $J^*(x, t)$ is independent of u gives

$$Ru + (\partial_x J^*) = 0 \qquad \Longrightarrow \qquad u^* = -R^{-1}(\partial_x J^*)\,.$$

Substituting $u = u^*$ into the HJB equation then leads to the nonlinear PDE

$$\partial_t J^*(x,t) - \tfrac{1}{2}R^{-1}\,(\partial_x J^*)^2 + D\partial_{xx} J^* = 0\,.$$

To change variables, we set $J^*(x, t) = -\lambda \log \psi(x, t)$. The derivatives are

$$\partial_x J^*(x,t) = -\frac{\lambda}{\psi}\,\partial_x\psi\,, \qquad \partial_t J^*(x,t) = -\frac{\lambda}{\psi}\,\partial_t\psi\,,$$
$$\partial_{xx} J^*(x,t) = \frac{\lambda}{\psi^2}\,(\partial_x\psi)^2 - \frac{\lambda}{\psi}\,\partial_{xx}\psi\,.$$

The HJB then becomes

$$-\frac{\lambda}{\psi}\,\partial_t\psi - \frac{1}{2R}\frac{\lambda^2}{\psi^2}(\partial_x\psi)^2 + D\left[\frac{\lambda}{\psi^2}\,(\partial_x\psi)^2 - \frac{\lambda}{\psi}\,\partial_{xx}\psi\right] = 0\,.$$
$$-\partial_t\psi - \frac{\lambda}{2R}\frac{1}{\psi}(\partial_x\psi)^2 + D\frac{1}{\psi}\,(\partial_x\psi)^2 = D\partial_{xx}\psi\,.$$

If we set the constant $\lambda = R\nu^2$, then the quadratic terms cancel, leaving

$$-\partial_t\psi = D\,\partial_{xx}\psi\,,$$

a diffusion equation in negative time.

The condition at the final time τ is that the particle must be at $x = \pm x_0$. This means that the cost should be infinite for $x \neq \pm x_0$. We can impose this condition by setting $\psi = 0$ for $x \neq \pm x_0$ and non-zero at $x = \pm x_0$. Then,

$$\psi(x,\tau) = \frac{1}{2}\left[\delta\left(x - x_0\right) + \delta\left(x + x_0\right)\right] .$$

We should not worry too much about the infinities produced by the delta functions, in that we can relax the target size to an interval $\pm\epsilon$ about $+1$ or -1 and take the limit $\epsilon \to 0$. Note the normalization: $\int_x \mathrm{d}x\, \psi(x,\tau) = 1$.

Delta-function initial conditions lead to the Green's function solution of the diffusion equation (going backwards in time, because of the $-\partial_t$ term). With $t' = \tau - t$, we have

$$\begin{aligned}
\psi(x,t') &= \left(\frac{1}{2\sqrt{4\pi D t'}}\right)\left[\exp\left(-\frac{(x-x_0)^2}{4Dt'}\right) + \exp\left(-\frac{(x+x_0)^2}{4Dt'}\right)\right] \\
&= \left(\frac{1}{2\sqrt{4\pi D t'}}\right)\exp\left(-\frac{x_0^2}{4Dt'}\right)\exp\left(-\frac{x^2}{4Dt'}\right)\left[\exp\left(\frac{2x\,x_0}{4Dt'}\right) + \exp\left(-\frac{2x\,x_0}{4Dt'}\right)\right] \\
&= \left(\frac{\exp\left(-\frac{x_0^2}{4Dt'}\right)}{\sqrt{4\pi D t'}}\right)\exp\left(-\frac{x^2}{4Dt'}\right)\cosh\left(\frac{x\,x_0}{2Dt'}\right) .
\end{aligned}$$

Next, we transform back to $J^* = -R\nu^2 \ln\psi$, giving

$$J^*(x,t') = R\nu^2\left[\frac{x^2}{4Dt'} - \ln\cosh\left(\frac{x\,x_0}{2Dt'}\right)\right] + f(t') ,$$

with $f(t') = R\nu^2(\frac{x_0^2}{4Dt'} + \frac{1}{2}\ln 4\pi t')$. The important point is that the optimal control $u^* = -R^{-1}\partial_x J^*$ is not affected by $f(t')$. Differentiating J^* with respect to x gives

$$u^*(x,t') = \left(\frac{x_0}{t'}\right)\left(\tanh\frac{x\,x_0}{2Dt'} - \frac{x}{x_0}\right) .$$

Taking $x_0 = 1$ gives the formulas in the main text. The symmetry-breaking transition condition, in dimensional form, is

$$\nu^2 t' = x_0^2 , \qquad \Longrightarrow \qquad t'_{\mathrm{c}} = \frac{x_0^2}{\nu^2} .$$

For numerical analysis, we set $R = \nu^2 = 2D = 1$ and drop $f(\tau)$. Thus,

$$\begin{aligned}
J^*(x,t') &= \frac{x^2}{2t'} - \ln\cosh\left(\frac{x}{t'}\right) \\
u^*(x,t') &= \left(\frac{1}{t'}\right)\left(\tanh\frac{x}{t'} - x\right) .
\end{aligned}$$

Finally, numerical trajectories can be generated by discretizing:

$$x_{k+1} = x_k + \Delta t\, u(x_k, t'_k) + \nu_k ,$$

where $\langle \nu_k\, \nu_\ell \rangle = \delta_{k\ell}\,(2D\Delta t)$.

8.9 **Toy model of state estimation, part 2**. Redo the calculations of Section 8.3.1, allowing for N noisy, independent measurements $y_i = x + \xi_i$. Find the posterior $p(x|y^N)$, where $y^N = \{y_i\}$ for $i = 1, \ldots, N$. Show that $\hat{x} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\xi^2/N}\,\bar{y}$, where $\bar{y} = \frac{1}{N}\sum_i y_i$ is the average of the N measurements. Find the corresponding variance.

Solution.

Following the calculations in Eq. (8.66) and using the independence of each measurements, we have

$$\begin{aligned}
p(x|y^N) &\propto p(y^N|x)\, p(x) \\
&= \left(\prod_{i=1}^{N} p(y_i|x)\right) p(x) \\
&= \left(\prod_{i=1}^{N} \mathcal{N}(y_i - x, \sigma_\xi^2)\right) \mathcal{N}(0, \sigma_x^2) \\
&\propto \prod_{i=1}^{N} \exp\left[-\frac{(y_i - x)^2}{2\sigma_\xi^2}\right] \exp\left[-\frac{x^2}{2\sigma_x^2}\right]. \\
&= \exp\left[-\sum_{i=1}^{N} \frac{(y_i - x)^2}{2\sigma_\xi^2}\right] \exp\left[-\frac{x^2}{2\sigma_x^2}\right]. \\
&\propto \exp\left[-\frac{\left(x - \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\xi^2/N}\,\bar{y}\right)^2}{2\sigma_0^2}\right],
\end{aligned}$$

where $\bar{y}$ is the arithmetic average

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

and where the variance of the posterior is

$$\frac{1}{\sigma_0^2} = \frac{N}{\sigma_\xi^2} + \frac{1}{\sigma_x^2}.$$

One way of interpreting the result is to define an effective signal-to-noise ratio

$$\mathrm{SNR}_{\mathrm{eff}}^2 = \frac{N\sigma_x^2}{\sigma_\xi^2},$$

in terms of which

$$\hat{x} = \left(\frac{\mathrm{SNR}_{\mathrm{eff}}^2}{\mathrm{SNR}_{\mathrm{eff}}^2 + 1}\right)\bar{y}$$

and

$$\sigma_0^2 = \frac{\sigma_x^2}{1 + \mathrm{SNR}_{\mathrm{eff}}^2}.$$

For large $N \gg 1$,

$$\hat{x} \to \bar{y}\,, \qquad \sigma_0^2 \to \frac{\sigma_\xi^2}{N}\,.$$

This is the usual formula for a measurement. We notice that it is independent of the prior estimate for x (equal to 0) and of σ_x^2. In other words, with enough measurements, the prior is "overwhelmed" by the data.

8.10 From Bayes to Kalman, in 1d. Equations (8.76) and (8.77) describe one-dimensional diffusion, $x_{k+1} = x_k + \nu_k$, with $y_k = x_k + \xi_k$, with i.i.d. Gaussian random noise terms $p(\nu_k) = \mathcal{N}(\nu_k; 0, \nu^2)$ and $p(\xi_k) = \mathcal{N}(\xi_k; 0, \xi^2)$.

a. Show that the PDF of $p(x_{k+1}|x_k)$ is $\mathcal{N}(x_{k+1} - x_k; 0, \nu^2) \equiv \frac{1}{\nu\sqrt{2\pi}} \exp\left[-\frac{(x_{k+1}-x_k)^2}{2\nu^2}\right]$.

b. Making the *ansatz* $p(x_k|y^k) = \mathcal{N}(x_k; \hat{x}_k, P_k)$ and marginalizing over the proper variable, show that $p(x_{k+1}|y^k) = \mathcal{N}(x_{k+1}; \hat{x}_k, P_k + \nu^2) \equiv \mathcal{N}(x_{k+1}; \hat{x}_{k+1}^-, P_{k+1}^-)$.

c. The Bayesian update step, $p(x_{k+1}|y^{k+1}) = \frac{p(y_{k+1}|x_{k+1})\, p(x_{k+1}|y^k)}{p(y_{k+1}|y^k)}$ requires three probability distributions. We know one. Derive the other two:

$$p(y_{k+1}|x_{k+1}) = \mathcal{N}(y_{k+1} - x_{k+1}; 0, \xi^2)\,, \quad p(y_{k+1}|y^k) = \mathcal{N}(y_{k+1}; \hat{x}_k, P_k + \nu^2 + \xi^2)\,.$$

d. Finally, evaluate the update step using the three distributions to show that the conditional distribution $p(x_{k+1}|y^{k+1}) = \mathcal{N}(x_{k+1}; \hat{x}_{k+1}, P_{k+1})$, where

$$\hat{x}_{k+1} = \hat{x}_k + L_{k+1}(y_{k+1} - \hat{x}_k) \qquad L_{k+1} = \frac{P_k + \nu^2}{P_k + \nu^2 + \xi^2}$$
$$P_{k+1}^{-1} = (P_k + \nu^2)^{-1} + (\xi^2)^{-1} \qquad \text{or} \qquad P_{k+1} = L_{k+1}\,\xi^2\,,$$

Hint: Complete the square or use a computer-algebra program.

Solution.

a. We use Eq. (8.77a) to write

$$p(x_{k+1}|x_k) = \mathcal{N}(x_{k+1} - x_k; 0, \nu^2)\,,$$

since $p(\nu_k) = \mathcal{N}(\nu_k; 0, \nu^2)$.

b.

$$\begin{aligned}
p(x_{k+1}|y^k) &= \int \mathrm{d}x_k\, p(x_{k+1}|x_k, y^k) p(x_k|y^k) && \text{marginalization} \\
&= \int \mathrm{d}x_k\, p(x_{k+1}|x_k) p(x_k|y^k) && \text{Markov} \\
&= \int_{-\infty}^{\infty} \mathrm{d}x_k\, \mathcal{N}(x_{k+1} - x_k; 0, \nu^2)\, \mathcal{N}(x_k; \hat{x}_k, P_k) && \text{Eq. (8.77}b\text{)} \\
&= \mathcal{N}(x_{k+1}; \hat{x}_k, P_k + \nu^2) && \text{sum of Gauss vars}\,.
\end{aligned}$$

The last identity uses Eq. (A.166) for the sum of Gaussian random variables, applied to the case $(x_{k+1} - x_k) + x_k = x_{k+1}$. The means and variances thus add.

c. From Eq. (8.77b), we have

$$p(y_{k+1}|x_{k+1}) = \mathcal{N}[(y_{k+1} - x_{k+1}); 0, \xi^2]\,,$$

using $p(\xi_{k+1}) = \mathcal{N}(\xi_{k+1}; 0, \xi^2)$.
We then have

$$\begin{aligned}
p(y_{k+1}|y^k) &= \int \mathrm{d}x_{k+1}\, p(y_{k+1}, x_{k+1}|y^k) && \text{marginalization}\\
&= \int \mathrm{d}x_{k+1}\, p(y_{k+1}|x_{k+1}, y^k)\, p(x_{k+1}|y^k) && \text{conditional prob.}\\
&= \int \mathrm{d}x_{k+1}\, p(y_{k+1}|x_{k+1})\, p(x_{k+1}|y^k) && \text{Markov}\\
&= \int \mathrm{d}x_{k+1}\, \mathcal{N}[(y_{k+1} - x_{k+1}); 0, \xi^2]\, \mathcal{N}(x_{k+1}; \hat{x}_k, P_k + \nu^2) && \text{previous results}\\
&= \mathcal{N}(y_{k+1}; \hat{x}_k, P_k + \nu^2 + \xi^2) && \text{sum of Gaussian vars}\,.
\end{aligned}$$

d. We have to show

$$\begin{aligned}
p(x_{k+1}|y^{k+1}) &= \frac{p(y_{k+1}|x_{k+1})\, p(x_{k+1}|y^k)}{p(y_{k+1}|y^k)}\\
&= \frac{\mathcal{N}(y_{k+1}; x_{k+1}, \xi^2)\, \mathcal{N}(x_{k+1}; \hat{x}_k, P_k + \nu^2)}{\mathcal{N}(y_{k+1}; \hat{x}_k, P_k + \nu^2 + \xi^2)}\\
&= \mathcal{N}(x_{k+1}; \hat{x}_{k+1}, P_{k+1})\,,
\end{aligned}$$

where

$$\hat{x}_{k+1} = \hat{x}^-_{k+1} + L_{k+1}(y_{k+1} - \hat{x}^-_{k+1}) \qquad L_{k+1} = \frac{P_k + \nu^2}{P_k + \nu^2 + \xi^2}$$
$$P^{-1}_{k+1} = (P_k + \nu^2)^{-1} + (\xi^2)^{-1} \qquad \text{or } P_{k+1} = L_{k+1}\,\xi_{k+1}\,,$$

The key step, in multiplying and dividing the Gaussians, is to show that

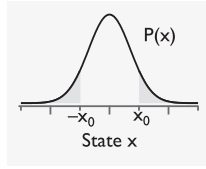
$$\frac{(y_{k+1} - x_{k+1})^2}{\xi^2} + \frac{(x_{k+1} - \hat{x}_k)^2}{P_k + \nu^2} - \frac{(y_{k+1} - \hat{x}_k)^2}{P_k + \nu^2 + \xi^2} = \frac{(x_{k+1} - \hat{x}_{k+1})^2}{P_{k+1}}\,,$$

where $\hat{x}^-_{k+1} = \hat{x}_k$ (since the dynamics are trivial) and $P^-_{k+1} = P_k + \nu^2$. The algebra is tedious, and computer-algebra software is helpful.

8.11 Instability with a hard-spring potential. Consider an overdamped particle in a hard-spring potential with input noise. If we ignore the nonlinear term, a quick analysis shows that the system is stable. Now add a nonlinear term that, in the absence of stochasticity, is stabilizing. Surprisingly, the state amplitude will eventually diverge, no matter how weak the noise. As shown at left, along with a normal quadratic potential (thin line), the *hard-spring potential* $V(x) = \frac{1}{2}x^2 + \frac{1}{4}x^4$, with force $F(x) = -\partial_x V = -(x + x^3)$. The spring is "hard" because the

local stiffness $\mathrm{d}F/\mathrm{d}x$ increases away from equilibrium. In a first-order numerical integration, the system state evolves as $x_{k+1} = f(x_k) = x_k - T_s(x_k + x_k^3) + \nu_k$, with $\nu_k \sim \mathcal{N}(0, \nu^2)$.

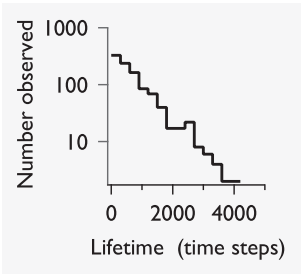
a. Simulate the above equation. Generate various times series for x_k for different values of ν_k, using $T_s = 0.5$. What happens as you increase the input noise?
b. Show that the linearized system is stable for $0 < T_s < 2$, for arbitrary ν.
c. Fix $\nu = 0.5$, and modify your code to track the *lifetime* of the state, the typical time before x_k goes unstable. Run your code many times (≈ 1000), measuring the lifetime in each case. Plot a histogram. What distribution does it follow, and why? You should find that the average lifetime ≈ 700.
d. Show that the state dynamics go unstable for $|x_k| > x_0 = \sqrt{3}$. Hint: Look at the conditions $f(x) > x$ and $f(x) < -x$. Why do these lead to instability?
e. To estimate the average lifetime, assume $x_k \sim \mathcal{N}(0, \nu^2)$, which is only approximately true because of the x_k^3 term. Assume, too, that each time step brings an independent perturbation, which implies that perturbations relax in one time step. The probability of instability then reduces to the probability $P(|x| > x_0)$ for a Gaussian distribution. If, in either tail of the distribution, a point lies in the shaded area $|x| > x_0$, then instability will likely result (see right). ("Likely" because a fluctuation just larger than x_0 might come back.) Then, either numerically or by an asymptotic expansion of the erfc$(\cdot)$ function, estimate the lifetime of the state x, in units of the time step k. You should find $\approx$ 1000–2000, slightly > 700.



f. For the linearization $x_{k+1} = \frac{1}{2}x_k + \nu_k$, show that the variance of x_k is $\frac{4}{3}\nu^2$. Compare to the numerically estimated variance for the full nonlinear equation.

Solution.

a. Graphs should show instabilities after a finite time, depending on the level of noise.
b. The linear equation is $x_{k+1} = (1 - T_s)x_k + \nu_k$. For the deterministic equation $x_{k+1} = (1 - T_s)x_k$, the fixed point is $x_k = 0$. The linear stability is determined by $f'(x = 0) = (1 - T_s)$ and must be between -1 and 1 (see, for example, Figure 5.9), which implies $0 < T_s < 2$. Because the equations are linear, it is then stable to *any* amplitude of perturbation.
c. You should see an exponential distribution resembling the one below. For $n = 1000$ runs, I get an average lifetime = 717. Since the distribution is exponential, the standard deviation is similar (696 in my case). The standard error of the mean then should be about $700/\sqrt{1000} \approx 22$. An exponential distribution implies that the probability of instability at each time step is independent of the previous time step. This independence leads to an exponential distribution for the interval, which characterizes a Poisson process (same argument as radioactive decay).



d. We have $x_{k+1} = f(x_k)$, for $f(x) = \frac{1}{2}(x - x^3)$. We have instability if $f(x) > x$ or $f(x) < -x$. The former leads to a directly growing instability, while the latter to an oscillatory instability.
   i. $f(x) > x \implies \frac{1}{2}(x - x^3) > x \implies -x^3 > x \implies -x^2 > 1$, which is impossible. So there is no instability in this case.
   ii. $f(x) < -x \implies \frac{1}{2}(x - x^3) < -x \implies -x^3 < -3x \implies x^2 > 3$. In other words, there is an oscillatory instability if $|x_0| > \sqrt{3}$.

   Note that the condition we impose is related to, but different from, the condition for linear stability of a fixed point x^*. The latter is defined by $x^* = f(x^*)$ and is linearly unstable if $|f'(x^*)| > 1$. The condition in this problem holds for nonlinear equations. The only fixed point in this problem is $x^* = 0$ and it is linearly stable, since $f'(0) = \frac{1}{2}$.

e. Now that we know the typical size of the x_k fluctuations, let us estimate the probability of an x_k fluctuation with $|x_k| \geq x_0$. We assume $x_k \sim \mathcal{N}(0, v^2)$ and estimate $P \equiv P(|x_k| > x_0)$. Then the typical number of time steps to instability will be simply $N_x \approx 1/P$.

   We thus need to evaluate the area in the two shaded tails in the figure. The cumulative distribution function

$$F(x_0') = \frac{2}{\sqrt{2\pi}} \int_{-\infty}^{x_0'} \mathrm{d}x\, \mathrm{e}^{-\frac{x^2}{2}} = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x_0'}{\sqrt{2}}\right)\right],$$

   with $x_0' \equiv x_0/v$, gives the area from $-\infty$ to x_0'. The area under one tail is thus $1 - F(x_0')$, and the area under both is $2(1 - F(x_0')) = 1 - \mathrm{erf}(x_0'/\sqrt{2}) \equiv \mathrm{erfc}(x_0'/\sqrt{2})$. Since the asymptotic expansion of $\mathrm{erfc}(x)$ is $\mathrm{erfc}(x) \sim \frac{\exp(-x^2)}{x\sqrt{\pi}}$, the typical lifetime N_x will be, dropping $O(1)$ numerical factors,

$$N_x = \mathrm{erfc}\left(\frac{x_0'}{\sqrt{2}}\right)^{-1} \sim x_0' \exp\left(\frac{x_0'^2}{2}\right).$$

   The erfc function is more accurate, but the asymptotic expansion is useful in driving home the point that the lifetime is a very sensitive function of x_0/v. Using the erfc function for $x_0' \approx \sqrt{3}/(0.5)$, I get about 1880. This is reasonably close to the observed lifetime of 717, as the estimate is extremely sensitive to the details. For example, substituting the observed variance of x_k of 0.513 leads to $N_x \approx 1360$. The slight non-Gaussian nature of the fluctuations accounts for the rest of the discrepancy.

f. Assuming that there is no instability and no nonlinearity, we calculate the typical value of x_k. We do so by writing

$$x_{k+1}^2 = \tfrac{1}{4}x_k^2 + x_k\,\nu_k + \nu_k^2\,.$$

Averaging this equation and defining $\sigma^2 = \langle x_k^2\rangle = \langle x_{k+1}^2\rangle$ (stationary stochastic process), we have

$$\sigma^2 = \tfrac{1}{4}\sigma^2 + 0 + \nu^2\,,$$

since x_k and ν_k are uncorrelated. Thus, $\sigma^2 = \frac{4}{3}\nu^2$. With $\nu = 0.5$, we have $\sigma \approx 0.58$. This is smaller than the observed fluctuations of 0.513, with the difference due to the extra restoring force at high x_k.

8.12 Conditioning matters! For the example in Figure 8.6, discuss and contrast

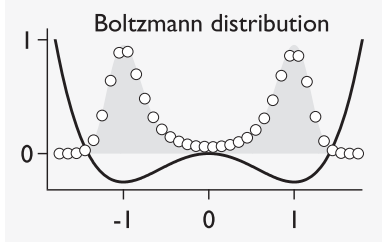
$$p(x_k)\,,\quad p(x_k|y_k)\,,\quad p(x_k|y^k)\,,\quad p(x_k|y^{k-1})\,.$$

Solution.

a. $p(x_k)$. This is the distribution of states independent of any observations. In the absence of other knowledge, we can solve the Fokker-Planck equation for the continuum equivalent (which gives the same answer as for the discrete case, as time plays no role), to find a Boltzmann distribution of the form

$$p(x_k) \propto \mathrm{e}^{-V(x_k)/\nu^2}\,,\qquad V(x_k) = -\tfrac{1}{2}x_k^2 + \tfrac{1}{4}x_k^4\,.$$

Another way to get such a distribution is to plot a long time series of states x_k and histogram them. The figure below represents 100 000 samples, and the solid line plots the Boltzmann distribution on top. The distribution has two maxima at ±1 and is larger in between than outside, as we expect.



b. $p(x_k|y_k)$. This is the probability that a given observation y_k was caused by a state x_k. In the absence of any other prior information concerning x_k, Bayes' Theorem implies that this probability is just $p(y_k|x_k) = \mathcal{N}(y_k - x_k; 0, \xi^2)$, a Gaussian with mean y_k.

c. $p(x_k|y^k)$. This is the optimal estimate that was the focus of our analysis. For one particular set of observations, the set of probability distributions is thus given in Figure 8.6. Qualitatively, the possibilities range from near certainty that the particle is in one well to bewilderment, with an equal-sized peak above each well.

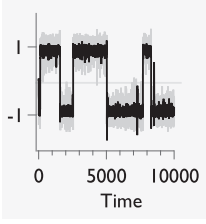
d. $p(x_k|y^{k-1})$. This is the prediction given the best estimate. With our potential, we see, again from Figure 8.6, that there is a local "flow" of probability "downhill" towards the closest local minimum of the potential. The evolution is somewhat like that of a (thick) liquid under gravity that oozes down to the local hill gradient.

8.13 Extended Kalman filter (EKF). Code the EKF algorithm for one-dimensional dynamics such as Examples 8.2 and 8.3. Explore in particular the double-well potential. Show, for example, that the EKF fails for the parameters in Figure 8.4 but succeeds for smaller time steps (smaller a), even when the thermal noise ν is increased. Explain, and confirm for $a = 0.7$, $\nu = 0.2$, and $\xi = 0.5$.

Solution.

See book website for EKF code. The value $a = 0.8$ is close to the "critical" value separating the regime where the EKF tracks the hops across the barrier. At smaller values of a (e.g., 0.7), the EKF can track hops. To make sure there are hops, we may need to increase ν. However, increasing ν too much leads to instability, as discussed in Problem 8.11. Thus, the parameters given are chosen with some care. But the general point remains valid: discretizing a continuous nonlinear system with a faster sampling frequency will make the discrete dynamics more linear and will improve the outcomes of the EKF.

8.14 Cubature Kalman filter (CKF). Code the CKF algorithm for one-dimensional dynamics such as Examples 8.2 and 8.3. Explore in particular the double-well potential.



a. Show that the $n = 1$ CKF algorithm matches the variance to $[\frac{1}{2}f''(0) + \frac{2}{3}f'(0)\,f'''(0)]$.
b. Plot true and estimated states for the CKF (see text and left).
c. Show that pushing out the two cubature points by a factor a can improve the CKF. In particular, find a value for a that works as well as the plot at right. The true state is shown in gray and the CKF estimate in black.

Solution.

a. See Mathematica notebook. As an intermediate checkpoint, you should find that the variance of the cubature-point approximation is

$$\frac{1}{2}\left[f(x+1)^2 + f(x-1)^2\right] - \left[\frac{1}{2}\left[f(x+1) + f(x-1)\right]\right]^2 = \left[f'(0) + \tfrac{1}{6}f'''(0)\right]^2 + \cdots .$$

b. See book website for CKF code.
c. The plot in the book uses $a = \sqrt{2}$. The same noise realization is used.

8.15 Unscented transform (UT). An alternative to the cubature Kalman filter (CKF) is the unscented Kalman filter (UKF), which is based on the "unscented" transform. It resembles the CKF but with different weights and an extra element. In particular, for a smooth function $y = f(x)$, with $x \sim \mathcal{N}(\mu, \sigma^2)$, there are

three ensemble elements (*sigma points*), $x_0 = \mu$ and $x_{\pm1} = \mu \pm a\sigma$ and weights w_0 and $w_{\pm1}$.

a. Show that $a = \sqrt{3}$, $w_0 = \frac{2}{3}$, and $w_{\pm1} = \frac{1}{6}$ matches the mean to fourth order.
b. Show that the variance is matched exactly to first order and partly to second order.
c. Find the mean $\bar{y}$ and variance $\overline{P}_y$ for $y = x^2$ and for $y = x^4$, assuming $x \sim \mathcal{N}(0, 1)$.

Solution.

Let us Taylor expand $f(x)$ about $x = \mu$. With $\delta = x - \mu$ and $f_n \equiv \left.\frac{\mathrm{d}^n f}{\mathrm{d}x^n}\right|_{x=\mu}$,

$$y = f(x) = f_0 + f_1\,\delta + \tfrac{1}{2}f_2\,\delta^2 + \tfrac{1}{6}f_3\,\delta^3 + \tfrac{1}{24}f_4\,\delta^4 + O(f_5)\,.$$

The moments are the same as for a standard Gaussian. The odd moments are zero, and the even moments are

$$\left\langle\delta^2\right\rangle = \sigma^2\,, \qquad \left\langle\delta^4\right\rangle = 3\sigma^4\,, \qquad \left\langle\delta^6\right\rangle = 15\sigma^6\,, \qquad \left\langle\delta^8\right\rangle = 105\sigma^8\,.$$

a. Using these moments, we can evaluate the estimate for the mean:

$$\langle y\rangle = f_0 + \tfrac{1}{2}f_2\sigma^2 + \tfrac{1}{8}f_4\sigma^4 + O(f_5\sigma^5)\,.$$

The unscented transform (UT) for the mean uses three sigma points:

$$x_0 = \mu\,, \qquad x_{\pm1} = \mu \pm a\,\sigma\,,$$

and weights w_0 and $w_{\pm1}$. By symmetry, $w_1 = w_{-1}$. Thus,

$$\begin{aligned}\bar{y} &\equiv \sum_{i=-1}^{+1} w_i\,y_i = w_0 f_0 + w_1(y_1 + y_{\pm1})\\ &= w_0 f_0 + w_1\left[2f_0 + f_2(a\sigma)^2 + \frac{1}{12}f_4(a\sigma)^4\right].\end{aligned}$$

Matching f_0, f_2, and f_4 coefficients then gives

$$\underbrace{1 = w_0 + 2w_1}_{f_0}\,, \qquad \underbrace{\frac{1}{2} = w_1 a^2}_{f_2}\,, \qquad \underbrace{\frac{1}{8} = \frac{1}{12}w_1 a^4}_{f_4}\,.$$

It is easy to verify that $a^2 = 3$, $w_0 = \frac{2}{3}$, and $w_1 = \frac{1}{6}$ solves these equations.

b. The corresponding variance equations are best calculated using a computer-algebra program. Writing P_y for the variance of y, we first calculate the expectation values.

$$\begin{aligned}y^2 = f_0^2 &+ 2f_0f_1\delta + (f_1^2 + f_0f_2)\delta^2 + (f_1f_2 + \tfrac{1}{3}f_0f_3)\delta^3 + (\tfrac{1}{4}f_x^2 + \tfrac{1}{3}f_1f_3 + \tfrac{1}{12}f_0f_4)\delta^4\\ &+ \tfrac{1}{12}f_1f_4\delta^5 + (\tfrac{1}{36}f_3^2 + \tfrac{1}{24}f_2f_4)\delta^6 + \tfrac{1}{72}f_3f_4\delta^7 + \tfrac{1}{576}f_4^2\delta^8\,.\end{aligned}$$

Taking expectations and using the moment formulas for $\langle \delta^n \rangle$ gives

$$\left\langle y^2 \right\rangle = f_0^2 + (f_1^2 + f_0 f_2)\sigma^2 + (\tfrac{3}{4} f_2^2 + f_1 f_3 + \tfrac{1}{4} f_0 f_4)\sigma^4 + (\tfrac{5}{12} f_3^2 + \tfrac{5}{8} f_2 f_4)\sigma^6 + \tfrac{35}{192} f_4^2 \sigma^8$$

$$\langle y \rangle^2 = f_0^2 + f_0 f_2 \sigma^2 + \frac{1}{4}(f_2^2 + f_0 f_4)\sigma^4 + \tfrac{1}{8} f_2 f_4 \sigma^6 + \tfrac{1}{64}\sigma^8$$

$$P_y = \left\langle y^2 \right\rangle - \langle y \rangle^2 = f_1^2 \sigma^2 + (\tfrac{1}{2} f_2^2 + f_1 f_3)\sigma^4 + (\tfrac{5}{12} f_3^2 + \tfrac{1}{2} f_2 f_4)\sigma^6 + \tfrac{1}{6} f_4^2 \sigma^8 .$$

The variance of the transformed sigma points is

$$\overline{P}_y = w_0 (y_0 - \overline{y})^2 + w_1 [(y_1 - \overline{y})^2 + (y_{-1} - \overline{y})^2 .$$

Substituting for the weights and y_{avg} then leads to

$$\left(P_y - \overline{P}_y \right) = (\tfrac{1}{6} f_3^2 + \tfrac{1}{4} f_2 f_4)\sigma^6 + \tfrac{13}{96} f_4^2 \sigma^8 ,$$

which shows that the terms through $O(\sigma^4)$ have been canceled using the *same* choice of weights as used to optimize the mean estimate. But we should be careful: we are not expanding in σ, which need not be small. Thus, the point to notice is that there is still a term proportional to f_2, which is the second derivative of $f(x)$ about $x = \mu$. Thus, the sigma-point approach here leads to an approximate representation that is significantly better for the mean than for the variance.

c. For $y = x^2$, the UT gives

$$\overline{y} = \frac{2}{3}(0)^2 + 2\frac{1}{6}\left(\sqrt{3}\right)^2 = 1$$

$$\overline{P}_y = \frac{2}{3}(0-1)^2 + 2\frac{1}{6}(3-1)^2 = 2 ,$$

which matches the exact values $\langle y \rangle = \left\langle x^2 \right\rangle = 1$ and $\left\langle y^2 \right\rangle = \left\langle x^4 \right\rangle = 3$. (The latter implies that $P_y = \left\langle y^2 \right\rangle - \langle y \rangle^2 = 3 - 1 = 2$.)

For $y = x^4$, the UT gives

$$\overline{y} = \frac{2}{3}(0)^4 + 2\frac{1}{6}\left(\sqrt{3}\right)^4 = 3$$

$$\overline{P}_y = \frac{2}{3}(0-3)^2 + 2\frac{1}{6}(9-3)^2 = 18 .$$

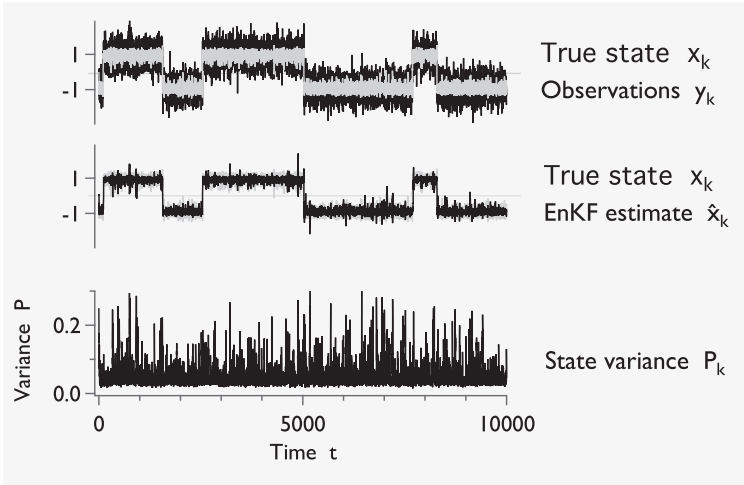
The exact moments for the transformed distribution are $\langle y \rangle = \left\langle x^4 \right\rangle = 3$ and $\left\langle y^2 \right\rangle = \left\langle x^8 \right\rangle = 105$, which implies $P_y = 105 - 9 = 96$. Thus, the UT matches the mean, as expected, but fails for the variance, again as expected.

In both these examples, it is worth noting that the EKF approach (local linearization) gives 0 for mean and variance estimates, a far worse prediction.

8.16 Ensemble Kalman filter (EnKF). Show that the EnKF can track motion in the double-well potential from Example 8.3. What is the effect of varying the number of elements n_{E} in the ensemble? (Remember that you choose n_{E} states and an equal number of noise elements ν and ξ.)

Solution.

See book website for code. Below shows how the EnKF works for the double-well potential with $a = 0.8$, $\nu = 0.15$, and $\xi = 0.5$. Indeed, we use the same noise (by setting the seed for the random-number generator), although we find that it always works. The figure below uses $n_E = 100$. Decreasing n_E increases the estimation variance and the disagreement between the actual error and the state variance. For example, when $n_E = 10$, the empirical variance is 0.064 whereas the average value of P_k is 0.041. But for $n_E = 100$, the corresponding values are 0.055 and 0.051. There is little change for $n_E = 1000$, where the values are 0.051 and 0.053. Since the observational noise variance $\xi^2 = 0.25$, the EnKF is a noticeable improvement over the naive observations, as can be seen in the figure below.



8.17 Grid method. We explore numerically the full Bayesian filtering solution for the double-well potential for various noise strengths ν and ξ.

a. Code the grid method and reproduce the equivalent of Figure 8.6.
b. Fix the input noise at $\nu = 0.15$ and study the state estimation problem as the measurement noise ξ goes from 0 (no noise) to large values. Discuss qualitatively.
c. Now fix the measurement noise $\xi = 1$ and vary the input noise from zero to large values. Again, describe qualitatively the different regimes.
d. In our problem of free diffusion, we found that the behavior of the Kalman filter depended only on the ratio $\alpha = \nu^2/\xi^2$ and not on the absolute values of the two noise strengths. Why does that conclusion not hold in this problem?

Solution.

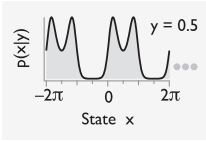
a. Should get something like the figures.
b. Qualitatively: we track better and better, but with an increasing delay as we increase the measurement noise. This makes sense: we need more time to

average over measurement readings to know whether a deviation is really or noisy. The optimal Bayes estimate shows how to do this in the best possible way, given the observations and noise statistics. When the measurement noise becomes larger than about 3, it becomes hard to distinguish between a fluctuation and a movement between wells. At this point, the estimation process becomes much more confused, with broad probability distributions that make it difficult to "know" which well the particle is in.

c. Qualitatively: For low ν, transitions are very rare. Their frequency is described by the Kramers relation, $e^{E_b \nu^2}$, where $E_b = \frac{1}{4}$ is the barrier height, which sets a natural scale (see below). For noise strengths comparable to this barrier scale, we lose the notion of hopping and have a continuous, modulated occupation of the entire space.

d. In the linear diffusion problem, the linear dynamics did not provide a scale to the state variable, the position x. Thus, a uniform rescaling does not change the physics. Those statements do not hold for the nonlinear double-well potential, where the barrier height is a natural energy scale, and we can measure the noise strength in terms of the fluctuations relative to the barrier position.

8.18 Sinusoidal nonlinearity in measurement function. The nonlinear measurement relation $y = \sin x + \xi$ can occur in interferometry experiments (Section 3.2.1). Note that a given y corresponds to an infinite number of possible x states. The graph at left for $P(x|y)$ was generated using $\xi^2 = 0.4^2$ and $y = 0.5$. Why the funny double bump? Explore the consequences of different noise strengths and observations, and explain what is going on in the different cases. Explain, in words, a strategy for dealing with the infinite number of possibilities. As usual, $\xi \sim \mathcal{N}(0, \xi^2)$.



Solution.

To eliminate the infinity of possibilities, you need to first establish, independently, which one is relevant. Then you need to have measurements that are rapid enough that there is no chance for the system to displace by 2π between updates. This is standard procedure for an interferometer, for example.

8.19 Fat tails. To understand how non-Gaussian noise can affect state estimation, consider a somewhat artificial example where both system and observation noise are drawn from Lorentz distributions whose "fat" tails ($\sim 1/x^2$ for $|x| \to \infty$) imply that large fluctuations are vastly more probable than with a normal distribution.

a. Show that $x \sim \mathrm{Lor}(x_0, \nu) = \frac{1}{\pi} \frac{\nu}{(x-x_0)^2+\nu^2}$ is normalized, but the mean and variance diverge. Show that the median equals x_0, $\mathrm{Prob}(x_0 - \nu, x_0 + \nu) = \frac{1}{2}$ (thus connecting ν with a notion of width), and the characteristic function is $\varphi_x(k) = e^{ikx_0 - \nu|k|}$.

b. Consider a toy state-estimation problem where we have a prediction x and an observation y that are both unbiased estimates of the true state of the

system. As with the Kalman filter, we seek the best linear combination $\hat{x}$ of the two. Here, both prediction and observation are "Lévy-flights" that obey $x \sim \mathrm{Lor}(x_t, \nu)$ and $y \sim \mathrm{Lor}(x_t, \xi)$, with x_t the true state value. Let $\hat{x} = (1 - K)x + Ky$ and match characteristic functions to show that $\hat{x} \sim \mathrm{Lor}(x_t, \gamma)$,with width $\gamma = |1 - K|\nu + |K|\xi$.

c. Conclude that the "optimal" choice of Kalman gain K that minimizes γ is 0 if $\nu < \xi$ and 1 if $\xi < \nu$. That is, unlike the ordinary Kalman filter, "blending" the prediction x with the observation y does not improve the accuracy of estimation. Because the fluctuations in x and y are so "wild," the best action is to select at each time step whichever variable has the smaller distribution width. For fixed ν and ξ, this means ignoring all observations if $\nu < \xi$. For $\xi < \nu$, we would use only the naive observation. In other words, the Kalman reduces to a trivial course of action. Fat tails that go as $|x|^{-(1+\mu)}$, with $1 < \mu < 2$ (so that the mean but not the variance is defined), lead to a nontrivial gain K that minimizes the width γ. Sornette and Ide (2001) dub the result the *Kalman-Lévy* filter.

Unfortunately, the "stable" property of the Lorentz distribution (sum of two Lorentzians is also Lorentzian) holds for only a few distributions (extensively studied by Paul Lévy). The more realistic case where the system dynamics has a fat tail but observations are Gaussian does not lead to simple analytic results.

Solution.

a. Many of the requested identities can be derived by showing that

$$\int_{x_1}^{x_2} \mathrm{d}x \, \frac{1}{\pi} \left[\frac{\nu}{(x - x_0)^2 + \nu^2} \right] = \frac{1}{\pi} \left[\tan^{-1} \left(\frac{x_0 - x_1}{\nu} \right) - \tan^{-1} \left(\frac{x_0 - x_2}{\nu} \right) \right]$$

Setting $x_1 = -\infty$ and $x_2 = +\infty$ proves normalization of the PDF. Setting $x_1 = -\infty$ and $x_2 = 0$ proves the statement about the median. Setting $x_1 = x_0 - \nu$ and $x_2 = x_0 + \nu$ proves the statement about the width. The mean diverges because the integrand $\to (1/x)$ for large x and hence gives a logarithmic term. The variance diverges more strongly.

The characteristic function is a standard Fourier transform (with a translation term of x_0).

If you have never played with fat-tailed distributions, it is worth simulating them to get an intuition. Many programs have built-in "Lorentz-noise" (or "Cauchy-noise") functions. If not, from Problem A.6.4, we know that if x and y are $\sim \mathcal{N}(0, 1)$, then $(y/x) \sim \mathrm{Lor}(0, 1)$. This gives a way to draw samples from a Lorentzian distribution for Monte Carlo simulations.

b. The characteristic functions of x and y are $\varphi_x = \mathrm{e}^{\mathrm{i}kx_t - \nu|k|}$ and $\varphi_y = \mathrm{e}^{\mathrm{i}kx_t - \xi|k|}$, respectively. (Recall that the latter is true because we assume that the position was exactly known at the time of observation and thus the only error in the observation is the observation noise. Obviously, the story will be more complicated for the full dynamical Kalman filter.)

Since $\varphi_{ax} = \varphi_x(ak)$, we have

$$\varphi_{(1-K)x} = \mathrm{e}^{\mathrm{i}(1-K)x_t - \nu|(1-K)k|} \quad \text{and} \quad \varphi_{Ky} = \mathrm{e}^{\mathrm{i}kx_t - \nu|Kk|} \,.$$

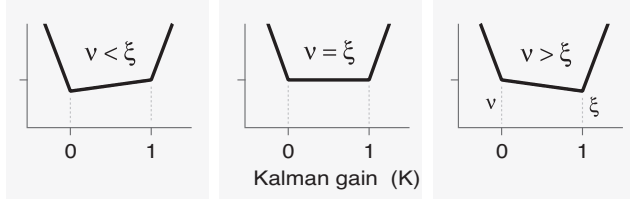
Then, we recall that the distribution of the sum of two random variables is given by the convolution of the respective probability distributions, implying that the characteristic function of the sum is the product of the individual characteristic functions. (See Example A.18.) Thus,

$$\begin{aligned}\varphi_{(1-K)x+Ky} &= \varphi_x[(1-K)k]\,\varphi_y[Kk] \\ &= \mathrm{e}^{\mathrm{i}k[(1-K)x_t + Kx_t] - \left[(|1-K|\nu + |K|\xi)\right]|k|} \\ &= \mathrm{e}^{\mathrm{i}kx_t - \gamma|k|} \,,\end{aligned}$$

where the width of the new Lorentzian distribution is

$$\gamma = |1-K|\nu + |K|\xi \,.$$

c. This is minimized by picking $K = 0$ or 1. See plots of $\gamma(K)$ for different combinations of ν and ξ values.



8.20 Bayesian RTS smoothing. In the text, we give a naive algorithm for smoothing that combines information from the forward and backwards dynamics to improve the estimate of a current state, $\boldsymbol{x}_k$. Here, we explore a more efficient, better-behaved smoother algorithm that does not need an explicit backwards dynamics.

a. By introducing the state $\boldsymbol{x}_{k+1}$ and applying causality, show that

$$p(\boldsymbol{x}_k|\boldsymbol{y}^N) = p(\boldsymbol{x}_k|\boldsymbol{y}^k) \int \mathrm{d}\boldsymbol{x}_{k+1} \, \frac{p(\boldsymbol{x}_{k+1}|\boldsymbol{x}_k)\, p(\boldsymbol{x}_{k+1}|\boldsymbol{y}^N)}{p(\boldsymbol{x}_{k+1}|\boldsymbol{y}^k)} \,,$$

which goes from $p(\boldsymbol{x}_{k+1}|\boldsymbol{y}^N)$ to $p(\boldsymbol{x}_k|\boldsymbol{y}^N)$. To apply it, first use forward Bayesian filtering, Eqs. (8.76), to find the $p(\boldsymbol{x}_k|\boldsymbol{y}^k)$ and their associated predictions $p(\boldsymbol{x}_{k+1}|\boldsymbol{y}^k)$.

b. Assuming linear dynamics and Gaussian probability distributions leads to the *Reich–Tung–Striebel* (RTS) smoother equations. Let $P(\boldsymbol{x}_k|\boldsymbol{y}^N) = \mathcal{N}\left[(\boldsymbol{x}^{\mathrm{s}})_k\,, (\boldsymbol{P}^{\mathrm{s}})_k\right]$ define the smoother state estimate $\boldsymbol{x}^{\mathrm{s}}$ and covariance matrix $\boldsymbol{P}^{\mathrm{s}}$. Show that these quantities may be found via the backwards recurrence relation

$$\begin{aligned}(\boldsymbol{G}^{\mathrm{s}})_k &= \boldsymbol{P}_k \boldsymbol{A}^{\mathsf{T}} \left(\boldsymbol{P}^-\right)^{-1}_{k+1} \\ (\boldsymbol{P}^{\mathrm{s}})_k &= \boldsymbol{P}_k + (\boldsymbol{G}^{\mathrm{s}})_k \left[(\boldsymbol{P}^{\mathrm{s}})_{k+1} - \boldsymbol{P}^-_{k+1}\right] (\boldsymbol{G}^{\mathrm{s}})_k{}^{\mathsf{T}} \\ (\boldsymbol{x}^{\mathrm{s}})_k &= \boldsymbol{x}_k + (\boldsymbol{G}^{\mathrm{s}})_k \left[(\boldsymbol{x}^{\mathrm{s}})_{k+1} - \boldsymbol{x}^-_{k+1}\right] ,\end{aligned}$$

where $\boldsymbol{x}$, $\boldsymbol{P}$, and $\boldsymbol{P}^-$ are first calculated using the forward Kalman filter. The recurrence relation starts at $k = N$, where $\boldsymbol{x}^{\mathrm{s}} = \boldsymbol{x}$ and $\boldsymbol{P}^{\mathrm{s}} = \boldsymbol{P}$ (Särkkä, 2013). Hint: Use Eq. (A.197) for conditional Gaussian distributions and computer algebra.

c. Using the RTS smoother equations, verify that the steady-state smoother variance of the one-dimensional diffusing particle is given by $P^{\mathrm{s}} = \xi^2\alpha/\sqrt{\alpha^2 + 4\alpha}$, in agreement with our previous result, Eq. (8.131).

Solution.

a. We introduce the state $\boldsymbol{x}_{k+1}$ (and will remove it later by marginalization). Using the definition of conditional probability then gives

$$p(\boldsymbol{x}_k, \boldsymbol{x}_{k+1}|\boldsymbol{y}^N) = p(\boldsymbol{x}_k|\boldsymbol{x}_{k+1}, \boldsymbol{y}^N)\, p(\boldsymbol{x}_{k+1}|\boldsymbol{y}^N)\,.$$

Then,

$$\begin{aligned} p(\boldsymbol{x}_k|\boldsymbol{x}_{k+1}, \boldsymbol{y}^N) &= p(\boldsymbol{x}_k|\boldsymbol{x}_{k+1}, \boldsymbol{y}^k) && \text{causality} \\ &= \frac{p(\boldsymbol{x}_k, \boldsymbol{x}_{k+1}|\boldsymbol{y}^k)}{p(\boldsymbol{x}_{k+1}|\boldsymbol{y}^k)} && \text{conditional probability} \\ &= \frac{p(\boldsymbol{x}_{k+1}|\boldsymbol{x}_k, \boldsymbol{y}^k)\, p(\boldsymbol{x}_k|\boldsymbol{y}^k)}{p(\boldsymbol{x}_{k+1}|\boldsymbol{y}^k)} && \text{Markov}\,, \end{aligned}$$

where Markov dynamics implies that if we know $\boldsymbol{x}_k$, then our knowledge of $\boldsymbol{x}_{k+1}$ is not improved by any of the observations in the set $\boldsymbol{y}^k$. Putting it all together,

$$p(\boldsymbol{x}_k, \boldsymbol{x}_{k+1}|\boldsymbol{y}^N) = \frac{p(\boldsymbol{x}_k|\boldsymbol{y}^k)\, p(\boldsymbol{x}_{k+1}|\boldsymbol{x}_k)\, p(\boldsymbol{x}_{k+1}|\boldsymbol{y}^N)}{p(\boldsymbol{x}_{k+1}|\boldsymbol{y}^k)}$$

Integrating both sides with respect to $\boldsymbol{x}_{k+1}$ then gives the identity.

b. See Särkkä (2013) for the solution.

c. From the Kalman analysis, the steady-state forward covariance is

$$P = \tfrac{\xi^2}{2}(\sqrt{\alpha^2 + 4\alpha} - \alpha)$$

and the steady-state predicted covariance is $P^- = \tfrac{\xi^2}{2}(\sqrt{\alpha^2 + 4\alpha} + \alpha)$. This implies that the steady-state smoother gain is

$$G^{\mathrm{s}} = \frac{P}{P^-} = \frac{\sqrt{\alpha^2 + 4\alpha} - \alpha}{\sqrt{\alpha^2 + 4\alpha} + \alpha}\,.$$

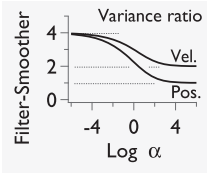
The steady-state smoother variance P^{s} obeys

$$P^{\mathrm{s}} = P - (G^{\mathrm{s}})^2(P^- - P^{\mathrm{s}})\,.$$

Solving for P^{s} gives

$$P^{\mathrm{s}} = \frac{P}{1 + G^{\mathrm{s}}} = \frac{P\,P^-}{P^- + P} = \xi^2\,\frac{\alpha}{\sqrt{\alpha^2 + 4\alpha}}\,.$$

8.21 The random rocket. In one dimension, a rocket with random forcing obeys $\ddot{x} = \nu(t)$. Assume that the measured positions are noisy: $y(t) = x(t) + \xi(t)$.



a. Discretize the system exactly for a time step that is scaled to $T_s = 1$.
b. Write a numerical code for a Kalman filter and then a Kalman smoother. Explore the case where the discrete-noise variances are $\nu^2 = \xi^2 = 1$.
c. Scale all variances by ξ^2 and define $\alpha = \nu^2/\xi^2$. Confirm that the ratio of filter to smoother variances is given by the plot at left.
d. Plot time series of position, measurements, and both filter and smoother estimates.

Hints: $\boldsymbol{A}_d = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$. For $\alpha = 1$ and steady state, $\boldsymbol{P} = \left(\begin{smallmatrix} 3/4 & 1/2 \\ 1/2 & 1 \end{smallmatrix}\right)$ and $\boldsymbol{P}^s = \left(\begin{smallmatrix} 1/3 & 0 \\ 0 & 1/3 \end{smallmatrix}\right)$.

Solution.

a. The continuous matrices are

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \qquad \boldsymbol{B} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \qquad \boldsymbol{C} = \begin{pmatrix} 1 & 0 \end{pmatrix}.$$

We find the discrete matrices by exponentiation, noting that $\boldsymbol{A}^2 = \boldsymbol{0}$. Then,

$$\boldsymbol{A}_d = e^{\boldsymbol{A}T_s} = \mathbb{I} + T_s\boldsymbol{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + T_s \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & T_s \\ 0 & 1 \end{pmatrix}$$

$$\boldsymbol{B}_d = \int_0^{T_s} dt\, e^{\boldsymbol{A}(T_s - t)}\, \boldsymbol{B} = \int_0^{T_s} dt \begin{pmatrix} 1 & T_s - t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \int_0^{T_s} dt \begin{pmatrix} T_s - t \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{T_s^2}{2} \\ T_s \end{pmatrix}.$$

We scale the update time so that $T_s = 1$; thus,

$$\boldsymbol{A}_d = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \qquad \boldsymbol{B}_d = \begin{pmatrix} 1/2 \\ 1 \end{pmatrix} \qquad \boldsymbol{C}_d = \begin{pmatrix} 1 & 0 \end{pmatrix}.$$

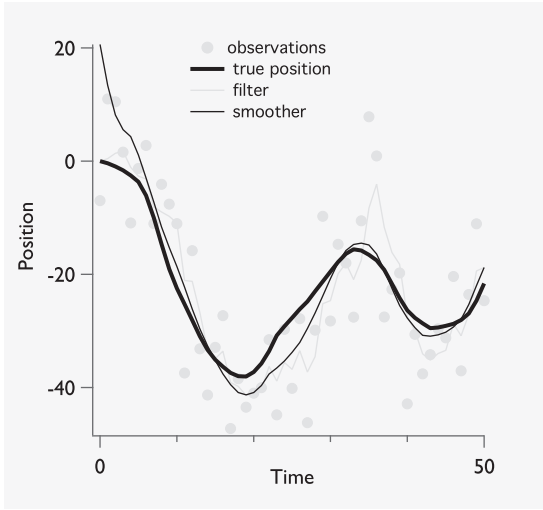
b. For $\alpha = 1$, you should find, for the steady-state filter,

$$\boldsymbol{P}^- = \begin{pmatrix} 3 & 2 \\ 2 & 2 \end{pmatrix} \qquad \boldsymbol{P} = \begin{pmatrix} 3/4 & 1/2 \\ 1/2 & 1 \end{pmatrix} \qquad \boldsymbol{L} = \begin{pmatrix} 3/4 \\ 1/2 \end{pmatrix},$$

and for the steady-state smoother,

$$\boldsymbol{P}^s = \begin{pmatrix} 1/3 & 0 \\ 0 & 1/3 \end{pmatrix}.$$

c. For general α, the graph of the ratio of steady-state variances is given in the problem's margin. It is possible expand for small and large α, but the results do not seem to be particularly informative. (Similarly, one can solve the $\alpha = 1$ problem analytically, but to no obvious benefit.)
d. Below is a plot of typical position measurements (light gray dots) for $\nu = 1$ and $\xi = 10$ (or, $\alpha = 10^{-2}$). The heavy dark line is the true position. The lightest gray trace is the filter estimate. The thin dark line is the smoother estimate, which, indeed, is smoother than the filter. Except at the beginning, the smoother is also more accurate.



8.22 Estimating a frequency. Consider the problem of estimating the frequency ω of a complex signal $y_k = e^{i\omega k} + \xi_k$, with $\xi_k \sim \mathcal{CN}(0, \sigma^2)$, for N data points.

a. Show that $\hat{\omega} = \text{argmax}_\omega \; F(\omega) \equiv \text{Re}\,(\frac{1}{N}\sum_{k=0}^{N-1} y_k\, e^{-i\omega k})$ is the maximum-likelihood estimator of ω. If you are not familiar with complex Gaussian distributions, start by considering the real and imaginary parts of the measurement equation, assuming that Re ξ_k and Im ξ_k are drawn from i.i.d. Gaussian distributions, $\mathcal{N}(0, \frac{1}{2}\sigma^2)$.

b. Use an FFT algorithm to find $\hat{\omega}$. Refine the estimate via a local-optimization routine. Plot $F(\omega)$ for good and bad SNR, and reproduce the figures in Example 8.4.

Solution.

a. With the notation $y^k = \{y_0, y_1, \ldots, y_{k-1}\}$, the log likelihood of $p(y^k|\omega)$, is given by the joint distribution of real and imaginary parts of $y_k = e^{i\omega k} + \xi_k$.

$$p(y^k|\omega) = \left(\frac{1}{\sqrt{2\pi(\sigma^2/2)}}\right)^N \prod_{k=0}^{N-1} e^{-\frac{\text{Re}\left(y_k - e^{i\omega k}\right)^2}{2(\sigma^2/2)}} \left(\frac{1}{\sqrt{2\pi(\sigma^2/2)}}\right)^N \prod_{k=0}^{N-1} e^{-\frac{\text{Im}\left(y_k - e^{i\omega k}\right)^2}{2(\sigma^2/2)}}$$
$$= \left(\frac{1}{\pi\sigma^2}\right)^N \prod_{k=0}^{N-1} e^{-\frac{\left|y_k - e^{i\omega k}\right|^2}{\sigma^2}} .$$

Notice how the complex notation and circular symmetry (the fact that real and imaginary parts are independent and identically distributed), greatly simplifies the joint distribution. Then, starting from the complex noise distribution, the log-likelihood is

$$\ln\, p(y^k|\omega) = \ln\left[\left(\frac{1}{\pi\sigma^2}\right)^N \prod_{k=0}^{N-1} e^{-\frac{\left|y_k - e^{i\omega k}\right|^2}{\sigma^2}}\right]$$
$$= -N \ln\left(\pi\sigma^2\right) - \frac{1}{\sigma^2}\sum_{k=0}^{N-1}\left|y_k - e^{i\omega k}\right|^2 .$$

Expanding the magnitude-squared term for time k gives

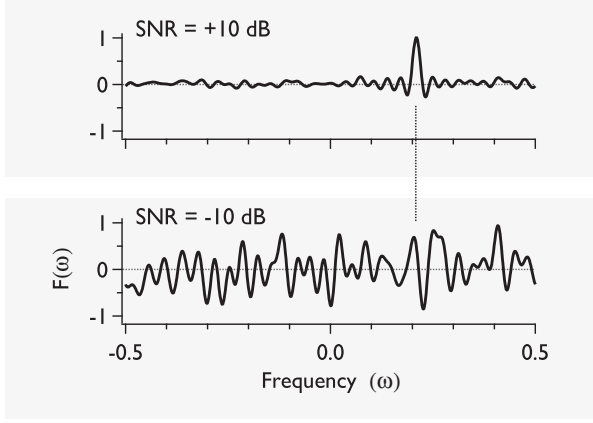
$$\begin{aligned}\left|y_k - \mathrm{e}^{\mathrm{i}\omega k}\right|^2 &= \left(y_k - \mathrm{e}^{\mathrm{i}\omega k}\right)\left(y_k^* - \mathrm{e}^{-\mathrm{i}\omega k}\right) \\ &= |y_k|^2 + 1 - y_k\,\mathrm{e}^{-\mathrm{i}\omega k} - y_k^*\,\mathrm{e}^{+i\omega k} \\ &= -2\ \mathrm{Re}\ \left(y_k\ \mathrm{e}^{-\mathrm{i}\omega k}\right) + \text{ terms independent of } \omega\,.\end{aligned}$$

Substituting and ignoring scaling factors (irrelevant for finding a maximum) and also terms independent of ω gives,

$$\ln\ p(y^k|\omega) \propto F(\omega) = +\mathrm{Re}\left(\frac{1}{N}\sum_{k=0}^{N-1} y_k\ \mathrm{e}^{-\mathrm{i}\omega k}\right),$$

with the normalization factor N^{-1} included by convention, to make $F = O(1)$.

b. The expression $F(\omega)$ is just the quantity evaluated by a fast Fourier Transform (FFT), although that algorithm evaluates only at a discrete set of frequencies. Still, it is a fast, easy algorithm, and the frequencies identified are closely spaced enough to identify the local peak. Plotting $F(\omega)$ for SNR of +10 and −10 dB shows how, in the former case, the global maximum is at the proper frequency while, for the low-SNR case, the global maximum is usually at a noise peak.



8.23 Wiener filtering. The Wiener filter is a frequency-domain technique equivalent to (and predating) the Kalman filter for time-invariant problems. We focus on the much-simpler smoothing case, where information is available for all times ($-\infty < t < +\infty$). Consider a signal $u(t)$ measured by an instrument with dynamical response $\mathcal{L}x(t) = u(t)$, where $\mathcal{L}$ is a differential operator. Its inverse in the Fourier domain is the transfer function $G(\omega)$, with $x(\omega) = G(\omega)u(\omega)$. The measured response $y(t) = x(t) + \xi(t)$ adds white noise $\xi(t)$ with spectral density ξ^2. The goal is to find an optimal linear "filter" that minimizes the mean-square estimation error. (Notice that the goal of estimating u rather than x is slightly different from that of a Kalman filter. But given $\hat{u}$, we have $\hat{x} = G\hat{u}$.) We thus define $\hat{u} = \frac{W}{G}y$, where all quantities are functions of ω. The mean-square estimation error is

$E = \int_0^\infty \mathrm{d}t\,[\hat{u}(t) - u(t)]^2 = \int_{-\infty}^\infty \frac{\mathrm{d}\omega}{2\pi}|\hat{u}(\omega) - u(\omega)|^2$, where the signal $u(t)$ is taken to be white noise with spectral density u^2.

a. Show that $\hat{u} = [G^{-1}\,(1 + \frac{1}{|G|^2\mathrm{SNR}^2})^{-1}]\,y$ minimizes the mean-square error, with $\mathrm{SNR} \equiv u/\xi$. The *whitening filter* G^{-1} compensates for the instrumental response.

b. Assume that $G = \frac{1}{1+\mathrm{i}\omega}$, a first-order, low-pass filter. For $\mathrm{SNR} \gg 1$, show that $(1 + \frac{1}{|G|^2\mathrm{SNR}^2})^{-1}]$ is a low-pass filter with cut-off frequency $\omega_\mathrm{c} \approx \mathrm{SNR}$. Thus, the Wiener filter cuts off the naive estimator $u(\omega) = G^{-1}(\omega)\,y(\omega)$ at ω_c.

The optimal filter contains the term $|G|^2$, making it acausal. Finding a causal Wiener filter that does not depend on future values of the signal turns out to be a harder problem and is solved more easily by the Kalman filter.

Solution.

a. The mean-square error, in the frequency domain is given by

$$\begin{aligned} E &= \int_{-\infty}^\infty \frac{\mathrm{d}\omega}{2\pi}|\hat{u}(\omega) - u(\omega)|^2 \\ &= \int_{-\infty}^\infty \frac{\mathrm{d}\omega}{2\pi}\left|\frac{W}{G}(x+\xi) - \frac{x}{G}\right|^2 \\ &= \int_{-\infty}^\infty \frac{\mathrm{d}\omega}{2\pi}\frac{1}{|G|^2}\left[|(W-1)x|^2 + |W|^2\xi^2\right], \end{aligned}$$

where the cross term $x(\omega)\xi(\omega)$ vanishes when integrated, since the signal and measurement noise are uncorrelated. With W^* the complex conjugate of W, setting the derivative $(\partial_{W^*})E = 0$ gives

$$(W-1)|x|^2 + W\xi^2 = 0\,,$$

which implies

$$W = \frac{|x|^2}{|x|^2 + \xi^2} = \frac{1}{1 + \xi^2/|x|^2} = \frac{1}{1 + \frac{1}{(|G|^2\mathrm{SNR}^2)}}\,.$$

b. For an instrument response that is a low-pass filter with unit cutoff, $G(\omega) = \frac{1}{1+\mathrm{i}\omega}$. Then

$$W = \frac{1}{1 + \frac{1+\omega^2}{\mathrm{SNR}^2}} \approx \frac{1}{1 + \frac{\omega^2}{\mathrm{SNR}^2}}\,,$$

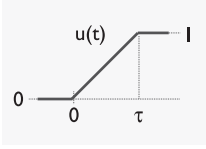
which is indeed the magnitude-squared frequency response of a low-pass filter with cutoff $\omega_\mathrm{c} = \mathrm{SNR}$, assuming $\mathrm{SNR} \gg 1$.

9 Robust Control

Problems

9.1 **Input shaping: Moving a load of uncertain mass**. From Section 9.1.1, we consider a transfer function $G(s) = \frac{1}{1+(s/\omega)^2}$, with ω an unknown oscillation frequency of nominal value $\omega_0 = 1$. The goal is to move from y from 0 to 1 in finite time using the input-shaping protocol: $n+1$ steps of amplitude $\mathbf{A} = \{A_i\}$, applied at times $\mathbf{t} = \{t_i\}$.

a. Show that the amplitude of residual oscillations is given by Eq. (9.3).
b. Zero Vibration (ZV): Show that $\boldsymbol{A} = \left\{\frac{1}{2}, \frac{1}{2}\right\}$ and $\mathbf{t} = \{0, \pi\}$ satisfies $J(\omega) = 0$ for $\omega = \omega_0 = 1$, which thus solves the control problem exactly if the system is known perfectly. Find the exact expression for $J_1(\omega)$. Here and below, set $t_0 \equiv 0$.
c. Zero Vibration Derivative (ZVD): Show that $\boldsymbol{A} = \left\{\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right\}$ and $\mathbf{t} = \{0, \pi, 2\pi\}$ satisfies $J_2 = J_2' = 0$ at $\varepsilon \equiv \omega - 1 = 0$.
d. Zero Vibration Double Derivative (ZVDD): Show that $\boldsymbol{A} = \left\{\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}\right\}$ and $\mathbf{t} = \{0, \pi, 2\pi, 3\pi\}$ satisfies $J_3 = J_3' = J_3'' = 0$ at $\varepsilon = 0$.
e. Show that the Taylor expansions of ZV, ZVD, and ZVDD solutions about $\omega = 1$ give $\left(\frac{\pi}{2}|\varepsilon|\right)^n$, with $n = 1, 2, 3$, respectively.
f. *Adiabatic limit*. Show that the ramp at left leads to residual oscillations whose typical amplitude is $(\omega\tau)^{-1}$, which is small for $\tau \gg \omega^{-1}$.



Solution.

a. In standard state-space notation, the dynamics $\ddot{y} + \omega^2 y = \omega^2 u(t)$ correspond to

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 \\ \omega^2 \end{pmatrix} u(t)\,, \qquad y(t) = x_1(t)$$

which has a solution, for $y(0) = \dot{y}(0) = 0$, of

$$\begin{aligned} y(t) &= \begin{pmatrix} 1 & 0 \end{pmatrix} \int_0^t \mathrm{d}t' \begin{pmatrix} \cos\omega(t-t') & \frac{1}{\omega}\sin\omega(t-t') \\ -\omega\sin\omega(t-t') & \cos\omega(t-t') \end{pmatrix}\begin{pmatrix} 0 \\ \omega^2 \end{pmatrix} u(t') \\ &= \omega \int_0^t \mathrm{d}t' \sin\omega(t-t')\, u(t')\,. \end{aligned}$$

Substituting $u(t) = \sum_i A_i\,\theta(t - t_i)$ and assuming $t > t_n$, the last step, gives

$$y(t) = \omega \sum_{i=0}^{n} A_i \int_{t_i}^{t} \mathrm{d}t' \sin\omega(t - t') = \frac{\omega}{\omega} \sum_{i=0}^{n} A_i[1 - \cos\omega(t - t_i)]$$
$$= 1 - \sum_{i=0}^{n} A_i \cos\omega(t - t_i)\,.$$

Here, we use the normalization of relative amplitudes, $\sum_i A_i = 1$. This last condition must be imposed to make $y(t)$ oscillate about the desired position $y = 1$ at the end of the protocol.

To find the amplitude of residual oscillations, we expand the cosine term:

$$\sum_{i=0}^{n} A_i \cos\omega(t - t_i) = \sum_{i=0}^{n} A_i\,(\cos\omega t \cos\omega t_i + \sin\omega t \sin\omega t_i)$$
$$= \left(\sum_{i=0}^{n} A_i \cos\omega t_i\right) \cos\omega t + \left(\sum_{i=0}^{n} A_i \sin\omega t_i\right) \sin\omega t\,,$$

which has amplitude

$$J_n = \sqrt{\left(\sum_{i=0}^{n} A_i \cos\omega t_i\right)^2 + \left(\sum_{i=0}^{n} A_i \sin\omega t_i\right)^2}\,.$$

To see this last identity, we note that we find the amplitude of

$$\delta y = \alpha\cos\omega t + \beta\sin\omega t$$

by finding the extremum. Setting the derivative to zero, we have

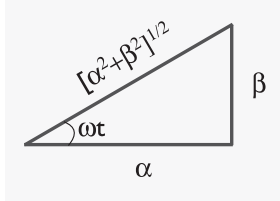
$$\frac{\mathrm{d}}{\mathrm{d}t}(\,\delta y) = \omega(-\alpha\sin\omega t + \beta\cos\omega t) = 0\,.$$

This implies

$$\tan\omega t = \beta/\alpha \quad \Longrightarrow \quad \cos\omega t = \tfrac{\alpha}{\sqrt{\alpha^2+\beta^2}}\,,\ \sin\omega t = \tfrac{\beta}{\sqrt{\alpha^2+\beta^2}}\,.$$

Thus, as seen below,

$$\delta y = \frac{\alpha^2}{\sqrt{\alpha^2+\beta^2}} + \frac{\beta^2}{\sqrt{\alpha^2+\beta^2}} = \sqrt{\alpha^2+\beta^2}\,.$$



b. We solve

$$J_n = \sqrt{\left(\sum_{i=0}^{n} A_i \cos \omega t_i\right)^2 + \left(\sum_{i=0}^{n} A_i \sin \omega t_i\right)^2} = 0\,,$$

for $n = 1$. For $J_1 = 0$, each sum must vanish. Recalling that $t_0 = 0$, we have

$$A_0 + A_1 \cos \omega t_1 = 0\,, \qquad A_1 \sin \omega t_1 = 0\,.$$

The second equation requires $\omega t_1 = m\pi$ for integer m. The smallest m (shortest protocol) is $m = 1$, which implies $t_1 = \pi$, assuming that $\omega = \omega_0 = 1$. The first equation is then $A_0 - A_1 = 0$, or $A_0 = A_1$. With the normalization $A_0 + A_1 = 1$, we conclude that

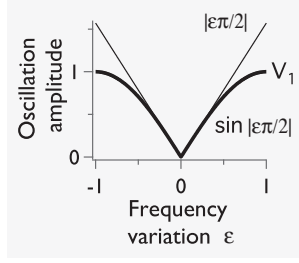
$$\mathbf{A} = \left\{\tfrac{1}{2}, \tfrac{1}{2}\right\}\,, \qquad \mathbf{t} = \{0,\, \pi\}\,.$$

Again, these values assume a nominal oscillator frequency $\omega_0 = 1$.

If a control based on this design is carried out on an oscillator whose actual frequency is ω, the exact expression for residual amplitude J_1 is

$$J_1 = \tfrac{1}{2}\sqrt{(1 + \cos \omega\pi)^2 + (\sin \omega\pi)^2} = \left|\cos \tfrac{\pi}{2}\omega\right| = \left|\sin \tfrac{\pi}{2}\varepsilon\right|\,,$$

where $\varepsilon = \omega - 1$. See below.



More generally, if the design frequency is ω_0 (not necessarily =1), then $t_1 = \pi/\omega_0$ and $\varepsilon = (\omega/\omega_0) - 1$.

c. For $n = 2$,

$$J_2 = \sqrt{(A_0 + A_1 \cos \pi\omega + A_2 \cos 2\pi\omega)^2 + (A_1 \sin \pi\omega + A_2 \sin 2\pi\omega)^2}\,.$$

But symmetry implies $A_2 = A_0$ and normalization implies $A_1 = 1 - 2A_0$. These lead to

$$J_2 = |1 - 2A_0 + 2A_0 \cos \pi\omega|\,.$$

To satisfy $J_2(\omega = 1) = 0$, we have $J_2 = |1 - 2A_0 - 2A_0| = 0$, which implies $A_0 = \frac{1}{4}$. Then,

$$J_2 = \cos\left(\tfrac{\pi}{2}\omega\right)^2\,.$$

d. The $n = 3$ (ZVDD) case is similar. With $A_0 = A_3$ and $A_1 = A_2$ (by symmetry), we have

$$J_3 = 2\left|\cos\left(\tfrac{\pi}{2}\omega\right)\left(-A_0 + A_1 + 2A_0\cos\pi\omega\right)\right| .$$

Taking two derivatives with respect to ω gives

$$J_3''(\omega) = -2\pi^2[A_1(2A_0 + A_1)\cos\pi\omega + A_0(8A_1\cos 2\pi\omega + 9A_0\cos 3\pi\omega)] .$$

Solving for $J_3''(1) = 0$ then implies $A_1 = 3A_0$. Normalization of the A_i then implies that $A_0 = A_3 = \frac{1}{8}$, while $A_1 = A_2 = \frac{3}{8}$. Then

$$J_3 = \left|\cos\left(\tfrac{\pi}{2}\omega\right)\right|^3 ,$$

e. For ZV, ZVD, and ZVDD, the cost functions are

$$J_n = \left|\cos\left(\tfrac{\pi}{2}\omega\right)\right|^n = \left|\sin\left(\tfrac{\pi}{2}\varepsilon\right)\right|^n ,$$

with $n = 1, 2, 3$, respectively. In the latter expression, we change variables to $\varepsilon = \omega - 1$. The Taylor expansions are then

$$J_n \approx \left|\tfrac{\pi}{2}\varepsilon\right|^n .$$

The expression generalizes to arbitrary n-th order, showing that, for a long-enough protocol, we can make the response arbitrarily robust to frequency mismatch.

f. *Adiabatic limit*. The easiest approach is to integrate directly for a ramp solution $u(t) = t/\tau$. That is, we solve

$$\ddot{y} + \omega^2 y = \omega^2(t/\tau) , \qquad y(0) = \dot{y}(0) = 0 .$$

The solution is simply,

$$y(t) = \frac{t}{\tau} - \frac{\sin\omega t}{\omega\tau} , \qquad \dot{y} = \frac{1}{\tau}(1 - \cos\omega t)$$

At the end of the ramp, at time $t = \tau$, this solution simplifies to

$$y(\tau) = 1 - \frac{\sin\omega\tau}{\omega\tau} , \qquad \dot{y}(\tau) = \frac{1 - \cos\omega\tau}{\tau} .$$

Intuitively, we can guess that, with this initial condition and setting $u(t) = 1$ for later times, the oscillations in $y(t)$ will be of order $(\omega\tau)^{-1}$ about the steady-state solution $y = 1$. Below, we show this in more detail.

If $u(t) = 1$ thereafter, the solution, by inspection, is

$$y(t) = 1 + [y(\tau) - 1]\cos\omega(t - \tau) + \frac{\dot{y}(\tau)}{\omega}\sin\omega(t - \tau) ,$$

which just oscillates with amplitude

$$\sqrt{[y(\tau)-1]^2+\left(\frac{\dot{y}(\tau)^2}{\omega}\right)^2} = \sqrt{\left(\frac{\sin\omega\tau}{\omega\tau}\right)^2+\left(\frac{1-\cos\omega\tau}{\omega\tau}\right)^2}$$
$$= \left(\frac{1}{\omega\tau}\right)\sqrt{2(1-\cos\omega\tau)}$$
$$= \left(\frac{2}{\omega\tau}\right)\sin\tfrac{1}{2}\omega\tau\,.$$

Thus, depending on the exact value of $\omega\tau \gg 1$, the oscillation amplitude ranges between 0 and $2(\omega\tau)^{-1}$. A "typical" value is then $(\omega\tau)^{-1}$. If the ramp is slow enough, then the residual oscillation will be small, whatever the frequency of the oscillator. There is a tradeoff between performance and robustness.

9.2 Swing up a pendulum robustly.

a. Derive the equations of motion for the four-dimensional augmented dynamics for $\boldsymbol{X} = (\boldsymbol{x}\ \boldsymbol{x}_\omega)^\mathsf{T}$ that augment Eqs. (9.6). Express them in the form $\dot{\boldsymbol{X}} = \boldsymbol{F}(\boldsymbol{X},\boldsymbol{u})$.
b. Write the eight-dimensional equations for the combined state and adjoint $(\boldsymbol{X}\ \boldsymbol{\Lambda})^\mathsf{T}$.
c. Write a boundary-value code to solve the eight-dimensional equations of motion and make plots similar to the ones given in the text.

Solution.

a. The augmented state vector is

$$\boldsymbol{X} = \begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{x}_\omega\end{pmatrix} = \begin{pmatrix}\theta\\ \dot{\theta}\\ \theta_\omega\\ \dot{\theta}_\omega\end{pmatrix} = \begin{pmatrix}x_1\\ x_2\\ x_{1,\omega}\\ x_{2,\omega}\end{pmatrix}$$

For the ordinary pendulum, $\ddot{\theta} = -\omega^2\sin\theta + u$. Taking a derivative of the right-hand side with respect to ω gives $-2\omega\sin\theta - \omega^2(\cos\theta)\,\theta_\omega$. The combined dynamics are then

$$\underbrace{\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}x_1\\ x_2\\ x_{1,\omega}\\ x_{2,\omega}\end{pmatrix}}_{\dot{\boldsymbol{X}}} = \begin{pmatrix}x_2\\ -\omega^2\sin x_1 + u\\ x_{2,\omega}\\ -2\omega\sin x_1 - \omega^2\cos x_1\, x_{1,\omega}\end{pmatrix} = \underbrace{\begin{pmatrix}x_2\\ -\sin x_1 + u\\ x_{2,\omega}\\ -2\sin x_1 - \cos x_1\, x_{1,\omega}\end{pmatrix}}_{\boldsymbol{F}(\boldsymbol{X},u)}.$$

In the second identity, we evaluate at the nominal frequency, $\omega = 1$.

b. Find the four-dimensional adjoint equations by computing the Jacobian of $\boldsymbol{F}$. Then write the eight-dimensional equations for the combined state-adjoint

system ($\boldsymbol{X}$ $\boldsymbol{\Lambda}$). As before, the functions are evaluated at the nominal frequency $\omega = 1$.

$$-(\partial_{\boldsymbol{X}}\boldsymbol{F})^{\mathsf{T}} = -\left(\begin{array}{cc:cc} 0 & 1 & 0 & 0 \\ -\cos x_1 & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & 1 \\ -2\cos x_1 + \sin x_1\, x_{1,\omega} & 0 & -\cos x_1 & 0 \end{array}\right)^{\mathsf{T}}$$

$$= +\left(\begin{array}{cc:cc} 0 & \cos x_1 & 0 & 2\cos x_1 - \sin x_1\, x_{1,\omega} \\ -1 & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & \cos x_1 \\ 0 & 0 & -1 & 0 \end{array}\right)^{\mathsf{T}}.$$

Then, since L does not depend on $\boldsymbol{x}$ (in this problem),

$$\frac{\mathrm{d}\boldsymbol{\Lambda}}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\lambda_1 \\ \lambda_2 \\ \lambda_{1,\omega} \\ \lambda_{2,\omega}\end{pmatrix} = \left(\begin{array}{cc:cc} 0 & \cos x_1 & 0 & 2\cos x_1 - \sin x_1\, x_{1,\omega} \\ -1 & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & \cos x_1 \\ 0 & 0 & -1 & 0 \end{array}\right)^{\mathsf{T}} \begin{pmatrix}\lambda_1 \\ \lambda_2 \\ \lambda_{1,\omega} \\ \lambda_{2,\omega}\end{pmatrix}$$

$$= \begin{pmatrix} \cos x_1 \lambda_2 + (2\cos x_1 - \sin x_1 x_{1,\omega})\lambda_{2,\omega} \\ -\lambda_1 \\ \cos x_1 \lambda_{2,\omega} \\ -\lambda_{1,\omega} \end{pmatrix}.$$

Finally, we put together the equations for $\boldsymbol{X}$ and $\boldsymbol{\Lambda}$ into a single system of eight coupled equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \\ x_{1,\omega} \\ x_{2,\omega} \\ \lambda_1 \\ \lambda_2 \\ \lambda_{1,\omega} \\ \lambda_{2,\omega} \end{pmatrix} = \begin{pmatrix} x_2 \\ -\sin x_1 + u \\ x_{2,\omega} \\ -2\sin x_1 - \cos x_1\, x_{2,\omega} \\ \cos x_1 \lambda_2 + (2\cos x_1 - \sin x_1 x_{1,\omega})\lambda_{2,\omega} \\ -\lambda_1 \\ \cos x_1 \lambda_{2,\omega} \\ -\lambda_{1,\omega} \end{pmatrix}, \qquad \begin{pmatrix} x_1(0) \\ x_2(0) \\ x_{1,\omega}(0) \\ x_{2,\omega}(0) \\ x_1(\tau) \\ x_2(\tau) \\ x_{1,\omega}(\tau) \\ x_{2,\omega}(\tau) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \pi \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Note that because the u dependence does not change, we will have the same equation, $u(t) = -\lambda_2(t)$. We then substitute $u = -\lambda_2$ in the second equation, for $\dot{x}_2$. The result is a system of eight coupled, nonlinear first-order ODEs with eight boundary conditions.

Although the manual procedure used here is tedious and would be even more so with more uncertain parameters, it is worth noting that all operations are simple, standard ones (differentiation, matrix algebra) and can, in principle, be automated in a symbolic-manipulation package.

c. See book website for code.

9.3 Robust rejection of disturbances for harmonic oscillator.

a. Derive the perturbative result that $\delta k \equiv (\frac{\delta K}{K_0}) = +\frac{3}{2}\,\varepsilon^2$.
b. Let $p(\omega)$ be lognormal, with $\langle\omega\rangle = 1$ and $\langle(\,\delta\omega)^2\rangle = \varepsilon^2$. Show that $\langle\omega^n\rangle = \mathrm{e}^{n\mu+(n\sigma)^2/2}$ and thus $\ln\omega \sim \mathcal{N}(\mu,\sigma^2)$, with $\mu = -(1/2)\ln\!\left(1+\varepsilon^2\right)$ and $\sigma^2 = -2\mu$.
c. Find exact expressions for the scaled optimal cost $\langle j^*\rangle$ and gain k^* as functions of ε. Taylor expand to confirm $\beta = \frac{3}{2}$.

Solution.

a. The problem to solve is

$$\ddot{x} + K\dot{x} + \omega^2 x = 0\,, \qquad x(0) = 0,\ \dot{x}(0) = 1\,,$$

where we have substituted the feedback derivative-control signal $u(t) = -K\dot{x}(t)$ into the system equation of motion and where the initial condition is derived from the delta-function "kick" at time 0. (Integrate $\ddot{x}(t)$ from $t = 0^-$ to 0^+.)

For the underdamped case ($\omega > K/2$), solving the equations of motion leads to

$$x(t) = \mathrm{e}^{-\zeta' t}\left(\frac{\sin\omega' t}{\omega'}\right), \qquad \omega' \equiv \sqrt{\omega^2 - \zeta'^2}\,, \qquad \zeta' \equiv K/2\,.$$

$$u(t) = -K\dot{x} = K\,\mathrm{e}^{-\zeta' t}\left[\left(-\frac{\zeta'}{\omega'}\right)\sin\omega' t + \cos\omega' t\right].$$

The overdamped solution is similar, with hyperbolic functions replaced by trigonometric functions and $(\omega^2 - \zeta'^2) \to (\zeta'^2 - \omega^2)$. As noted in the main text, evaluating *both* cases in the cost function leads to the same result,

$$J(\omega, K) = \int_0^\infty \mathrm{d}t\left[x^2(t) + u^2(t)\right] = \frac{1}{2}\left(K + \frac{1}{\omega^2 K}\right),$$

To find the optimal value of the feedback gain K, we solve

$$\frac{\mathrm{d}J}{\mathrm{d}K} = 0\,, \quad \Longrightarrow \quad K = \frac{1}{\omega}\,, \quad \Longrightarrow \quad K_0 = 1\,.$$

In the last step, we evaluate K at the nominal frequency, $\omega_0 = 1$. The optimal cost is then $J_0 = J(\omega_0, K_0) = 1$.

To find the shift $\delta K/K_0$ in optimal gain as a function of the relative frequency uncertainty $\langle\,\delta\omega^2\rangle$, we evaluate

$$\beta = -\frac{1}{2}\left(\frac{\partial_{k\vartheta\vartheta} J}{\partial_{k,k} J}\right).$$

Differentiating (and using $\vartheta = \omega$ and $k = K$),

$$\partial_{k\vartheta\vartheta}\, J = \frac{\partial^3 J}{\partial k\,\partial\vartheta^2} = -\frac{3}{k^2\vartheta^4} \to -3\,, \qquad \partial_{kk}\, J = \frac{\partial^2 J}{\partial k^2} = \frac{1}{k^3\vartheta^2} \to +1\,,$$

where we substitute the nominal values $k = \vartheta = 1$. Finally,

$$\beta = -\left(\frac{1}{2}\right)\frac{-3}{1} = +\frac{3}{2}\,.$$

It is interesting to note that if we were to redo the calculation allowing for a relative cost R of control in J, we would find out that the coefficient β does not depend on R. It appears in the expression for K_0 but not in its *relative* shift.

b. First, let us first show that $\langle\omega^n\rangle = \mathrm{e}^{n\mu+(n\sigma)^2/2}$ for a lognormal distribution with $\ln\omega \sim \mathcal{N}(\mu,\sigma^2)$. We start by defining

$$y = (\ln\omega) \text{ and } \frac{y-\mu}{\sigma} \equiv z \sim \mathcal{N}(0,1)\,.$$

Then

$$\langle\omega^n\rangle = \langle\mathrm{e}^{ny}\rangle = \langle\mathrm{e}^{n\sigma z+n\mu}\rangle = \mathrm{e}^{n\mu}\,\langle\mathrm{e}^{n\sigma z}\rangle = \underbrace{\mathrm{e}^{n\mu+(n\sigma)^2/2}}_{\text{our result}}\left\langle\mathrm{e}^{n\sigma z-(n\sigma)^2/2}\right\rangle\,.$$

But

$$\begin{aligned}\left\langle\mathrm{e}^{n\sigma z-(n\sigma)^2/2}\right\rangle &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathrm{d}z\,\mathrm{e}^{-z^2/2}\;\mathrm{e}^{n\sigma z-(n\sigma)^2/2}\\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathrm{d}z\,\mathrm{e}^{-(z-n\sigma)^2/2} = 1\,,\end{aligned}$$

which proves our result.

Using the $n = 1$ and $n = 2$ cases, we have

$$\begin{aligned}\langle\omega\rangle &= \mathrm{e}^{\mu+\sigma^2/2} = 1\\ \langle(\,\delta\omega)^2\rangle &= \langle\omega^2\rangle - \langle\omega\rangle^2 = \mathrm{e}^{2\mu+2\sigma^2} - 1 = \varepsilon^2\,.\end{aligned}$$

Solving these two equations for μ and σ^2 gives

$$\mu = -\frac{1}{2}\ln\left(1+\varepsilon^2\right)\,,\quad \sigma^2 = \ln\left(1+\varepsilon^2\right) = -2\mu\,.$$

c. In the cost function, J contains a factor of $1/\omega^2$, which is the $n = -2$ case of the result from (b). Thus,

$$\left\langle\omega^{-2}\right\rangle = \mathrm{e}^{-2\mu+(4\sigma^2)/2} = \mathrm{e}^{-6\mu} = \left(1+\varepsilon^2\right)^3\,.$$

Substituting this formula into the cost function leads to

$$\langle J\rangle = \frac{1}{2}\left(k + \frac{\left\langle\omega^{-2}\right\rangle}{k}\right) = \frac{1}{2}\left(k + \frac{\left(1+\varepsilon^2\right)^3}{k}\right)\,.$$

Solving $\partial_k\langle J\rangle = 0$ gives

$$k^*_{\text{exact}} = (1+\varepsilon^2)^{3/2} \quad\Longrightarrow\quad \langle J^*\rangle_{\text{exact}} = (1+\varepsilon^2)^{3/2}\,.$$

We see immediately that

$$\delta k^* = k^* - 1 = \frac{3}{2}\varepsilon^2 + O\left(\varepsilon^4\right),$$

which confirms that $\beta = 3/2$ calculation above. Similarly,

$$\delta J^* = J^* - 1 = \frac{3}{2}\varepsilon^2 + \frac{3}{8}\varepsilon^4 + O\left(\varepsilon^6\right).$$

The plot below shows that the perturbative approximation agrees well up to relative frequency variations of about 60%. The left plot shows the exact result for k (thin line), compared to the perturbative result in our formalism, thick dashed line. The right plot shows the optimal scaled cost function,

$$\langle J\rangle(\varepsilon, k) = \frac{1}{2}\left[k + \frac{1}{k}\left(1+\varepsilon^2\right)^3\right],$$

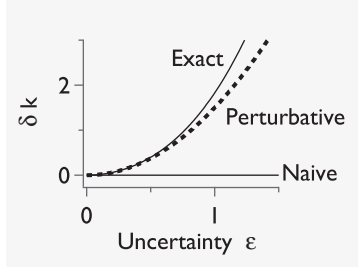
We use three different formulas for k, which is minimized either exactly or with approximations.

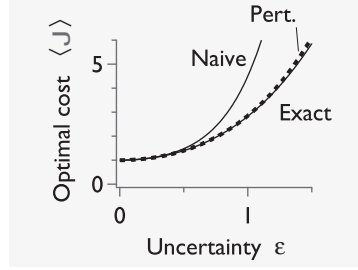
- Exact solution: $k^*_{\text{exact}} = \left(1+\varepsilon^2\right)^{3/2}$ and $\langle J^*\rangle_{\text{exact}}(\varepsilon) = \left(1+\varepsilon^2\right)^{3/2}$.
- Naive solution: $k^*_{\text{naive}} = 1$ (uncorrected) and $\langle J^*\rangle_{\text{naive}} = \frac{1}{2}[1 + (1+\varepsilon^2)^3]$.
- Perturbative solution: $k^*_{\text{pert}} = 1 + \frac{3}{2}\varepsilon^2$ and

$$\langle J^*\rangle_{\text{pert}} = \frac{1}{2}\left[1 + \frac{3}{2}\varepsilon^2 + \frac{(1+\varepsilon^2)^3}{1+\frac{3}{2}\varepsilon^2}\right].$$

As you can see, all three expressions agree to $O(\varepsilon^2)$. In the "naive" solution, we fix $k = 1$ and calculate $\langle J^*\rangle$. Amazingly, the perturbative expression, in this case, exceeds the exact calculation only at $O(\varepsilon^8)$, which is high order! Note the ordering: for a given ε,

$$\langle J^*\rangle_{\text{naive}} > \langle J^*\rangle_{\text{pert}} > \langle J^*\rangle_{\text{exact}}.$$





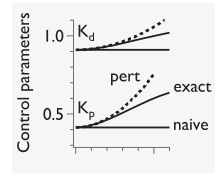
9.4 Harmonic oscillator with PD control. Redo Prob. 9.3 using PD control (two gains – proportional gain K_p and derivative gain K_d).

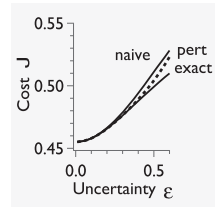
a. Find the solution $x(t)$ for the disturbance response. Evaluate the cost function J. Find the optimal values of the derivative gain K_d and proportional gain K_p.

b. Write out more explicitly Eq. (9.16) for the present case of two control parameters and one uncertain system parameter.

c. Find the scaled gain shifts, δk_d and δk_p as a function of ε^2.

d. Evaluate the cost function $\langle J\rangle(\varepsilon, K_\mathrm{d}, K_\mathrm{p})$ numerically by computing the expectation over the lognormal distribution for ω and minimizing over gains $K_\mathrm{d} = K_\mathrm{d}^*$ and $K_\mathrm{p} = K_\mathrm{p}^*$ to find $\langle J\rangle^*(\varepsilon)$. Plot and compare to the perturbative result. Do the same with the cost function and plot three curves (right): a *naive* control where you fix $K_\mathrm{d} = K_{\mathrm{d},0}$ and $K_\mathrm{p} = K_{\mathrm{p},0}$, the optimal values for the nominal model, a *perturbative* control using the techniques of this book, and an *exact* control using the gains determined using the numerically determined K_d^* and K_p^*.





Solution.

For proportional-derivative control of an undamped oscillator, the equations of motion are

$$\ddot{x} + K_\mathrm{d}\dot{x} + (K_\mathrm{p} + \omega^2)x = 0\,, \qquad x(0) = 0,\ \dot{x}(0) = 1\,,$$

where we have substituted the feedback derivative-control signal $u(t) = -K_\mathrm{p}x(t) - K_\mathrm{d}\dot{x}(t)$ into the system equation of motion.

a. The solution for $x(t)$ to the above initial-value problem is, for the underdamped case,

$$x(t) = \mathrm{e}^{-\zeta' t}\left(\frac{\sin\omega' t}{\omega'}\right), \qquad \omega' \equiv \sqrt{\omega^2 + K_\mathrm{p} - \zeta'^2}\,, \qquad \zeta' \equiv K_\mathrm{d}/2\,.$$

Putting this into the cost function gives

$$J(\omega, K_\mathrm{d}, K_\mathrm{p}) = \frac{1}{2}\int_0^\infty \mathrm{d}t\left[x^2(t) + u^2(t)\right] = \frac{K_\mathrm{d}}{4} + \frac{1 + K_\mathrm{p}^2}{4K_\mathrm{d}(K_\mathrm{p} + \omega^2)}\,,$$

Setting the gradient of J to zero gives

$$\partial_{\boldsymbol{K}} J = \begin{pmatrix}\partial_{K_\mathrm{d}}\\ \partial_{K_\mathrm{p}}\end{pmatrix} J = \begin{pmatrix}0\\0\end{pmatrix} \quad\Longrightarrow\quad \begin{pmatrix}K_{\mathrm{d},0}\\ K_{\mathrm{p},0}\end{pmatrix} = \begin{pmatrix}\sqrt{2\,(\sqrt{2}-1)}\\ \sqrt{2}-1\end{pmatrix} \approx \begin{pmatrix}0.91\\0.41\end{pmatrix}.$$

The optimal value of the cost function is $J_0 = J(K_{\mathrm{d},0}, K_{\mathrm{p},0}) = \sqrt{(\sqrt{2}-1)/2} \approx 0.46$.

We rescale J by defining

$$j(\omega, k_\mathrm{d}, d_\mathrm{p}) \equiv \frac{1}{J_0}J(\omega, k_\mathrm{d}K_{\mathrm{d},0}, k_\mathrm{p}K_{\mathrm{p},0}) = \frac{1}{2}\left[k_\mathrm{d} + \frac{1}{2k_\mathrm{d}}\left(\frac{\left(\sqrt{2}+1\right) + (\sqrt{2}-1)k_\mathrm{p}^2}{\left(\sqrt{2}-1\right)k_\mathrm{p} + \omega^2}\right)\right]$$

Then $j(1,1,1) = 1$, which simplifies further numerics.

b. The general formula for the corrections to the control parameter gains, Eq. (9.16), is given by

$$\delta\boldsymbol{K} = -\frac{1}{2}(\partial_{\boldsymbol{KK}}J)^{-1}\,\partial_{\boldsymbol{K}}\mathrm{Tr}\left[\boldsymbol{\Sigma}\,(\partial_{\boldsymbol{\theta\theta}}J)\right]\,.$$

For our present case of one uncertain parameter ω with nominal value = 1 and with two control parameters, we can write

$$\begin{pmatrix}\delta K_{\mathrm{d},0}\\ \delta K_{p,0}\end{pmatrix} = -\frac{1}{2}\begin{pmatrix}\partial_{kk}\, j & \partial_{kK_\mathrm{p}}\, j\\ \partial_{kK_\mathrm{p}}\, j & \partial_{K_\mathrm{p}K_\mathrm{p}}\, j\end{pmatrix}^{-1}\begin{pmatrix}\partial_{k\omega\omega} j\\ \partial_{K_\mathrm{p}\omega\omega} j\end{pmatrix}\varepsilon^2\,,$$

with all derivatives evaluated at $k = K_\mathrm{p} = \omega = 1$.

c. The Hessian matrix for the control parameters is

$$\begin{pmatrix}\partial_{k_\mathrm{d}k_\mathrm{d}} j & \partial_{k_\mathrm{d}k_\mathrm{p}} j\\ \partial_{k_\mathrm{d}k_\mathrm{p}} j & \partial_{k_\mathrm{p},k_\mathrm{p}} j\end{pmatrix}^{-1} = \begin{pmatrix}1 & 0\\ 0 & (2-\sqrt{2})/4\end{pmatrix}^{-1} = \begin{pmatrix}1 & 0\\ 0 & 2(2+\sqrt{2})\end{pmatrix}.$$

It is nice that the matrix happens to be diagonal! The gradient matrix is

$$\begin{pmatrix}\partial_{k_\mathrm{d}\omega\omega}\, j\\ \partial_{k_\mathrm{p}\omega\omega}\, j\end{pmatrix} = \begin{pmatrix}1/\sqrt{2}-2\\ 5/\sqrt{2}-9/2\end{pmatrix}.$$

Putting everything together gives

$$\begin{aligned}\begin{pmatrix}\delta k_\mathrm{d}\\ \delta k_\mathrm{p}\end{pmatrix} &= -\frac{1}{2}\begin{pmatrix}1 & 0\\ 0 & 2(2+\sqrt{2})\end{pmatrix}\begin{pmatrix}1/\sqrt{2}-2\\ 5/\sqrt{2}-9/2\end{pmatrix}\varepsilon^2\\ &= \begin{pmatrix}1-1/(2\sqrt{2})\\ 4-1/\sqrt{2}\end{pmatrix}\varepsilon^2\\ &\approx \begin{pmatrix}0.65\\ 3.3\end{pmatrix}\varepsilon^2\,.\end{aligned}$$

The result tells us how to modify each gain K_d and K_p relative to the nominal optimal control values $K_{\mathrm{d},0}$ and $K_{\mathrm{p},0}$.

9.5 Harmonic oscillator with uncertain frequency and damping. Analyze the disturbance response of a harmonic oscillator $\ddot{x} + 2\zeta\omega\dot{x} + \omega^2 x = u$, with $u = -K\dot{x}$, for two uncertain parameters $\omega \approx 1$ and $\zeta \approx 0.1$ and cost function $J = \int_0^\infty \mathrm{d}t\,(x^2+u^2)$.

a. Show that the nominal optimal control problem leads to $K_0 = J_0 \approx 0.82$.
b. Write out more explicitly Eq. (9.16) for the present case of one control parameter and two uncertain system parameters.
c. Find the shift in relative gain $\delta K/K$ in terms of $\sqrt{\langle\,\delta\omega^2\rangle}/\langle\omega\rangle$ and $\sqrt{\langle\,\delta\zeta^2\rangle}/\langle\zeta\rangle$. Are uncertainties in both parameters important?

Solution.

a. The solution to the equations of motion is, for the underdamped case,

$$x(t) = \mathrm{e}^{-\zeta' t}\left(\frac{\sin\omega' t}{\omega'}\right), \qquad \omega' \equiv \sqrt{\omega^2-\zeta'^2}\,, \qquad \zeta' \equiv \zeta + K/2\,.$$

The cost function J is given by

$$J(\omega, \zeta, K) = \int_0^\infty \mathrm{d}t \left[x^2(t) + u^2(t)\right] = \frac{1 + K^2\omega^2}{2(K + 2\zeta)\omega^2} \,.$$

Solving $\partial_K J = 0$ at the nominal values $\omega = 1$ and $\zeta = 0.1$ gives $K_0 = J_0 \approx 0.819804$.

The rescaled cost function

$$J(\omega, z, k) = J(\omega, z\,\zeta_0, k\,K_0) = \frac{1 + 0.672078k^2\omega^2}{(1.63961k + 0.4z\omega)\,\omega^2} \,.$$

In this problem, we had to rescale the uncertain system parameter ζ since its nominal value is 0.1. No such scaling is needed for ω, since its nominal value happens to be 1.

b. The general formula for the corrections to the control parameter gains, Eq. (9.16), is given by

$$\delta \boldsymbol{K} = -\frac{1}{2}(\partial_{\boldsymbol{K},\boldsymbol{K}} J)^{-1}\, \partial_{\boldsymbol{K}} \mathrm{Tr}\left[\boldsymbol{\Sigma}\,(\partial_{\boldsymbol{\theta},\boldsymbol{\theta}} J)\right] \,.$$

In our present case, we have one control parameter k and two uncertain system parameters ω and ζ, all scaled to have nominal values = 1. We will assume uncorrelated uncertainties for the unknown parameters, since the problem mentions that we know $\langle\, \delta\omega^2 \rangle$ and $\langle\, \delta\zeta^2 \rangle$ but does not mention the cross-correlation $\langle\, \delta\omega\, \delta\zeta \rangle$. The trace term then becomes simply $(\partial_{\omega,\omega} J)\langle\, \delta\omega^2 \rangle + (\partial_{\zeta,\zeta} J)\langle\, \delta\zeta^2 \rangle$. Thus,

$$\delta k = \beta_\omega \left\langle\, \delta\omega^2 \right\rangle + \beta_\zeta \left\langle\, \delta\zeta^2 \right\rangle \,,$$

where

$$\beta_\omega \equiv -\frac{1}{2}\left(\frac{\partial_{k\omega\omega} J}{\partial_{kk} J}\right) \,, \qquad \beta_\zeta \equiv -\frac{1}{2}\left(\frac{\partial_{k\zeta\zeta} J}{\partial_{kk} J}\right) \,,$$

with all derivatives evaluated at $\omega = \zeta = k = 1$.

c. Evaluating the derivatives for the two β coefficients using a computer-algebra program gives

$$\beta_\omega \approx 2.34025 \,, \qquad \beta_\zeta \approx 0.0769231 \,.$$

Since $\beta_\omega/\beta_\zeta \approx 30$, the solution is more sensitive to uncertainty in ω. This seems intuitive, as the gain K directly alters the damping ζ but only indirectly the frequency. Probably, it would be enough to account for the uncertainty in frequency alone.

Note that because the unknown damping must still be positive (we know the uncontrolled system is stable), we would again take a lognormal distribution for ζ, and we know a priori that our control has no danger of destabilizing the system.

9.6 **One-step LQR**. The problem considered in Section 9.3.3 can be solved analytically.

a. One striking feature of the probability distributions is that they are zero when the cost $J < J_{\min}$. Explain why (without calculation) and find $J_{\min}(u)$.
b. By changing variables, transform the normal distribution for ν and show that

$$p(J) = \left(\frac{1}{\sigma^2}\right) \frac{1}{\sqrt{\pi(\,\delta J)}}\, \mathrm{e}^{-\left(\,\delta J + \frac{1}{2}(\,\delta x)^2\right)} \cosh\left(\sqrt{2(\,\delta J)}\,\delta x\right) \theta(\,\delta J)\,,$$

where $\theta(\cdot)$ is the step function, $\delta J = (J - J_{\min})/\sigma^2$, and $\delta x = (x-u)/\sigma$.
c. Find analytic expressions for the mean $\langle J\rangle$, standard deviation σ_J, and tail probability $P(J > J_{\max})$. Can you find a simpler approximation to this last quantity?

Solution.

a. The cost function

$$J(\nu) = \tfrac{1}{2}\left((x - u + \nu)^2 + Ru^2\right)$$

is minimized by a lucky fluctuation $\nu = -(x-u)$ that makes the first term zero. The minimal cost is thus $J_{\min} = \frac{1}{2}Ru^2$.
b. The noise ν obeys a normal distribution of mean zero and variance σ^2:

$$p(\nu) = \frac{1}{\sqrt{2\pi}\sigma}\, \mathrm{e}^{-\frac{\nu^2}{2\sigma^2}}\,,$$

Following the logic of Example A.17, we write

$$p(J) = \frac{p\,(\nu_+)}{|J'(\nu_+)|} + \frac{p\,(\nu_-)}{|J'(\nu_-)|}\,,$$

where $\nu_\pm(J)$ is the inverse function of $J(\nu)$ and $J'(\nu) = \mathrm{d}J/\mathrm{d}\nu$. Notice that there are two roots to the inverse, and the pdf $p(J)$ picks up contributions from both. Inverting the relationship $J(\nu)$ gives,

$$\nu_\pm = -(x-u) \pm \sqrt{2(J - J_{\min})}\,,$$

and

$$|J'(\nu^\pm)| = |x - u + \nu_\pm| = |\pm\sqrt{2(J - J_{\min})}| = \sqrt{2(J - J_{\min})}\,.$$

Putting all the pieces together, we have

$$p(J) = \frac{1}{\sqrt{2\pi}\sigma\sqrt{2(J - J_{\min})}}\left(\mathrm{e}^{-\frac{\nu_-^2}{2\sigma^2}} + \mathrm{e}^{-\frac{\nu_+^2}{2\sigma^2}}\right).$$

Substituting for $\nu_\pm$ and recalling that $2\cosh x = \mathrm{e}^x + \mathrm{e}^{-x}$ and then defining δJ and δx leads directly to the expressions given in the main text.

c. The mean was already given in the main text:

$$\langle J\rangle = \frac{1}{2}\left((x-u)^2 + Ru^2 + \sigma^2\right),$$

which is minimized by choosing

$$u^* = \left(\frac{1}{1+R}\right)x \quad \Longrightarrow \quad \langle J\rangle_{\min} = \frac{x^2}{2}\left(\frac{R}{1+R}\right) + \frac{1}{2}\sigma^2 .$$

For the variance, we first define

$$\begin{aligned}\delta J = J - \langle J\rangle &= \frac{1}{2}\left((x-u+\nu)^2 + Ru^2\right) - \frac{1}{2}\left((x-u)^2 + Ru^2 + \sigma^2\right)\\ &= \sigma(x-u) + \tfrac{1}{2}\left(\nu^2 - \sigma^2\right)\end{aligned}$$

and, using $\langle \nu^4\rangle = 3\sigma^4$ for a Gaussian variable of mean zero, we have

$$\begin{aligned}\sigma_J^2 \equiv \left\langle (\delta J)^2\right\rangle &= \sigma^2(x-u)^2 + \tfrac{1}{4}\left(\left\langle \nu^4\right\rangle - 2\sigma^2\sigma^2 + \sigma^4\right)\\ &= \sigma^2(x-u)^2 + \tfrac{1}{2}\sigma^4 .\end{aligned}$$

Note that the cross term is $\sim \left\langle \sigma(x-u)(\nu^2-\sigma^2)\right\rangle = \sigma(x-u)\langle \nu^2-\sigma^2\rangle = 0$. The expression for σ_J is minimized for $u = x$, where $\sigma_J = \sigma^2/\sqrt{2}$.

The last quantity to calculate (using Mathematica) is

$$\begin{aligned}P(J > J_{\max}) &= \int_{J_{\max}}^{\infty} \mathrm{d}J\, p(J)\\ &= \tfrac{1}{2}\left[\operatorname{erfc}\left(\sqrt{\delta J_{\max}} - (\delta x)/\sqrt{2}\right) + \operatorname{erfc}\left(\sqrt{\delta J_{\max}} + (\delta x)/\sqrt{2}\right)\right],\end{aligned}$$

where $J_{\max}/\sigma^2 = \delta J_{\max} + J_{\min}/\sigma^2$.

To simplify this expression for $P(J > J_{\max})$, we note that the second term is negligible for almost all values of its arguments. (When $\delta x = 0$, or $u = x$, there is an error of a factor of 2. But for smaller u, the term is truly small.) Thus,

$$P(J > J_{\max}) \approx \tfrac{1}{2}\left[\operatorname{erfc}\left(\sqrt{\delta J_{\max}} - (\delta x)/\sqrt{2}\right)\right].$$

For large values of the argument, this simplifies even more: $\operatorname{erfc}(x) \sim \mathrm{e}^{-x^2}/(\sqrt{\pi}x)$.

9.7 Harmonic oscillator statistics. We analyze the cost distribution $p(J)$ for an undamped harmonic oscillator with uncertain frequency, continuing Problem 9.3.

a. Derive the analytic form for $p(J)$ and plot for $K = 1.4, 2, 3$.
b. Derive analytic expressions for $\langle J\rangle$, σ_J, and the tail probability $p(J > J_{\max})$.

c. Show that the gain $K^{**} = 2J_{\max}$ minimizes the tail probability.
d. Plot $p(J)$ for $K = 1.4, 2, 3$; mean $\langle J \rangle$ and σ_J versus gain K; gain K^* that minimizes $\langle J \rangle$ versus ε; and tail probability $P(J > J_{\max} = 1)$ versus K.

Solution.

a. As usual, we derive $p(J)$ by change of variables. We invert $J(\omega)$ to $\omega(J)$:

$$J(\omega) = \tfrac{1}{4}\left(K + \frac{1}{\omega^2 K}\right) \quad \Longrightarrow \quad \omega_\pm(J) = \pm\frac{1}{\sqrt{K(4J-K)}}\,.$$

Although there are formally two roots, $\omega_\pm$, only $\omega_+ = 1/\sqrt{K(4J-K)}$ is relevant, as the support is $\omega \in [0, \infty)$. We also need the Jacobian:

$$|J'(\omega_+)| = \left|\frac{-2}{4K\omega_+^3}\right| = \frac{1}{2K\omega_+^3}\,.$$

Then, using the standard form for a lognormal distribution, we have

$$\begin{aligned} p(J) = \frac{p(\omega_+)}{|J'(\omega_+)|} &= \left(\frac{2K\omega_+^3}{\omega_+\sigma\sqrt{2\pi}}\right)\exp\left(-\frac{(\ln\omega_+ - \mu)^2}{2\sigma^2}\right) \\ &= \sqrt{\frac{2}{\pi}}\left(\frac{K\omega_+^2}{\sigma}\right)\exp\left(-\frac{(\ln\omega_+ - \mu)^2}{2\sigma^2}\right). \end{aligned}$$

We can further substitute for ω_+, $\mu = -\frac{1}{2}\ln\left(1+\varepsilon^2\right)$, and $\sigma^2 = -2\mu$, but the overall expressions remain complicated. So we stop here.

b. The result for the mean is derived for scaled units in Problem 9.3c. It is based on the identity

$$\left\langle \omega^{-2} \right\rangle = \mathrm{e}^{-2\mu + (4\sigma^2)/2} = \mathrm{e}^{-6\mu} = \left(1+\varepsilon^2\right)^3\,.$$

Thus,

$$\langle J \rangle = \frac{1}{4}\left(K + \frac{1}{K}\left\langle \omega^{-2} \right\rangle\right) = \frac{1}{4}\left[K + \frac{1}{K}\left(1+\varepsilon^2\right)^3\right],$$

which is minimized for $K^* = \left(1+\varepsilon^2\right)^{3/2} \approx 1.398$ for $\varepsilon = 0.5$. The minimum mean cost is $\langle J \rangle^* = \frac{1}{2}K^*$.

To calculate the variance σ_J^2, we first define

$$\delta J = J - \langle J \rangle = \frac{1}{4K}\left(\frac{1}{\omega^2} - (1+\varepsilon^2)^3\right),$$

so that

$$\begin{aligned} \sigma_J^2 = \langle (\,\delta J)^2 \rangle &= \frac{1}{16K^2}\left\langle \left(\omega^{-2} - (1+\varepsilon^2)^3\right)^2 \right\rangle \\ &= \frac{1}{16K^2}\left[\left\langle \omega^{-4} \right\rangle - 2\left\langle \omega^{-2} \right\rangle (1+\varepsilon^2)^3 + (1+\varepsilon^2)^6\right] \\ &= \frac{1}{16K^2}\left[(1+\varepsilon^2)^{10} - (1+\varepsilon^2)^6\right], \end{aligned}$$

where we use the moment identity described above, $\langle\omega^{-4}\rangle = (1+\varepsilon^2)^{10}$. Thus, the standard deviation is

$$\sigma_J = \frac{(1+\varepsilon^2)^3}{4K}\sqrt{(1+\varepsilon^2)^4 - 1}\,.$$

Notice that $\sigma_J \sim 1/K$, which always decreases with K. For small uncertainty in frequency, $\sigma_J \approx \varepsilon/(2K)$.

To calculate the tail probability $P(J > J_{\max})$, we need the complement of the cumulative distribution function. Since J is monotonic in ω, we transform the cumulative distribution function $F_\omega(\omega)$ into $F_J(J)$ as follows:

$$F_J(J) = \int_J^\infty \mathrm{d}J'\, p_J(J') = \int_\omega^\infty p_\omega(\omega') = F_\omega(\omega)\,.$$

Thus, $F_\omega(\omega) = F_\omega[\omega(J)]$. Since $\omega \sim \mathcal{N}(\mu, \sigma^2)$, we have

$$F_\omega[\omega(J)] = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{\ln\omega - \mu}{\sqrt{2}\sigma}\right).$$

The tail probability is the complement:

$$P(J > J_{\max}) = 1 - F_\omega[\omega(J)] = \tfrac{1}{2}\mathrm{erfc}\left(\frac{\ln\omega - \mu}{\sqrt{2}\sigma}\right).$$

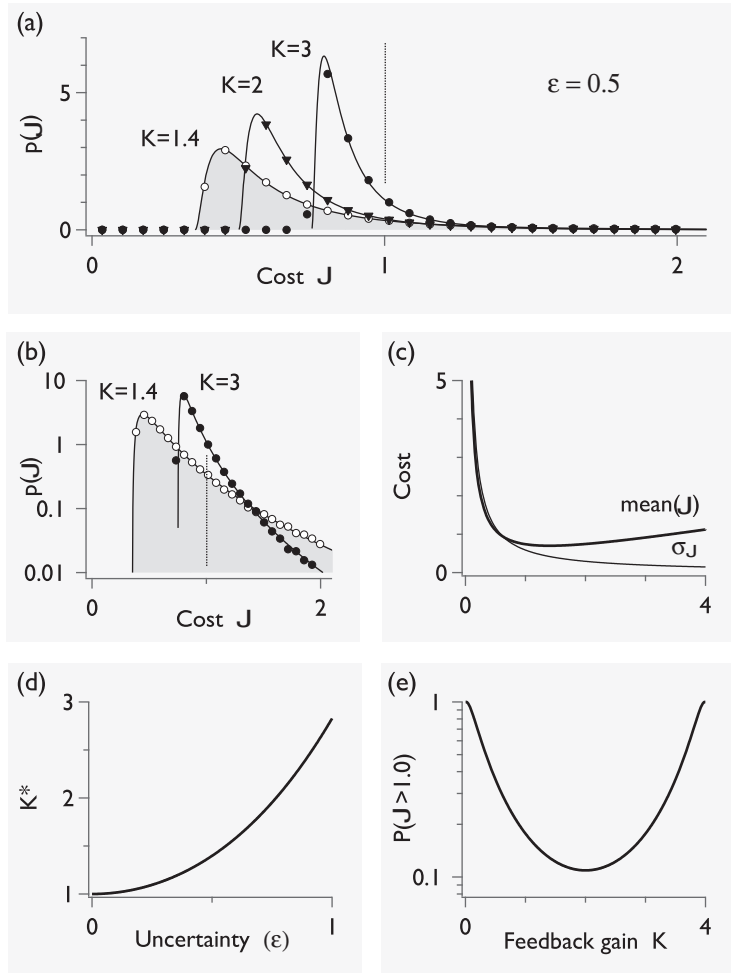
We then substitute for $\omega_+(J) = 1/\sqrt{K(4J - K)}$.

c. To find the minimum of the tail probability $P(J > J_{\max})$, we differentiate with respect to K. The problem simplifies enormously when we realize that the factors containing K all vary monotonically with K, except the "core" part of $\omega'(K)$, which is $\omega \sim [K(J - K/4)]^{-1/2}$.

$$\frac{\mathrm{d}\omega}{\mathrm{d}K} = -\frac{1}{2}(\cdots)^{-3/2}(J - K/2)|_{J=J_{\max}} = 0\,,$$

which implies $K^{**} = 2J_{\max}$.

d. Below, we plot the cost distribution $p(J)$ for an undamped harmonic oscillator with uncertain frequency $p(\omega)$ satisfying $\langle\omega\rangle = 1$ and $\varepsilon = 0.5$. (a) $p(J)$ for three gains. $K \approx 1.4$ minimizes $\langle J\rangle$. $K = 2$ minimizes $P(J > 1)$. $K = 3$ has lower variance but higher mean. (b) Same, on log scale, with $K = 2$ hidden for clarity. (c) Mean and standard deviation, vs. gain. (d) Gain that minimizes mean, as function of oscillator frequency uncertainty ε. (e) Tail probability for $J_{\max} = 1$.

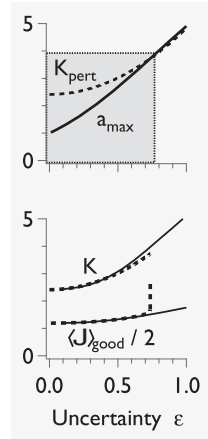


9.8 **Unstable first-order system**. Consider an unstable first-order system,

$$\dot{x} = ax + u\,, \qquad u = -Kx\,, \qquad x(0) = 1\,, \qquad J = \int_0^\infty \mathrm{d}t \left(x^2 + u^2\right). \tag{9.1}$$

The closed-loop system, $\dot{x} + (K - a)x = 0$, is stable for feedback gain $K > a$. Now assume an uncertain a with lognormal distribution, as in Problem 9.3: $\ln a \sim \mathcal{N}(\mu, \sigma^2)$, with $\mu = -\frac{1}{2}\ln\left(1 + \varepsilon^2\right)$ and $\sigma = \sqrt{-2\mu}$. Here, ε^2 is the variance of a, and $\langle a \rangle = 1$.

a. Show that the optimal control is $K_0 = \sqrt{2} + 1$ and that $J_0(K_0) = K_0$, for $a = 1$.
b. Use perturbation theory to find the optimal feedback gain to $O(\varepsilon^2)$, ignoring the possibility of instability. Show that $\beta = 1$ and, thus, $K^*(\varepsilon) = K_0(1 + \varepsilon^2)$.
c. For $\alpha = 0.01$, show that perturbation theory is limited to $\varepsilon < \varepsilon_{\max} \approx 0.74$.
d. Minimize $\langle J \rangle_{\text{good}}(K)$ for $\alpha = 0.01$ and $\varepsilon < \varepsilon_{\max}$. Plot $K(\varepsilon)$ and $\langle J \rangle_{\text{good}}$ versus ε for both the numerical minimization and the perturbation theory.



For the graphs at right, dashed quantities are from the perturbation theory. Intuitively, to stabilize systems up to $a = a_{\max}$, we need to set the gain K a "little bit" higher than $a_{\max}$. At $\varepsilon = \varepsilon_{\max}$, for example, $a_{\max} \approx 3.7$ for $\alpha = 0.01$. The perturbation-determined gain $K_{\text{pert}} \approx a_{\max}$ is too low and leads to the divergence of average cost. The numerically determined value that optimizes $\langle J \rangle_{\text{good}} \approx 3.0$ is $K^* \approx 4.0$.

Solution.

a. For fixed, known a, we can choose $K > a$, and the problem is defined by

$$\dot{x} - ax = u\,, \qquad u = -Kx\,, \qquad x(0) = 1\,, \qquad J = \int_0^\infty \mathrm{d}t \left(x^2 + u^2\right)\,.$$

The solution to the equations of motion is $x(t) = \mathrm{e}^{(a-K)t}$, which leads to an integrated cost of

$$J = \left(1 + K^2\right) \int_0^\infty \mathrm{d}t \underbrace{\mathrm{e}^{2(a-K)t}}_{x^2} = \frac{1}{2}\left(\frac{1+K^2}{K-a}\right)\,.$$

Differentiating and setting $\partial_K J = 0$ implies $K_0 = J_0 = a + \sqrt{1+a^2}$.
For $a = 1$, this gives $K_0 = J_0 = \sqrt{2} + 1 \approx 2.4$.
Note that this part of the problem repeats a calculation done for a slightly more general cost function in Problem 7.2.

b. The scaled cost function is

$$j(k,a) = \left(\sqrt{2} - 1\right) \frac{1 + \left(3 + 2\sqrt{2}\right)k^2}{\left(\sqrt{2}+1\right)k - a}\,,$$

which satisfies $j(1,1) = 1$. Evaluating derivatives of the cost function leads to

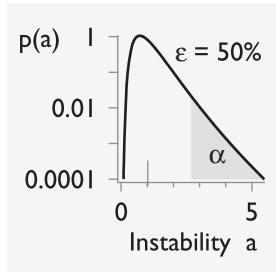
$$\beta = -\frac{1}{2}\left(\frac{\partial_{k,a,a} j}{\partial_{k,k} j}\right)_{a=k=1} = 1\,.$$

Thus, $\delta k = (K^* - K_0)/K_0 = \varepsilon^2 + O(\varepsilon^4)$.

c. The perturbation theory tells to choose $K^*(\varepsilon) = K_0(1 + \beta\varepsilon^2)$, with $K_0 = \sqrt{2}+1$ and $\beta = 1$. However, stability requires $K^* > a$. With an unbounded $p(a)$, some fraction of systems will be unstable and thus "fail." If we start with a prescription that we need to choose K so that only a defined fraction α of failures will occur, we can have a conflict with perturbation theory. We visualize this issue by plotting $a_{\max}$ and K^* as a function of ε and looking for the crossing point. Note that $a_{\max}$ is defined as the *inverse survival function* (or inverse to the complementary cumulative distribution),

$$\alpha = \int_{a_{\max}}^\infty \mathrm{d}a\, p(a,\varepsilon)\,,$$

with $p(a,\varepsilon)$ the lognormal distribution with $\langle a \rangle = 1$ and variance ε^2. The graph below illustrates $\alpha = 0.01$ for $\varepsilon = 0.5$.



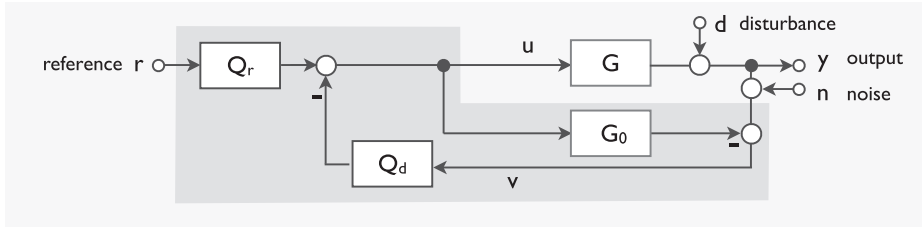
d. The cost for feedback control of the "good" systems is given by

$$J_{\text{good}}(K,\varepsilon) = \int_0^{a_{\max}} \mathrm{d}a\, p(a)\, J(a,K)$$
$$= \int_0^{a_{\max}} \mathrm{d}a\, \frac{1}{\sqrt{2\pi}\sigma a}\, \mathrm{e}^{-\frac{(\mu-\ln a)^2}{2\sigma^2}} \left(\frac{1+K^2}{2}\right) \frac{1}{K-a}\,,$$

with $\mu = -\frac{1}{2}\ln\left(1+\varepsilon^2\right)$ and $\sigma = \sqrt{-2\mu}$. The upper limit $a_{\max}$ is chosen as a function of ε as described in (c).

It is straightforward to minimize this function numerically (over K for fixed ε). The starting point for an iterative solution can either be the perturbative solution or the solution found numerically for the previous, neighboring value of ε. Here, there is a single, global minimum, and the solution is easily found.

9.9 Internal Model Control (IMC) and feedforward. Consider the "two degrees of freedom" variant of IMC shown below, with system model and two controller transfer functions Q_d and Q_r. The latter is the feedforward filter defined in Section 3.4.1.



Show that the transfer function of the error signal $e = r - y$ is given by

$$e = \left(1 - \frac{GQ_r}{1+Q_d(G-G_0)}\right) r - \left(\frac{1-G_0Q_d}{1+Q_d(G-G_0)}\right) d + \left(\frac{GQ_d}{1+Q_d(G-G_0)}\right) n\,.$$

For a perfect model, $G_0 = G$, this reduces to $e = (1-GQ_r)\,r - (1-GQ_d)\,d + GQ_d\,n$, which shows very clearly the role of Q_r in tracking the reference and Q_d in rejecting disturbances. Without feedforward ($Q_r = 1$), we cannot, in general, do both. We also see that rejecting disturbances generally adds noise to the error signal.

Solution.

From the block diagram,

$$v = (G - G_0)u + d + n\,, \qquad u = Q_r r - Q_d v\,, \qquad y = Gu + d\,.$$

Then,

$$u = Q_r r - Q_d[(G - G_0)u + d + n] = Q_r r - Q_d d - Q_d n - Q_d(G - G_0)u\,.$$

Solving for u gives

$$u = \frac{Q_r r - Q_d d - Q_d n}{1 + Q_d(G - G_0)}\,.$$

The error is

$$\begin{aligned} e = r - y = r - Gu - d &= r - \frac{GQ_r r - GQ_d d - GQ_d n}{1 + Q_d(G - G_0)} - d \\ &= \left(1 - \frac{GQ_r}{1 + Q_d(G - G_0)}\right) r - \left(1 - \frac{GQ_d}{1 + Q_d(G - G_0)}\right) d + \left(\frac{GQ_d}{1 + Q_d(G - G_0)}\right) n \\ &= \left(1 - \frac{GQ_r}{1 + Q_d(G - G_0)}\right) r - \left(\frac{1 - G_0 Q_d}{1 + Q_d(G - G_0)}\right) d + \left(\frac{GQ_d}{1 + Q_d(G - G_0)}\right) n\,. \end{aligned}$$

For a perfect model, $G_0 = G$, we have then

$$e = (1 - GQ_r)\, r - (1 - GQ_d)\, d + Q_d\, n\,.$$

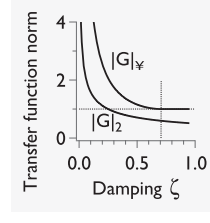
Choosing $Q_r = Q_d = G^{-1}$ would thus eliminate errors in the tracking due to the reference r and disturbances d, although not to measurement noise n. But, as we have seen repeatedly, it will not, in any case, be possible to perfectly invert a system. We recall the main limitations: the inverse of G can be acausal or unstable; the required inputs may be too big, $K = 1/(1 - GQ_d) \to \infty$; and, of course, we may not have a perfect model, $G_0 \neq G$. In some limits, for example at low frequencies, good approximations can be feasible.

9.10 **Transfer function norms**. Define the 2 and ∞-norms of a transfer function $G(\mathrm{i}\omega)$ as $\|G\|_2 \equiv \left[\int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} |G(\mathrm{i}\omega)|^2\right]^{1/2}$ and $\|G\|_\infty \equiv \sup_\omega |G(\mathrm{i}\omega)|$.

a. *2-norm*. Using Parseval's theorem (Problem A.4.3) and representing $G(s)$ in statespace via $\{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}\}$, show that $\|G\|_2 = \sqrt{\boldsymbol{CPC}^\mathsf{T}}$, where $\boldsymbol{P} = \int_0^\infty \mathrm{d}t\, \mathrm{e}^{\boldsymbol{A}t}\, \boldsymbol{B}\, \boldsymbol{B}^\mathsf{T}\, \mathrm{e}^{\boldsymbol{A}^\mathsf{T} t}$ is the Gramian matrix introduced in Example 4.4. Recall that you can compute $\boldsymbol{P}$ directly or solve the Lyapunov equation, $\boldsymbol{AP} + \boldsymbol{PA}^\mathsf{T} = -\boldsymbol{BB}^\mathsf{T}$ (Problem 2.15).

b. For $G(s) = \frac{1}{1+\tau s}$, show that $\|G\|_2 = 1/\sqrt{2\tau}$ and $\|G\|_\infty = 1$.

c. Show that the 2- and ∞-norms for $G(s) = \frac{1}{1+2\zeta s+s^2}$ give the curves at right.



Solution.

a. From Parseval's theorem, the square of the 2-norm is given by

$$\|G\|_2^2 = \int_{-\infty}^{\infty} \mathrm{d}t\, |G(t)|^2\,.$$

But $G(t)$ is the response function to an impulse $\delta(t)$ at time $t = 0$. From Eq. (2.66) and using causality, it is given by

$$G(t) = \begin{cases} \boldsymbol{C}\, \mathrm{e}^{\boldsymbol{A}t}\, \boldsymbol{B} & t > 0 \\ 0 & t \le 0 \end{cases} .$$

Thus,

$$\|G\|_2^2 = \int_0^\infty \mathrm{d}t\, \boldsymbol{C}\, \mathrm{e}^{\boldsymbol{A}t}\, \boldsymbol{B}\, \boldsymbol{B}^\mathsf{T}\, \mathrm{e}^{\boldsymbol{A}^\mathsf{T}t}\, \boldsymbol{C}^\mathsf{T} = \boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^\mathsf{T} .$$

b. For the first-order system $G(s) = \frac{1}{1+\tau s}$, the state-space representation has $A = -1/\tau$, $B = 1$, and $C = 1/\tau$. Then the Gramian is

$$P = \int_0^\infty \mathrm{d}t\, \mathrm{e}^{-t/\tau}(1)(1)\, \mathrm{e}^{-t/\tau} = \int_0^\infty \mathrm{d}t\, \mathrm{e}^{-2t/\tau} = \frac{\tau}{2} .$$

The 2-norm is then

$$\|G\|_2 = \sqrt{\left(\frac{1}{\tau}\right)\left(\frac{\tau}{2}\right)\left(\frac{1}{\tau}\right)} = \frac{1}{\sqrt{2\tau}} .$$

Another way to calculate $\|G\|_2$ is to do a contour integral in the complex s-plane (Doyle et al., 1992).

To calculate $\|G\|_\infty$, we note that the magnitude is

$$|G(\mathrm{i}\omega)| = \frac{1}{\sqrt{1 + \tau^2\omega^2}} ,$$

which clearly has a maximum value =1 (for $\omega = 0$).

c. For the second-order system $G(s) = \frac{1}{1+2\zeta s+s^2}$ and referring to Eq. (2.17), the state-space representation has

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ -1 & -2\zeta \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \boldsymbol{C} = \begin{pmatrix} 1 & 0 \end{pmatrix} .$$

We can find the Gramian $\boldsymbol{P} = \boldsymbol{P}^\mathsf{T}$ by solving the Lyapunov equation:

$$\begin{aligned} \boldsymbol{A}\boldsymbol{P} + \boldsymbol{P}\boldsymbol{A}^\mathsf{T} + \boldsymbol{B}\boldsymbol{B}^\mathsf{T} &= \begin{pmatrix} 0 & 1 \\ -1 & -2\zeta \end{pmatrix}\begin{pmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{pmatrix} + \begin{pmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & -2\zeta \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 2p_{12} & -p_{11} + p_{22} - 2p_{12}\zeta \\ -p_{11} + p_{22} - 2p_{12}\zeta & -2p_{12} - 4p_{22}\zeta + 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} , \end{aligned}$$

which leads to $p_{11} = p_{22} = \frac{1}{4\zeta}$ and $p_{12} = 0$. Thus,

$$\|G\|_2 = \sqrt{\frac{1}{4\zeta}\begin{pmatrix} 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix}} = \frac{1}{2\sqrt{\zeta}} .$$

For the ∞-norm, we take $\frac{\mathrm{d}}{\mathrm{d}\omega}$. It is simpler algebraically to differentiate the the magnitude squared of $|G|^2$, which is given by

$$|G(\mathrm{i}\omega)|^2 = \left|\frac{1}{1 + 2\zeta\,\mathrm{i}\omega - \omega^2}\right|^2 = \frac{1}{(1-\omega^2)^2 + 4\zeta^2\omega^2}\,,$$

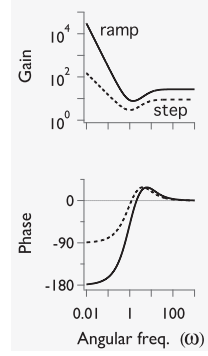
This condition implies

$$\frac{\mathrm{d}}{\mathrm{d}\omega}|G(\mathrm{i}\omega)|^2 = 0 \qquad \Longrightarrow \qquad (\omega^*)^2 = \{0,\ 1 - 2\zeta^2\}\,,$$

which implies, being careful to select the physical root ω^* that gives a maximum,

$$\|G\|_\infty = |G(\mathrm{i}\omega^*)| = \begin{cases} \frac{1}{2\zeta\sqrt{1-\zeta^2}} & \zeta < \frac{1}{\sqrt{2}} \\ 1 & \zeta \geq \frac{1}{\sqrt{2}} \end{cases}\,.$$

Note that for $\zeta \geq 1/\sqrt{2}$, we evaluate $|G|$ at the $\omega = 0$ root. The two norms are plotted against ζ in the main text.



9.11 Tracking a ramp. For the system discussed in Example 9.1, use Internal Model Control to design a controller to track a ramp. Put all closed-loop poles at $s = -1/\tau$, and choose $\tau = 1/3$ for plots. Make a time-domain plot to show that the output does track a ramp. Make a Bode plot for $K(s)$. At right we add a dashed line showing the controller from Example 9.1, which tracks a step input but not a ramp.

Solution.

The solution follows Example 9.1 closely and is a more sophisticated version of Example 3.7. The system is

$$G = \frac{1}{(1+s)^2}\,.$$

We want the sensitivity function S to have the inverse model, i.e., s^2 for tracking a ramp, whose Laplace-domain signal is $r(s) \sim 1/s^2$. We also want it to replace the two poles at -1 with poles at $-1/\tau$. We can do this with an IMC function $Q(s)$

$$Q = \frac{(1+s)^2(a+bs)}{(1+s\tau)^3}\,,$$

where a and b are coefficients that need to be determined. The complementary sensitivity function $T(s)$ is then

$$T(s) = Q(s)\,G(s) = \frac{a+bs}{(1+s\tau)^3}\,,$$

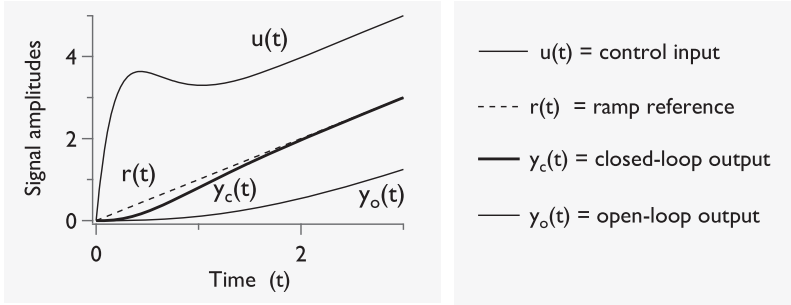
which implies

$$S(s) = 1 - T = \frac{\tau^3 s^3 + 3\tau^2 s^2 + (3\tau - b)s + (a-1)}{(1+s\tau)^3} \quad \rightarrow \quad \frac{(s\tau)^2(3+s\tau)}{(1+s\tau)^3},$$

where we have taken $b = 3\tau$ and $a = 1$. Finally, the controller $K(s)$ is given by

$$K(s) = \frac{Q}{S} = \frac{(1+s)^2(1+3s\tau)}{(s\tau)^2(3+s\tau)}.$$

which is plotted in the main text, along with the controller from Example 9.1. Note that for low frequencies, $K(s) \sim s^{-2}$, showing it will track a ramp. (Compare $K \sim s^{-1}$ in Example 9.1, which tracked a step but not a ramp.) The controller is biproper and thus realizable: $K(s) \to 3/\tau^2$ as $s \to \infty$. The response is shown below.



9.12 Norms and transfer functions. Why is a transfer function with finite ∞-norm proper but one with finite 2-norm strictly proper?

Solution.

Let us start with the 2-norm,

$$\|G\|_2 \equiv \left[\int_{-\infty}^{\infty} \frac{d\omega}{2\pi} |G(i\omega)|^2\right]^{1/2}.$$

If a transfer function $G(s)$ does not vanish at infinity, then clearly the integral diverges. Being strictly proper is thus necessary for having finite 2-norm.

Now consider the ∞-norm,

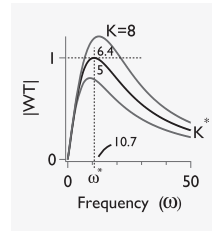
$$\|G\|_\infty \equiv \sup_\omega |G(i\omega)|.$$

If $G(s \to i\infty)$ is constant, then either it will determine the ∞-norm or a finite-frequency maximum will. In either case, if it has a finite norm, it will not diverge at $\omega = \infty$ and hence is proper.

9.13 First-order system with uncertain delay. Consider a nominal $G_0(s) = \frac{1}{1+s} e^{-st_{\max}/2}$ and multiplicative uncertainty $G(s) = G_0[1 + \Delta(s)W(s)]$, with $W = \frac{2.1 s t_{\max}}{1+s t_{\max}}$ and $|\Delta| \le 1$. The controller $K(s) = K$, for $t \in [0, t_{\max}]$, with $t_{\max} = 0.1$ and $\tau = 1$.

a. From the robust stability limit $\|WT\|_\infty = 1$, find (numerically) the maximum allowable gain for which stability is guaranteed. Hint: Find the frequency ω^* and gain $K_{\max}$ such that $|WT| = 1$ and $\frac{d}{d\omega}|WT| = 0$. Here, $T = \frac{L}{1+L}$ and $L = KG$.

b. Plot $K(\omega)$ to confirm the values of $K_{\max}$ and ω^* (right). Next, make a Bode magnitude plot of $T(\mathrm{i}\omega)$ and $W^{-1}(\mathrm{i}\omega)$. Finally, plot in the complex plane the circles of possible loop transfer functions $L(\mathrm{i}\omega)$ for a few different frequencies and a gain $K_{\max} \approx 6.4$. Highlight the circular domain corresponding to $K_{\max}$ and ω^*.



c. Calculate $K_{\max}$ for known delays $\frac{1}{2}t_{\max}$ and $t_{\max}$.

Solution.

a. The condition $\|WT\|_\infty = 1$ means that we seek the highest controller gain $K_{\max}$ such that the magnitude $|W(\mathrm{i}\omega)\,T(\mathrm{i}\omega)| \equiv f(K,\omega)$ has a maximum at some ω^* that is just equal to one. With

$$G_0(s) = \frac{\mathrm{e}^{-st_{\max}/2}}{1+s\tau}\,, \qquad K(s) = K\,,$$

the complementary sensitivity function is

$$T(s) = \frac{KG}{1+KG} = \frac{K}{K+\mathrm{e}^{st_{\max}/2}(1+s\tau)}\,.$$

Using Mathematica to write out the explicit expression for the magnitude $|WT| \equiv f(K,\omega)$, we have

$$f(K,\omega) \equiv \frac{2.1K\,\omega\,t_{\max}}{\sqrt{\left(\omega^2 t_{\max}^2+1\right)}}\,\frac{1}{\sqrt{\left(K^2-2K\tau\omega\sin\left(\frac{\omega t_{\max}}{2}\right)+2K\cos\left(\frac{\omega t_{\max}}{2}\right)+\tau^2\omega^2+1\right)}}\,.$$

We then seek the lowest $K_{\max} > 0$ such that

$$f(K_{\max},\omega^*) = 1\,, \qquad \frac{\partial f}{\partial\omega}(K_{\max},\omega^*) = 0\,.$$

Numerically solving in Mathematica then gives $\omega^* \approx 10.678$ and $K_{\max} \approx 6.446$.

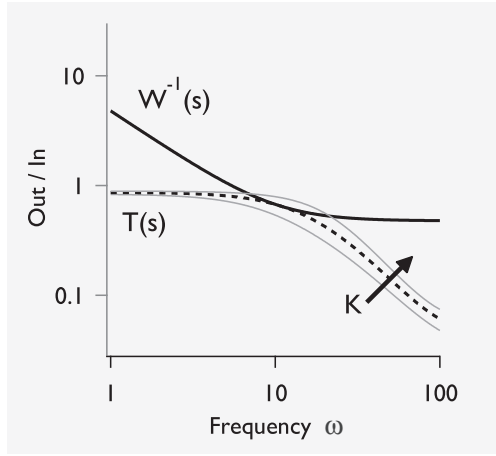
b. The inverse bound is

$$W^{-1}(s) = \frac{1+st_{\max}}{2.1st_{\max}}\,.$$

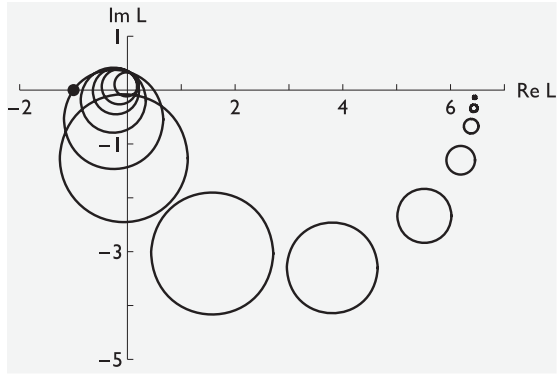
The closed-loop transfer function (complementary sensitivity function) is

$$T(s) = \frac{KG}{1+KG} = \frac{K\,\mathrm{e}^{-st_{\max}/2}}{K\,\mathrm{e}^{-st_{\max}/2}+(1+s\tau)}\,.$$

Below are Bode magnitude plots for $t_{\max} = 0.1$, $\tau = 1$, and $K = \{5, 6.45, 8\}$. At $K_{\max} \approx 6.45$, the plot for $T(s)$ first hits $W^{-1}(s)$.



Below, we plot in the complex L plane, elements of the response function with the multiplicative uncertainty "circles." Shown in light gray are the frequencies $\omega = 0.02, 0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 5, 20, 30, 40, 60$. The black circle represents $\omega^* \approx 10.678$. All plots are at the critical gain $K_{\max} \approx 6.446$. The black circle intersects the point $(-1, 0)$ that marks the onset of instability.



c. If we assume that we know the delay exactly, we simply solve $L(\mathrm{i}\omega) = -1$ by separating into two equations, one for the magnitude, $|L| = 1$, the other for the phase $\tan\phi = \tan\pi = 0 = (\mathrm{Im}\, L/\mathrm{Re}\, L)$. The magnitude condition gives

$$|L| = \left| \frac{K\ \mathrm{e}^{-\mathrm{i}\omega t_{\max}/2}}{1 + \mathrm{i}\omega\tau} \right| = \frac{K}{\sqrt{1+\omega^2\tau^2}} = 1 \qquad \Longrightarrow \qquad K_{\max} = \sqrt{1 + (\omega^*)^2\tau^2}\,.$$

The condition that the phase $\phi = \pi$ implies

$$\begin{aligned} 0 &= \mathrm{Im}\, L(\mathrm{i}\omega) \propto \mathrm{Im}\,[(\cos\tfrac{1}{2}\omega t_{\max} - \mathrm{i}\sin\tfrac{1}{2}\omega t_{\max})\,(1 - \mathrm{i}\omega\tau)] \\ &= -\omega\tau\cos\tfrac{1}{2}\omega t_{\max} - \sin\tfrac{1}{2}\omega t_{\max}\,, \end{aligned}$$

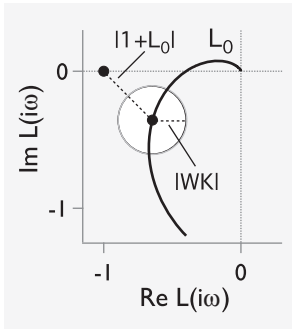
which implies

$$\tan\left(\tfrac{1}{2}\omega^* t_{\max}\right) + \omega^*\tau = 0\,.$$

Solving these two simultaneous equations for $\tau = 1$ and $t_{\max} = 0.1$ gives $\omega^* \approx 32.04$ and $K_{\max} \approx 32.06$. For $\tau = 1$ and $t_{\max} = 0.2$, they give $\omega^* \approx 16.32$ and $K_{\max} \approx 16.35$.

9.14 Robust stability for additive noise. Show that for a set of systems with additive noise limits, $\mathcal{G}(s) = G_0(s) + \Delta(s)\, W(s)$, that the condition for robust stability in a control loop with controller $K(s)$ is $\|WKS\|_\infty \leq 1$, where $S = (1 + L_0)^{-1}$ is the sensitivity function of the nominal loop dynamics, $L_0 = K(s)\, G_0(s)$.

Solution. The easiest solution is graphical and is illustrated below.



Comparing with the multiplicative case, we see that the radius of the circular uncertainty domain is now $|WK|$ rather than $|W|$. Algebraically, for all frequencies ω,

$$\begin{aligned} |1 + \mathcal{L}| &= |1 + K(s)\,[G_0(s) + \Delta(s)\, W(s)]| \\ &= |1 + L_0(s) + [\Delta(s)\, W(s)\, K(s)]| \neq 0\,. \end{aligned}$$

Then, choosing the worst possible case, with $|\Delta| = 1$ and an appropriate phase,

$$|1 + L_0(s)| - |[W(s)\, K(s)]| \geq 0\,,$$

which implies that, for all frequencies $s = \mathrm{i}\omega$,

$$\frac{|W(s)\, K(s)|}{|1 + L_0(s)|} = \left|\frac{W(s)\, K(s)}{1 + L_0(s)}\right| = |W(s)\, K(s)\, S(s)| \leq 1\,,$$

with $S = \frac{1}{1+L_0}$. This condition is equivalent to $\|WKS\|_\infty \leq 1$.

9.15 Bounding functions. The bounding functions $W_1(s)$ and $W_2(s)$ of Section 9.4.4 are typically lag and lead compensators, respectively. Find forms that give good approximations to arbitrary low- and high-frequency limits.

Solution.

For the lag compensator $W_1(s)$, we want $|W_1| \approx a \gg 1$ for $\omega \ll \omega_0$ and $|W_1| \approx b \ll 1$ for $\omega \gg \omega_0$. We also want $|W_1| \approx 1$, for $\omega = \omega_0$. The most general lag compensator is

$$W_1(s) = \frac{\alpha + s/\omega_0}{\beta + s/\omega_1}\,.$$

Imposing these limits, defining terms appropriately, and playing around leads to

$$W_1(s) = a\left(\frac{1+\frac{s}{\omega_0}b}{1+\frac{s}{\omega_0}a}\right) = W_2(s)\,.$$

We can verify that $W_1(0) = a$, $W_1(\mathrm{i}\infty) = b$, and

$$|W_1(\mathrm{i}\omega_0)| = a\left|\frac{1+ib}{1+ia}\right| = a\sqrt{\frac{1+b^2}{1+a^2}} = \sqrt{\frac{1+b^2}{1+a^{-2}}} \approx 1\,,$$

using $a \gg 1$ and $b \ll 1$.

For $W_2(s)$, we can use the same formula but with $a \ll 1$ and $b \gg 1$.

9.16 Loop-shaping criteria. Robust performance requires $\Gamma \equiv \| |W_1 S| + |W_2 T| \|_\infty < 1$. Typically, W_1 is large at low frequencies and small at high frequencies and W_2 the reverse. Here, $S = \frac{1}{1+L}$ and $T = \frac{L}{1+L}$. (Notice that $S + T = 1$.)

a. Show that $\Gamma < 1$ implies that $\mathrm{Min}(|W_1|, |W_2|) < 1$ at all frequencies.
b. Show that $|L| > |W_1|$ at low frequencies and $|L| < |W_2|^{-1}$ at high frequencies.

Solution.

a. Let us assume, at some frequency ω, that $|W_1| < |W_2|$. Then

$$|W_1| = |W_1(S+T)| \le |W_1 S| + |W_1 T| \le |W_1 S| + |W_2 T| < 1\,.$$

Alternatively, if $|W_2| < |W_1|$, we have

$$|W_2| = |W_2(S+T)| \le |W_2 S| + |W_2 T| \le |W_1 S| + |W_2 T| < 1\,.$$

Thus, $\mathrm{Min}(|W_1|, |W_2|) < 1$ at all frequencies. One or the other weight must have magnitude < 1 at each and every frequency. Notice that we prove only necessity, and the converse is not true: Setting one of the weights W_1 or W_2 to have magnitude < 1 at some frequency does not imply that $|W_1 S| + |W_2 T| < 1$.

b. The robust-performance criterion $\Gamma < 1$ is equivalent to

$$||W_1 S| + |W_2 T|| < 1\,, \forall\omega\,.$$

In the low-frequency limit ($\omega \ll 1$), we have $|W_1| \gg 1$ and $|W_2| \ll 1$. The robust performance criterion reduces to

$$|W_1 S| = \left|\frac{W_1}{1+L}\right| \approx \left|\frac{W_1}{L}\right| < 1 \quad \Longrightarrow \quad |L| > |W_1|\,.$$

In the high-frequency limit ($\omega \gg 1$), we have $|W_2| \gg 1$ and $|W_1| \ll 1$. The robust performance criterion reduces to

$$|W_2 T| = \left|\frac{W_2 L}{1+L}\right| \approx |W_2 L| < 1 \quad \Longrightarrow \quad |L| < |W_2|^{-1}\,.$$

A more-refined version of these limits imposes the looser requirements that at low frequencies, $|W_1| \gg 1$ but only that $|W_2| < 1$ (not $\ll 1$) and similarly for

high frequencies. The reasoning is that the uncertainty bounds are unlikely to be negligibly small at any frequency. The corresponding limits are

$$|L| > \frac{|W_1|}{1-|W_2|}, \quad \text{and} \quad |L| < \frac{1-|W_1|}{|W_2|},$$

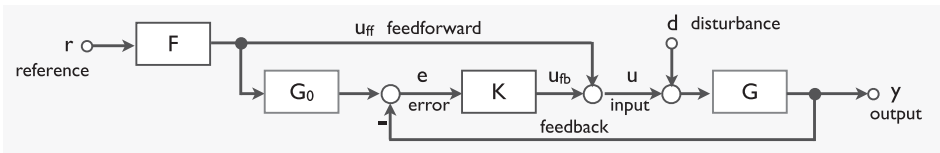
which push up further the required gain at low frequencies and push it further down at high frequencies. The added demands on the loop shape come from accommodating the "other" shaping function in each limit. See Doyle et al. (1992), Chapter 7. Remember, too, that in order to preserve stability the slope of $|L|$ on the Bode plot must be shallower than -2 near the crossover frequency where $|L| = 1$

9.17 Feedforward with model uncertainty (Devasia, 2002). What happens when an actual transfer function deviates from its model $G_0(s)$? Let $G(s) = G_0(s) + \Delta G(s)$.

a. Assuming an invertible model and a feedforward block $F = G_0^{-1}$, show that the tracking error $e(s) = r(s) - y(s)$ to a command signal $r(s)$ is $e(s) = -\frac{\Delta G/G_0}{1+KG}\, r(s)$.
b. Argue that feedforward helps only for frequencies where $\Delta G/G_0 < 1$.
c. Let $G_0 = \frac{1}{1+s}$ with uncertainty $\Delta G = \varepsilon$, a constant. What does such an uncertainty represent physically? Design a feedforward filter $F(s) = G_0^{-1}(s)\, G_{\text{lp}}(s)$, where $G_{\text{lp}}(s)$ is a low-pass cutoff whose frequency respects the criterion derived in (b).

Solution.

For convenience, the Figure 3.5 block diagram is reproduced here:



From the block diagram,

$$\begin{aligned} y &= G(u+d) & u_{\text{ff}} &= Fr \\ u &= u_{\text{ff}} + u_{\text{fb}} & u_{\text{fb}} &= Ke \\ &= Fr + K(FG_0 r - y) & e &= FG_0 r - y\,. \end{aligned}$$

Then

$$\begin{aligned} y &= G[Fr + K(FG_0 r - y) + d] \\ &= FGr + FGKG_0 r - KGy + Gd \\ &= FG\left(\frac{1+KG_0}{1+KG}\right) r + \left(\frac{G}{1+KG}\right) d\,. \end{aligned}$$

a. For command tracking errors $e_{\text{command}} = r - y$, we can set $d = 0$ and find

$$e_{\text{command}} = \left(\frac{(1 - FG) + KG(1 - FG_0)}{1 + KG} \right) r$$
$$= -\left(\frac{\Delta G \, G_0^{-1}}{1 + KG} \right) r ,$$

where the second line uses $F = G_0^{-1}$ and $G = G_0 + \Delta G$.

b. A pure feedback solution would lead to tracking error

$$e_{\text{command}} = \left(\frac{1}{1 + KG} \right) r ,$$

The ratio of the performance with and without feedforward is then, simply, $-\Delta G / G_0$. At frequencies where the magnitude of this quantity is less than one, feedforward is advantageous. When the criterion is violated, it is not.

c. Adding a constant uncertainty is a way to model high-frequency modeling errors. The idea is that they persist at all frequencies and dominate at high frequencies, where the model has small magnitude response. The criterion from (b) is

$$|(1 + s)\varepsilon| < 1 .$$

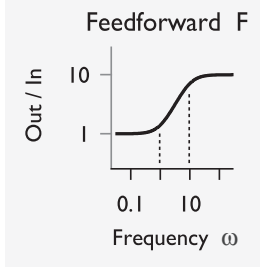
For small ε, this implies that feedforward to be limited to a bandwidth

$$\omega < \omega_{\text{ff}} = \varepsilon^{-1} .$$

A simple feedforward filter that inverts the system to as high a frequency as is reasonable uses $G_{\text{lp}}(s) = \frac{1}{1 + s/\omega_{\text{ff}}}$, which implies

$$F(s) = \frac{1 + s}{1 + s/\omega_{\text{ff}}} = \frac{1 + s}{1 + \varepsilon s} ,$$

which is a lead compensator that is active up to the frequency ω_{ff}, whose magnitude response is illustrated below for $\varepsilon = 0.1$. The general lesson is that the architecture of Figure 3.5 is useful and allows us to respect the limitations of model uncertainty at (typically) high frequencies in a simple way.



9.18 Input shaping by minimax. Go through Example 9.5.

a. Argue that symmetry dictates that the command response function is symmetric under time reversal, implying that J_2 reduces to $J_2 = |1 - 2A_0 + 2A_0 \cos \omega t_1|$.

b. Conclude that the minimax solution has $A_0 = [2(1 + (\cos \frac{\pi}{2}\varepsilon)^2)]^{-1}$ and $t_1 = \pi$.
c. Generate the minimax plot of Example 9.5.

Solution.

a. Because the equations of motion are invariant under time reversal, an optimal step going forward in time must be optimal for the time-reversed dynamics. Thus, for the optimal $u(t)$, it is possible to define an origin $t = 0$ such that the response function of the input-shaping filter obeys $F(t) = F(-t)$. Here, it is easier not to work with this origin explicitly, but for

$$u(t) = A_0\theta(t) + A_1\theta(t - t_1) + A_2\theta(t - t_2)\,, \qquad \sum_{i=0}^{n} A_i = 1\,,$$

we ask that $A_0 = A_2$. The normalization condition then implies $A_1 = 1 - 2A_0$. Similarly, the time interval between step 0 and 1 and then 1 and 2 must be equal: With $t_0 = 0$, we have $t_2 = 2t_1$. Substituting these constraints into

$$J_2 = \sqrt{(A_0 + A_1 \cos \pi\omega + A_2 \cos 2\pi\omega)^2 + (A_1 \sin \pi\omega + A_2 \sin 2\pi\omega)^2}\,.$$

then leads to

$$J_2 = |1 - 2A_0 + 2A_0 \cos \omega t_1|\,.$$

The absolute value comes after taking the square root. Since amplitudes are by definition positive, we "reflect" all negative solutions about zero.

b. The minimax solution satisfies

$$J_2(\omega = 1 - \varepsilon) = J_2(\omega = 1) = J_2(\omega = 1 + \varepsilon)\,.$$

The condition $J_2(\omega = 1 - \varepsilon) = J_2(\omega = 1 + \varepsilon)$ implies that

$$\cos[(1 - \varepsilon)t_1] = \cos[(1 + \varepsilon)t_1]\,.$$

Using the relation $\cos(x - y) = \cos x \cos y + \sin x \sin y$ then implies

$$\sin t_1 \sin \varepsilon t_1 = -\sin t_1 \sin \varepsilon t_1\,,$$

which implies $\sin t_1 = 0$, since ε is arbitrary. Thus, $t_1 = \pi$ is the shortest solution. We then need to determine the A_0 that minimizes the maximum value of

$$J_2 = |1 - 2A_0 + 2A_0 \cos \omega\pi| = \left|1 - 4A_0 \left(\sin \tfrac{\pi}{2}\omega\right)^2\right|$$

over the interval $1 - \varepsilon < \omega < 1 + \varepsilon$. The condition $J_2(\omega = 1 - \varepsilon) = J_2(\omega = 1)$ then implies

$$1 - 4A_0 = 4A_0 \left(\sin \tfrac{\pi}{2}(1 + \varepsilon)\right)^2 - 1 = 4A_0 \left(\cos \tfrac{\pi}{2}\varepsilon\right)^2 - 1\,,$$

or

$$A_0 = \frac{1}{2\left(1 + \left(\cos\frac{\pi}{2}\varepsilon\right)^2\right)}.$$

c. The plot is of

$$J_2(\omega) = |1 - 2A_0 + 2A_0 \cos\omega\pi|,$$

with the above value of A_0.

Adaptive Control 10

Problems

10.1 MRAC stability for a feedforward gain. Analyze the stability of a constant solution for the feedforward adaptive control of an underdamped oscillator presented in Eq. (10.8). That is, let $u_c = u_0 + \delta u_c(t)$, $\theta(t) = (k_m/k) + \delta\theta(t)$, and so on.

a. Define $\mu \equiv \gamma k k_m (u_0)^2$. Show that the gain perturbations $\delta\theta$ obey, to lowest order, $d_{ttt}(\delta\theta) + d_{tt}(\delta\theta) + d_t(\delta\theta) + \mu\,(\delta\theta) = -\gamma k_m u_0 (\delta y)$.

b. The Laplace transform of the $\delta\theta$ dynamics is $s^3 + s^2 + s + \mu = 0$. Deduce that stability of the MRAC system requires $0 < \mu < 1$. Hint: Look up the Routh–Hurwitz theorems, graph the roots, or just prove directly.

In the text, we use a square-wave input, not a constant. However, since the period of the square wave is long compared to the oscillation time scales, the square wave acts as a sequence of steady-state conditions, with "jump perturbations" at the start.

Solution.

a. Recalling that $\mathcal{L} = (d_{tt} + d_t + 1)$ and $\mathcal{L}\, y_m = k_m\, u_c$ and $\mathcal{L}\, y = k\, u = k\theta u_c$, we write

$$\begin{aligned}
\mathcal{L}\dot{\theta} &= -\gamma\, \mathcal{L}[y_m(y - y_m)] \\
&= -\gamma\, [(\mathcal{L} y_m) y + y_m(\mathcal{L} y) - 2 y_m(\mathcal{L} y_m)] \\
&= -\gamma\, [(k_m u_c) y + y_m(k\theta u_c) - 2 y_m(k_m u_c)] \\
&= -\gamma\, [(k_m u_c) y + y_m[k(k_m/k) + \delta\theta] u_c - 2 y_m(k_m u_c)] \\
&= -\gamma u_c\, [k_m(y - y_m) + k y_m(\delta\theta)]\ .
\end{aligned}$$

We note that $\mathcal{L}\dot{\theta} = \mathcal{L}(\delta\dot{\theta})$ and that, to lowest order, $y_m(\delta\theta) = k_m u_c(\delta\theta)$. Thus,

$$\mathcal{L}\delta\dot{\theta} + \mu\, \delta\theta = -\gamma u_c k_m \delta y(t)\,,$$

with $\mu = \gamma k k_m (u_0)^2$. Here, we use $[\text{square}^2(t)] = 1$, implying $u_c(t)^2 = u_0^2$.

b. Taking the Laplace transform of the above relation shows that stability is governed by the roots in the complex plane of $s^3 + s^2 + s + \mu = 0$, the characteristic equation. There are several ways to see that $0 < \mu < 1$ implies that all roots are in the LHS of the s-plane:

- *Routh-Hurwitz criterion.* This is a straightforward application of a rather complicated algebra algorithm for counting the numbers of roots in the LHS and RHS of the complex plane (and on the imaginary axis). The general algorithm is complicated (Dutton et al., 1997), but for a cubic polynomial, it is simple. If

$$\sigma(s) = s^3 + a_1 s^2 + a_2 s + a_3 \,,$$

then the roots of $\sigma(s) = 0$ are in the left-hand side of the complex plane if

$$a_i > 0 \quad \text{and} \quad a_1 a_2 > a_3 \,.$$

Applying the two criteria to the present case implies stability for $0 < \mu < 1$.
- *Graphical approach.* Plot the roots in the complex plane as a function of μ, and verify directly the claim.
- *Direct proof.* Even without recourse to general algebra theorems, it is possible to argue that $0 < \mu < 1$ is necessary. Because the stability is governed by a cubic equation with real coefficients, there must either be three real roots or one real and one complex-conjugate pair.
For three real roots,

$$\begin{aligned}\sigma(s) =& (s + \lambda_1)(s + \lambda_2)(s + \lambda_3) \\ &= s^3 + (\lambda_1 + \lambda_2 + \lambda_3)s^2 + (\lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_1)s + \lambda_1\lambda_2\lambda_3 \,,\end{aligned}$$

stability implies negative roots ($s = -\lambda_i < 0$), which implies $\lambda_i > 0$ and thus $a_i > 0$. In particular, $a_3 = \mu = \lambda_1\lambda_2\lambda_3$ needs $\mu > 0$. Then, given that $\lambda_1 + \lambda_2 + \lambda_3 = 1$ and $\lambda_i > 0$, we conclude that $0 < \lambda_i < 1$ and thus that the product $\lambda_1\lambda_2\lambda_3 = \mu < 1$. The two conditions together give $0 < \mu < 1$.

Assuming one real and one complex-conjugate pair, we consider the limiting cases. Either the real root = 0 or the complex-conjugate pair is $\pm \mathrm{i}\omega$. In the former case, we have

$$s(s + \lambda + \mathrm{i}\omega)(s + \lambda - \mathrm{i}\omega) = s[(s + \lambda)^2 + \omega^2] \,.$$

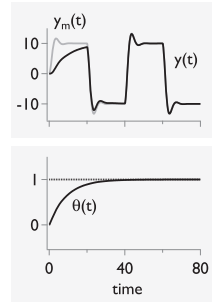
Equating this to $s^3 + s^2 + s + \mu$, we see that $\mu = 0$, $\lambda = \frac{1}{2}$, and $\omega = \frac{\sqrt{3}}{2}$, which is a possible solution. In the other case $\lambda = 0$ and $\mu = 1$, and it is easy to see $\omega = 1$ and the other root is at -1.

These two cases define the limiting values of μ. Since there are no bifurcations—these would imply three real roots, with two degenerate and cannot occur given that the coefficients of s^2 and s are unity—the intermediate cases connect them by continuity.

10.2 Normalized MRAC. Normalizing the model reference adaptive control algorithm can stabilize it for all operating amplitudes.

a. Simulate the MRAC system of Figure 10.2. Show stability for $\gamma = 0.1$ and $u_0 = 1$.

b. Confirm numerically that $u_0 > \sqrt{10}$ leads to instability for $\gamma = 0.1$.
c. The normalized MRAC algorithm replaces $e'(\theta) \to e'(\theta)/(\alpha + |e'(\theta)|^2)$, where α is a small constant. Find the normalized equation for $\dot{\theta}$ analogous to Eq. (10.5).
d. Explain how the algorithm works, using the result in Problem 10.1.
e. Reproduce the plot at right for $\gamma = 0.1$, $k = k_m = a_m = 1$, $a = 2$, $\alpha = 0.001$, and $u_0 = 10$. Without normalization, the response would be unstable.



Solution.

a. See Figure 10.2.
b. Confirm with your code. Note that the $\sqrt{10}$ comes from the linear stability analysis in Problem 10.1, where we show that stability is governed by $\mu \equiv \gamma k k_m (u_0)^2$. Here, $k = k_m = 1$, so that the stability parameter μ is

$$\mu = \gamma (u_0)^2 ,$$

so that $\gamma = 0.1$ implies $(u_0)_{max} = \sqrt{10}$ to have $\mu < 1$.
c. We have

$$e'(\theta) = y_m \qquad \Longrightarrow \qquad \dot{\theta} = -\frac{\gamma e y_m}{\alpha + y_m^2} .$$

Since $y_m = k_m u_c$, we can also write

$$\mu = \frac{\gamma (k/k_m) u_0^2}{\alpha + u_0^2} ,$$

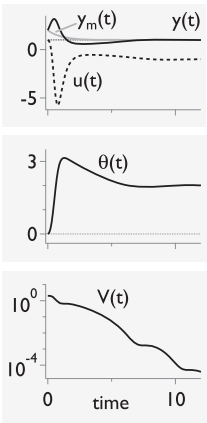
where we redefine α slightly.
d. To simplify, let $k = k_m = 1$, as in the numerical examples. Then

$$\mu = \frac{\gamma u_0^2}{\alpha + u_0^2} ,$$

which shows that $\mu \to \gamma$ for large u_0. Thus, the normalization implies that stability is independent of u_0 for $u_0^2 \gg \alpha$. The constant α is needed to prevent the $\dot{\theta}$ equation from blowing up when $y_m = 0$.
e. Reproduce the plots in the book.

10.3 MRAC to stabilize a first-order system. Consider the system $\dot{y} = ay + u$, with $a > 0$ unknown. Let the desired stable dynamics be $\dot{y}_m = a_m y_m + u_c$, with $a_m < 0$ and $u_c(t)$ an arbitrary input function. Define the control $u = u_c - \theta y$.

a. Show that the error $e = y - y_m$ obeys $\dot{e} = a_m e - (\theta - \theta^*) y$, with $\theta^* \equiv a - a_m$.
b. Show that $V(t) = \frac{1}{2}[e^2 + \frac{1}{\gamma}(\theta - \theta^*)^2]$ is a Lyapunov function if $\dot{\theta} = \gamma y e$.
c. For $u_c(t) = u_0 \neq 0$, show that the stationary solution $\theta(t) = \theta^*$. Why is this solution not valid for $u_0 = 0$?



d. Simulate the adaptive dynamics and show that the steady-state solution of Part (b) is reached in the long-time limit. The solution for parameters $\gamma = a = 1$ and $a_m = -1$, initial conditions $y(0) = 2$, $y_m(0) = \theta(0) = 0$, and input $u_c(t) = 1$ should resemble the plots at left. Note that $\theta^* = 0$. What happens as $u_0 \to 0$?

e. Solve analytically for $u_0 = 0$ for $u_c = y_m(0) = 0$ and $\theta(0) = -a$. Then show that $y_m(t) = 0$ and $\theta(t) = \sqrt{\gamma}\, y_0 \tanh\left(\sqrt{\gamma}\, y_0 t\right) - a$ and $y(t) = \sqrt{\gamma}\, y_0\, \mathrm{sech}(\sqrt{\gamma}\, y_0 t)$. Notice that $\theta(\infty) = \sqrt{\gamma}\, y_0 - a$, but $y(\infty) = y_m(\infty) = 0$: an input signal $u(t)$ that vanishes as $t \to \infty$ will not force $\theta \to \theta^*$, even though $y \to y_m$.

Solution.

a. The error dynamics $e = y - y_m$ obey

$$\begin{aligned}\dot{e} &= ay + \cancel{u_c} - \theta y - a_m y_m - \cancel{u_c} \\ &= (a - \theta)y - a_m y_m \\ &= (\underbrace{a - \theta^*}_{a_m} - \theta + \theta^*)y - a_m y_m \\ &= a_m e - (\theta - \theta^*)y\,.\end{aligned}$$

b. The candidate Lyapunov function $V(t) = \frac{1}{2}[e^2 + \frac{1}{\gamma}(\theta - \theta^*)^2]$ is clearly ≥ 0 and $= 0$ when the error vanishes ($e = 0$) and the parameters have converged to the correct value, $\theta = \theta^*$. Then, we just need to show $\dot{V} \leq 0$.

$$\begin{aligned}\dot{V} &= e\,\dot{e} + \frac{1}{\gamma}(\theta - \theta^*)\dot{\theta} \\ &= e\,[a_m e - (\theta - \theta^*)y] + \frac{1}{\gamma}(\theta - \theta^*)\dot{\theta} \\ &= a_m\, e^2 + \frac{1}{\gamma}(\theta - \theta^*)(\dot{\theta} - \gamma y e) \\ &= a_m\, e^2 \qquad\qquad \text{if } \dot{\theta} = \gamma y e \\ &\leq 0\,,\end{aligned}$$

since $a_m < 0$. Thus, choosing $\dot{\theta} = \gamma y e$ ensures that $V(t)$ is a Lyapunov function.

c. For $u_c = u_0 \neq 0$, the equations are

$$\dot{y} = ay + u_0 - \theta y\,, \qquad \dot{y}_m = a_m y_m + u_0\,, \qquad \dot{\theta} = \gamma y(y - y_m)\,.$$

The stationary solution is obtained by setting the time derivatives = 0. Then

$$y = y_m = -\frac{u_0}{a_m}\,, \qquad \text{or} \qquad e = 0\,.$$

Substituting these into the y equation gives

$$\theta = a + \frac{u_0}{y} = a - a_m = \theta^*\,,$$

meaning that the steady-state solution is consistent with the correct parameter values. The simulations in Part (d) will show that the solution actually reaches this steady-state solution. When $u_0 = 0$, we are dividing 0/0 and must be careful. Indeed, the analysis in Part (e) confirms that θ goes to a different value when $u_0 = 0$.

d. The equations to be solved are

$$\dot{y} = y + 1 - \theta y\,, \qquad \dot{y}_\mathrm{m} = -y_\mathrm{m} + 1\,, \qquad \dot{\theta} = y(y - y_\mathrm{m})\,,$$

with initial conditions $y_\mathrm{m}(0) = \theta(0) = 0$ and $y(0) = 2$. You can simplify the numerics further by using the analytical solution $y_\mathrm{m}(t) = 1 - \mathrm{e}^{-t}$.

It is interesting to examine the solution for $u_0 \to 0$, to confirm numerically the conclusions of Parts (c) and (e) that when $u_0 \neq 0$, $\theta \to \theta^*$, but when $u_0 = 0$, $\theta \to -a + \sqrt{\gamma}\, y_0$. As $u_0 \to 0$, θ takes a diverging time to reach θ^*.

e. When $u_\mathrm{c}(t) = u_0 = 0$, $\dot{y}_\mathrm{m} = -a_\mathrm{m} y_\mathrm{m} + 0$, with $y_\mathrm{m}(0) = 0$, which implies $y_\mathrm{m}(t) = 0$, too. The equations then simplify to

$$\begin{aligned} \dot{y} &= ay - \theta y = -(\theta - a)y \equiv -\theta_1 y \qquad \text{with } \theta_1(t) = \theta(t) - a \\ \dot{\theta} &= \dot{\theta}_1 = \gamma y e = \gamma y^2\,. \end{aligned}$$

Differentiating the equation for θ_1 gives

$$\ddot{\theta}_1 = 2\gamma y\, \dot{y} = -2\gamma\theta_1 y^2 = -2\theta_1\dot{\theta}_1 = -\frac{\mathrm{d}}{\mathrm{d}t}\theta_1^2\,.$$

Integrating, we find

$$\begin{aligned} \dot{\theta}_1(t) + \theta_1^2(t) &= \text{const} \\ &= \dot{\theta}_1(0) + \theta_1^2(0) \\ &= \gamma y_0^2 + 0 \\ &= \gamma y_0^2\,. \end{aligned}$$

Scaling by defining $\tau = \sqrt{\gamma} y_0\, t$ and $\theta_1 = \sqrt{\gamma} y_0\, \overline{\theta}$ reduces the equation to

$$\overline{\theta}'(\tau) + \overline{\theta}^2(\tau) = 1\,, \qquad \overline{\theta}(0) = 0\,, \qquad \Longrightarrow \qquad \overline{\theta}(\tau) = \tanh\tau\,.$$

In dimensional variables, $\theta_1(t) = \sqrt{\gamma} y_0 \tanh \sqrt{\gamma} y_0 t$, or

$$\theta(t) = -a + \sqrt{\gamma} y_0 \tanh \sqrt{\gamma} y_0 t\,.$$

We then substitute $\theta_1(t)$ into $\dot{y} = -\theta_1 y$ and integrate again, to get

$$y(t) = \sqrt{\gamma}\, y_0\, \mathrm{sech}(\sqrt{\gamma}\, y_0 t)$$

As discussed, for $t \to \infty$, $y(t) \to 0$, but $\theta(t) \to -a + \sqrt{\gamma}\, y_0$, which depends on the learning rate γ, the initial state y_0, and a. For $\theta(0) \neq -a$, the limit depends on that initial condition, too. The important point is that $\theta(t) \not\to \theta^*$.

10.4 Lyapunov function for 2nd-order MRAC. In Example 10.3 the Lyapunov construction works if we observe the full state vector $\boldsymbol{e}_x$, or, equivalently, y and $\dot{y}$.

a. Find a Lyapunov function by showing that $\boldsymbol{Q} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right) \implies \boldsymbol{P} = \left(\begin{smallmatrix} \frac{3}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{smallmatrix}\right)$.
b. Show that the adaptation law for $\theta(t)$ becomes $\dot{\theta} = -\gamma(e + 2\dot{e})\, u_\mathrm{c}$.
c. Why might it be better to integrate the law for $\dot{\theta}$ in a practical implementation?
d. Verify by simulation that the resulting adaptive system is stable for step commands of any amplitude.

Solution.

a. Here are three ways to find the matrix $\boldsymbol{P}$ given a positive-definite matrix $\boldsymbol{Q}$.
   i. You can show that the Lyapunov equation is satisfied by substituting the matrices $\boldsymbol{P}$, $\boldsymbol{Q}$, and $\boldsymbol{A}$ and confirming that
   $$\boldsymbol{A}^\mathsf{T}\boldsymbol{P} - \boldsymbol{P}\boldsymbol{A} = -\boldsymbol{Q}\,.$$
   ii. Control software usually has a Lyapunov equation solver built in.
   iii. Write $\boldsymbol{P}$ as a 3-component vector and solve the resulting matrix equation.

   Once you have found $\boldsymbol{P}$, you can find the Lyapunov function (with $k = 1$):
   $$\begin{aligned} V &= \frac{1}{2}\left[\left(\boldsymbol{e}_x^\mathsf{T}\boldsymbol{P}\boldsymbol{e}_x\right) + \frac{1}{\gamma}(\theta - \theta^*)^2\right] \\ &= \frac{1}{2}\left[\begin{pmatrix} e & \dot{e} \end{pmatrix}\begin{pmatrix} \frac{3}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}\begin{pmatrix} e \\ \dot{e} \end{pmatrix} + \frac{1}{\gamma}(\theta - \theta^*)^2\right] \\ &= \frac{1}{2}\left(\frac{3}{2}e^2 + e\,\dot{e} + \dot{e}^2\right) + \frac{1}{2\gamma}(\theta - \theta^*)^2\,. \end{aligned}$$

b. From the text, the control is of the form
   $$\dot{\theta} = -\gamma\boldsymbol{B}^\mathsf{T}\boldsymbol{P}\boldsymbol{e}_x\, u_\mathrm{c}$$
   with
   $$\boldsymbol{e}_x = \begin{pmatrix} e \\ \dot{e} \end{pmatrix},$$
   with $e = y - y_\mathrm{m}$. This gives
   $$\begin{aligned} \dot{\theta} &= -\gamma\begin{pmatrix} 0 & 1 \end{pmatrix}\begin{pmatrix} \frac{3}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}\begin{pmatrix} e \\ \dot{e} \end{pmatrix} u_\mathrm{c} = -\gamma(e/2 + \dot{e})u_\mathrm{c} \\ &= -\gamma(e + 2\dot{e})u_\mathrm{c}\,, \end{aligned}$$
   where we rescale γ in the last step.

c. The problem with the adaptation law
   $$\dot{\theta} = -\gamma(e + 2\dot{e})\, u_\mathrm{c}$$
   is that we need to evaluate $\dot{e} = \dot{y} - \dot{y}_\mathrm{m}$. There is no problem calculating the latter, since $y_\mathrm{m}(t)$ is determined by $u_\mathrm{c}(t)$, which we choose. Thus, we can calculate $y_\mathrm{m}(t)$ either analytically or numerically as accurately as required. But the observation $y(t)$ is another story, as taking a numerical derivative will amplify

any measurement noise. In this case, integrating the law for $\dot{\theta}$ can improve the situation. Doing so gives

$$\theta(t) = -\gamma \left[\int_0^t \mathrm{d}t'\, e(t') + 2e(t)\, u_\mathrm{c}(t) \, - 2 \int_0^t \mathrm{d}t'\, e(t')\, \dot{u}_c(t') \right],$$

where we have integrated by parts and assumed $u_\mathrm{c}(0) = 0$ for convenience. Although this law requires the time derivative of $u_\mathrm{c}(t)$, we know u_c exactly, since we choose it. Thus, we can also compute its derivative $\dot{u}_c$ exactly.

Notice that even though our simple construction from Eq. (10.14) failed, we nonetheless have found an integral-control-like feedback law that requires only the observation $y(t)$, in addition to $u_\mathrm{c}(t)$, $\dot{u}_c(t)$, and $y_\mathrm{m}(t)$, which are all known exactly. This success hints that a more sophisticated approach can systematically lead to stable, observation-based algorithms for higher-order systems. Åström and Wittenmark (2008) and Slotine and Li (1991) show how to proceed.

d. Simulating the dynamics confirms the claim. It is interesting to see that dropping the $\dot{e}$ term in the law for $\dot{\theta}$ can lead to instability when the amplitude of $u_\mathrm{c}(t)$ is too large.

10.5 MRAC with unmodeled dynamics. Consider a Lyapunov control for a model system $y_\mathrm{m} = k_\mathrm{m} G_0(s) u_\mathrm{c}$, with $G_0(s) = \frac{1}{1+s}$ where the actual dynamics are given by $G_1(s) = \left(\frac{1}{1+s}\right)\left(\frac{1}{1+s/\alpha}\right)$, and the feedforward gain $\theta(t)$ is adjusted according to $\dot{\theta} = -\gamma\, u_\mathrm{c}\, e$, with $e = y - y_\mathrm{m}$ and command input $u_\mathrm{c}(t) = u_0 \cos \omega t$. Now analyze the stability of the θ dynamics by assuming a separation of time scales, $\gamma \ll \omega$, so that $\theta(t)$ evolves very slowly compared to the oscillation period of the forcing, $\tau = 2\pi/\omega$.

a. By averaging over the time τ, show that $\dot{\bar{\theta}} \approx -\gamma \bar{\theta} k \left(\overline{u_\mathrm{c}\, G_1\, u_\mathrm{c}}\right) + \gamma k_\mathrm{m} \left(\overline{u_\mathrm{c}\, G_0\, u_\mathrm{c}}\right)$, where the overline, $\overline{x} \equiv \frac{1}{\tau}\int_0^\tau \mathrm{d}t\, x(t)$, denotes averaging over one period.

b. Show that $\bar{\theta}(t)$ is unstable when $u_0^2\, \mathrm{Re}\, G(\mathrm{i}\omega) < 0$.

c. Show that this happens for the example here when the input frequency $\omega > \sqrt{\alpha}$.

This *method of averaging* is widely used to analyze nonlinear oscillating systems.

Solution.

The feedforward gain $\theta(t)$ obeys

$$\dot{\theta} = -\gamma\, u_\mathrm{c}\, (y - y_\mathrm{m})\,.$$

Given a sinusoidal input,

$$u_\mathrm{c} = \frac{1}{2}\left(u_0\, \mathrm{e}^{\mathrm{i}\omega t} + u_0^*\, \mathrm{e}^{-\mathrm{i}\omega t}\right).$$

a. The response $y_\mathrm{m}(t)$ is strictly sinusoidal at frequency ω, since it is a linear equation. The response $y(t)$ is not strictly at ω, since $\theta(t)$ changes in time. However, in the spirit of averaging, we take ω constant over the time scale of

averaging, and then $y(t)$ is also sinusoidal. Averaging the equation of motion for $\theta(t)$ then leads to

$$\dot{\bar{\theta}} = -\gamma\bar{\theta}k\left(\overline{u_c\, G_1\, u_c}\right) + \gamma k_m\left(\overline{u_c\, G_0\, u_c}\right),$$

where overline denotes average over τ. Also, $y_m = k_m G_0\, u_c$ and $y = kG_1\,\theta u_c$.

b. Only DC terms formed by $e^{i\omega t}\times e^{-i\omega t}$ survive the averaging. This gives, for the first term,

$$\overline{u_c\, G_1\, u_c} = \frac{1}{2}u_0^2\,\mathrm{Re}\, G_1(i\omega)\,.$$

The θ equation then becomes

$$\dot{\bar{\theta}} = -\gamma\bar{\theta}k\tfrac{1}{2}u_0^2\,\mathrm{Re}\, G_1(i\omega) + \gamma k_m\tfrac{1}{2}u_0^2\,\mathrm{Re}\, G_0(i\omega)\,,$$

The first term on the RHS leads to an instability if it is negative. Thus, the criterion for instability is

$$u_0^2\,\mathrm{Re}\, G(i\omega) < 0\,.$$

c. Since $u_0^2 > 0$, instability occurs when $\mathrm{Re}\, G(i\omega) < 0$. For

$$u_0^2\,\mathrm{Re}\, G(i\omega) < 0\,.$$

In this case, $G_1(s) = \left(\frac{1}{1+s}\right)\left(\frac{1}{1+s/\alpha}\right)$, and thus

$$\mathrm{Re}\, G_1(i\omega) \propto \mathrm{Re}\,(1-i\omega)(1-i\omega/\alpha) = (1-\omega^2/\alpha)\,,$$

which leads to the condition that stable adaptation requires that $\omega < \sqrt{\alpha}$.

10.6 Extremum-seeking control. Assuming that $J(\theta) - J_0 + \frac{1}{2}J''(\theta_0)(\theta-\theta_0)^2$ exactly, analyze Eqs. (10.18), with $k = 50$, $a = 0.2$, $\omega_0 = 2\pi\times 10$, $\omega_h = 2\pi$.

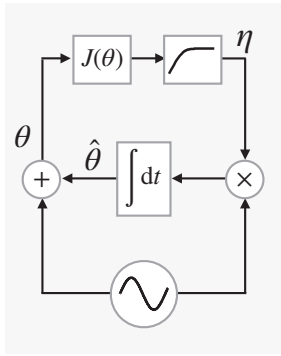
a. By averaging over the modulation period $2\pi/\omega_0$, show that $\dot{\hat{\theta}} \approx -\gamma J'(\hat{\theta})$, Hint: Expand $J[\theta(t)]$ to first order in a, filter DC terms, modulate, filter AC terms.

b. Solve Eqs. (10.18) numerically and confirm the plots of $J(t)$ and $\hat{\theta}(t)$ in Section 10.1.2. Investigate the effects of periodically modulating the coefficient J''.

Solution.

a. For convenience, the block diagram is reproduced below. Recall the equations,

$$\begin{aligned} J(\theta) &= J(\hat{\theta} + a\cos\omega_0 t) && \text{modulate}\\ \dot{\eta} &= -\omega_h\eta + \dot{J} && \text{high pass}\\ \dot{\hat{\theta}} &= -k\,(a\cos\omega_0 t)\,\eta && \text{demodulate}\,. \end{aligned}$$



We proceed through the equations. Expanding about $\hat{\theta}$ to first order in a, we have

$$J(\theta) \approx J(\hat{\theta}) + J'(\hat{\theta}) a \cos \omega_0 t \, .$$

The high-pass filter, with transfer function $\frac{s}{s+\omega_h}$ removes the (nearly) DC terms, leaving

$$\eta(t) \approx J'(\hat{\theta}) a \cos \omega_0 t \, .$$

The demodulation step is accomplished here by multiplying by $a \cos \omega_0 t$ and then integrating. (Some implementations of extremum seeking use a low-pass filter, as is common in lock-in amplifiers.) This gives,

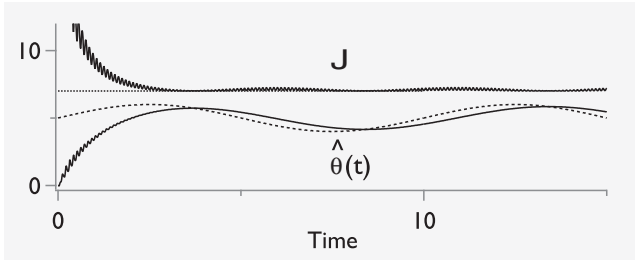
$$\begin{aligned} \dot{\hat{\theta}}(t) &= -k(a \cos \omega_0 t)\eta \\ &= -ka^2 J'(\hat{\theta}) \cos^2 \omega_0 t \\ &\approx -\tfrac{1}{2} ka^2 J'(\hat{\theta}) \\ &\equiv -\gamma J'(\hat{\theta}) \, . \end{aligned}$$

In the third line, we use $\cos^2 \omega_0 t = \frac{1}{2}(1 + \cos 2\omega_0 t) \approx \frac{1}{2}$, using the averaging. The implicit picture (supported below by numerics) is of a slow relaxation with a small rapid ripple. Mathematically, we need $\gamma \ll \omega_0$. Notice how the unknown $J''(\theta_0)$ coefficient affects the relaxation rate but not the convergence of $\hat{\theta}$ (up to the ripple contribution, which is neglected in the final equation).

The above arguments are clearly only heuristic: where we "drop" various DC and AC terms, we need to more carefully assess their size and influence. More rigorous arguments draw on singular perturbation theory, as dropping terms reduces the dimension of the state space. (Singular perturbations occur when a small term increases the order of a dynamical system, as, for example, in the WKB theory of quantum mechanics.)

b. The plot below is for a "disturbance" $\theta_0 \to \theta_0(t) = \theta_0 + a_d \cos(\omega_d t)$, with $\theta_0 = 5$, $a_d = 1$, and $\omega_d = 2\pi \times 0.1$. Notice that $\hat{\theta}(t)$ (solid line) tracks the time-varying parameter $\theta_0(t)$ (dashed line) with only a small phase lag while the cost (J) remains very close to its minimum value (7). Notice, too, that the approximation implicit in the demodulation equation is reasonable: The $\hat{\theta}$ behavior

really does consist of a slow relaxation plus small added ripple. Increasing the frequency and/or amplitude of the $\theta_0(t)$ disturbance will increase the phase lag (and increase the small cost oscillations). Increasing k makes the system track parameters better but also makes the system more sensitive to noise.



10.7 Simplest LS. In Example 10.5, we wrote down the formulas for estimating the gain of a linear relationship $y_k = \theta u_k + \xi_k$. Here, we simplify further by choosing $u_k = 1$.

a. Interpret the formulas for $\hat{\theta}$, $\hat{\xi}$, and P in terms of elementary statistics.
b. Introduce a *forgetting factor* λ, with $0 < \lambda < 1$. Assume that observations last long enough that $(1-\lambda) \gg \frac{1}{k}$. Show that $\hat{\theta}_k \approx (1-\lambda)\sum_i \lambda^{k-i} y_i$.
c. Show that the recursive version for $\hat{\theta}_k$ is the moving average given in Eq. (10.39).

Solution.

a. The least-squares formulas are

$$\hat{\theta} = \frac{\sum_{k=1}^N y_k u_k}{\sum_{k=1}^N u_k^2} = \frac{\sum_{k=1}^N y_k (1)}{\sum_{k=1}^N (1)^2} = \frac{\sum y_k}{N} \equiv \overline{y}\,,$$

where $\overline{y}$ is the arithmetic average of the N terms y_k. Then

$$(N-1)\hat{\xi}^2 = \sum y_k^2 - \hat{\theta}\sum u_k y_k = \sum y_k^2 - \hat{\theta}\sum (1) y_k = \sum y_k^2 - N\overline{y}^2 = \sum (y_k - \overline{y})^2\,.$$

In other words, $\hat{\xi}^2 = \frac{1}{N-1}\sum (y_k - \overline{y})^2$, which is just the unbiased estimate of the sample variance. Finally, $P = \frac{\hat{\xi}^2}{N}$ is the variance of the mean.

b. We minimize the weighted least-squares sum

$$\chi^2 = \sum_{i=1}^k \lambda^{k-i}(y_i - \theta_k)^2\,.$$

Taking $\partial_{\theta_k}\chi^2 = 0$ gives

$$\sum_{i=1}^k \lambda^{k-i}(y_i - \theta_k)(-2) = 0\,,$$

which implies

$$\hat{\theta}_k = \frac{\sum \lambda^{k-i} y_i}{\sum \lambda^{k-i}} = (1-\lambda)\sum_{i=1}^k \lambda^{k-i} y_i\,.$$

In the last step, we assume that k is large enough that we can approximate the sum by extending to ∞.

c. The recursive version is then

$$\begin{aligned}
\hat{\theta}_k &= (1-\lambda)\sum_{i=1}^{k}\lambda^{k-i}y_i \\
&= (1-\lambda)\left[\sum_{i=1}^{k-1}\lambda^{k-i}y_i + y_k\right] \\
&= (1-\lambda)\left[\lambda\sum_{i=1}^{k-1}\lambda^{k-i-1}y_i + y_k\right] \\
&\approx \lambda(1-\lambda)\sum_{i=0}^{k-1}\lambda^{k-i-1}y_i + (1-\lambda)\,y_k \\
&= \lambda\,\hat{\theta}_{k-1} + (1-\lambda)\,y_k\,.
\end{aligned}$$

10.8 Recursive estimation of the noise strength. Derive Eq. (10.27). Reformulate the recursion relation as $\hat{\xi}_0^2 = 0$ and $\hat{\xi}_k^2 = \hat{\xi}_{k-1}^2 + \varepsilon_k^2$ for $1 \le k \le N_p + 1$. For $k \ge N_p + 2$, we have $\hat{\xi}_k^2 = (\frac{k-N_p-1}{k-N_p})\hat{\xi}_{k-1}^2 + (\frac{1}{k-N_p})\varepsilon_k^2$. Hint: write the first few cases, for $N_p = 1$.

Solution.

Let's first write out the first few terms, assuming $N_p = 1$:

$$\begin{aligned}
\hat{\xi}_2^2 &= \varepsilon_1^2 + \varepsilon_2^2 \\
\hat{\xi}_3^2 &= \frac{3-1-1}{3-1}\hat{\xi}_2^2 + \frac{1}{3-1}\varepsilon_3^2 = \frac{1}{2}\left(\varepsilon_1^2 + \varepsilon_2^2\right) + \frac{1}{2}\varepsilon_3^2 = \frac{1}{2}\left(\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2\right) \\
\hat{\xi}_4^2 &= \frac{4-1-1}{4-1}\hat{\xi}_3^2 + \frac{1}{4-1}\varepsilon_4^2 = \frac{2}{3}\cdot\frac{1}{2}\left(\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2\right) + \frac{1}{3}\varepsilon_4^2 = \frac{1}{3}\left(\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + \varepsilon_4^2\right)
\end{aligned}$$

Now that we see the pattern, the general case is

$$\begin{aligned}
\hat{\xi}_k^2 &= \left(\frac{k-N_p-1}{k-N_p}\right)\hat{\xi}_{k-1}^2 + \left(\frac{1}{k-N_p}\right)\varepsilon_k^2 \\
&= \left(\frac{\cancel{k-N_p-1}}{k-N_p}\right)\left[\left(\frac{k-N_p-2}{\cancel{k-N_p-1}}\right)\hat{\xi}_{k-2}^2 + \left(\frac{1}{\cancel{k-N_p-1}}\right)\varepsilon_{k-1}^2\right] + \left(\frac{1}{k-N_p}\right)\varepsilon_k^2 \\
&= \left(\frac{k-N_p-2}{k-N_p}\right)\hat{\xi}_{k-2}^2 + \left(\frac{1}{k-N_p}\right)\left(\varepsilon_k^2 + \varepsilon_{k-1}^2\right) \\
&\;\;\vdots \\
&= \left(\frac{k-N_p-(k-N_p-1)}{k-N_p}\right)\hat{\xi}_{k-(k-N_p-1)}^2 + \left(\frac{1}{k-N_p}\right)\left(\varepsilon_k^2 + \varepsilon_{k-1}^2 + \cdots + \varepsilon_{k-(k-N_p-2)}^2\right) \\
&= \left(\frac{1}{k-N_p}\right)\left(\hat{\xi}_{N_p+1}^2 + \varepsilon_k^2 + \varepsilon_{k-1}^2 + \cdots + \varepsilon_{N_p+2}^2\right) \\
&= \left(\frac{1}{k-N_p}\right)\sum_{i=1}^{k}\varepsilon_i^2\,.
\end{aligned}$$

Notice that in the last step we have set

$$\hat{\xi}^2_{N_p+1} = \sum_{i=1}^{N_p+1} \varepsilon_i^2 ,$$

which is the initialization condition for the recurrence relation.

10.9 LS vs. RLS. The goal is to compare the ordinary least squares (LS) equations with their recursive (RLS) counterparts for $y_k = \theta u_k + \xi_k$. Cf. Examples 10.5 and 10.6. To simplify the analysis, choose a step-function input, $u_k = 1$, and let $\xi^2 = 1$.

a. Show that the LS solution for k steps is $\hat{\theta}_k = \frac{1}{k}\sum y_k$, with variance $P_k = 1/k$.
b. Show that the RLS solution is the same, if you correctly choose P_1 and $\hat{\theta}_1$.
c. How does altering the step-input amplitude to $u_k = u$ affect the convergence?
d. How does choosing a pulse, $u_k = u$ for $k \leq K$ and 0 otherwise, affect convergence?
e. (i) Write a simulation to illustrate that LS = RLS only if the RLS initial conditions are chosen correctly. (ii) Compare the convergence for step amplitudes $u = 1$ and $u = 10$. (iii) Show that $\hat{\theta}$ for a finite pulse gets "stuck" when the pulse ends.

Solution.

a. The LS solution is $\hat{\theta}_k = \frac{\sum y_k u_k}{\sum u_k^2} = \frac{\sum y_k}{k}$. This is the arithmetic average of $\{y_1, y_2, \ldots, y_k\}$. Similarly, the variance is $P_k = \frac{1}{\sum u_k^2} = \frac{1}{k}$.
b. The RLS solution for P is

$$P_{k+1} = \frac{P_k}{1 + u_{k+1}^2 P_k} = \frac{P_k}{1 + P_k} ,$$

Iterating this recurrence relation gives

$$P_{k+1} = \frac{P_k}{1+P_k} = \frac{\frac{P_{k-1}}{1+P_{k-1}}}{1+\frac{P_{k-1}}{1+P_{k-1}}} = \frac{P_{k-1}}{1+2P_{k-1}} = \cdots = \frac{P_1}{1+kP_1} .$$

Thus, $P_k = 1/k$ for all k if and only if we choose $P_1 = 1$ (or ξ^2 if not scaled). On the other hand, $P_k \to 1/k$ asymptotically for $k \to \infty$, for all P_1.
From P_k, we can calculate L_{k+1} as

$$L_{k+1} = \frac{P_k u_{k+1}}{1 + u_{k+1}^2 P_k} = \frac{P_k}{1+P_k} = \frac{P_1}{1+kP_1} .$$

Again, if $P_1 = 1$, we have $L_k = 1/k$. For the parameter estimate itself, we have

$$\begin{aligned} \hat{\theta}_{k+1} &= \hat{\theta}_k + L_{k+1}(y_{k+1} - \hat{\theta}_k) \\ &= (1 - L_{k+1})\,\hat{\theta}_k + L_{k+1}\, y_{k+1} \\ &= \left(1 - \frac{P_1}{1+kP_1}\right)\hat{\theta}_k + \frac{P_1}{1+kP_1}\, y_{k+1} . \end{aligned}$$

If we choose $P_1 = 1$, the expression simplifies to

$$\hat{\theta}_{k+1} = \left(\frac{k}{k+1}\right)\hat{\theta}_k + \frac{1}{1+k}\, y_{k+1} \ .$$

With this choice of P_1, the estimate $\hat{\theta}_{k+1}$ is given by an exponential moving average of y_k, which is just the result of LS. The corollary, of course, is that if we make the wrong choice for P_1 or $\hat{\theta}_1$, the two estimates will not agree. However, the recurrence relations will converge to the same asymptotic solution as the batch algorithm, for $k \to \infty$. Better choices of initial conditions converge faster.

c. Redoing the calculations with $u_k = u$, we find

$$\hat{\theta}_k = \frac{\sum y_k u_k}{\sum u_k^2} = \theta + \frac{\sum \xi_k u_k}{\sum u_k^2} \to \theta + \frac{\sum \xi_k}{uk} \,,$$

where we substitute for y_k in terms of the (unobservable) noise ξ_k. We see that the estimate is unbiased and that the correction term converges more quickly, by a factor u. We can reach a similar conclusion by examining the variance:

$$P_k = \frac{1}{\sum u_k^2} = \frac{1}{ku} \,.$$

Thus, $\hat{\theta}$ will converge faster if the input signal is larger. The relevant ratio is $\langle u_k \rangle / \xi$ in dimensional units.

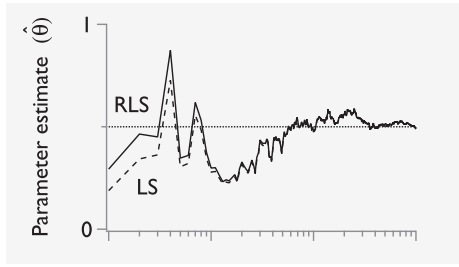
d. If the input is a pulse of amplitude u that lasts until $k = K$, the estimate is "stuck" at its $k = K$ value and does not change thereafter:

$$\hat{\theta}_k = \begin{cases} \frac{\sum_{i=1}^{k} y_i}{ku} & k \le K \\ \frac{\sum_{i=1}^{K} y_i}{Ku} & k > K \end{cases} \,, \qquad P_k = \begin{cases} \frac{1}{ku} & k \le K \\ \frac{1}{Ku} & k > K \end{cases} \,.$$

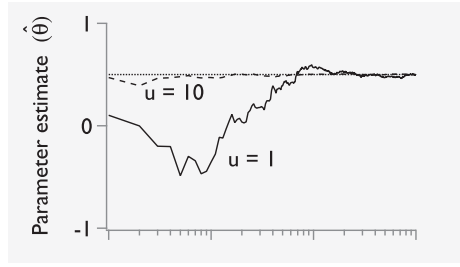
Intuitively, with no new information coming in, the estimate remains what it was at the last time step where new information was available. This is perhaps the simplest example illustrating the need for *persistent excitation* in the input signal u_k. From the formula for P_k, we see that for $P_k \to 0$, we need $\sum u_k^2 \to \infty$. Since u_k is bounded (it's the input), it must not go too quickly to zero as $k \to \infty$.

e. The graphs below show simulations for $\theta = 0.5$. The top, for different initial conditions, shows the LS curve (dashed) and RLS curve (solid). The RLS curve uses the "wrong" initial conditions.

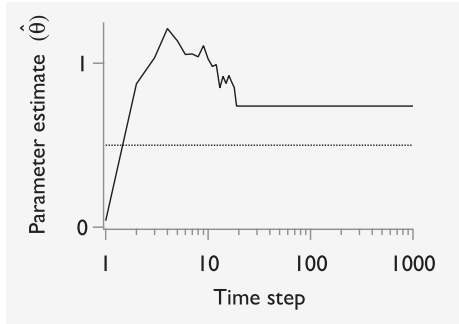
Different initial conditions



Different step amplitudes



Finite pulse length



10.10 Continuous-time recursive-least-squares (RLS). The continuous-time algorithm is simpler than the discrete-time case and illuminates its structure. For scalar output $y(t)$, the model is $y(t) = \boldsymbol{\varphi}^{\mathsf{T}}(t)\,\boldsymbol{\theta} + \xi_c(t)$, with Gaussian noise $\langle \xi_c(t)\,\xi_c(t')\rangle = \xi_c^2\delta(t-t')$. If we scale $y \to y/\xi_c$ and $\boldsymbol{\varphi} \to \boldsymbol{\varphi}/\xi_c$, the loss function $\chi^2(\boldsymbol{\theta}) = \int_0^t \mathrm{d}t'\,[y(t') - \boldsymbol{\varphi}^{\mathsf{T}}(t')\,\boldsymbol{\theta}]^2$.

a. Show that $\hat{\boldsymbol{\theta}}(t) = \boldsymbol{P}(t)\int_0^t \mathrm{d}t'\,\boldsymbol{\varphi}(t')\,y(t')$, with $\boldsymbol{P}^{-1}(t) \equiv \int_0^t \mathrm{d}t'\,\boldsymbol{\varphi}(t')\,\boldsymbol{\varphi}^{\mathsf{T}}(t')$ minimizes χ^2. Note that $\boldsymbol{P}(t)$ is the covariance matrix for the estimate $\hat{\boldsymbol{\theta}}(t)$ (see Problem A.8.2).

b. Differentiating the equations for $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{P}^{-1}$, derive an equivalent recursive algorithm. Hint: $\mathrm{d}_t\boldsymbol{I} = \mathrm{d}_t(\boldsymbol{P}\,\boldsymbol{P}^{-1}) = \boldsymbol{0}$. Show that $\mathrm{d}_t\hat{\boldsymbol{\theta}} = \boldsymbol{P}\,\boldsymbol{\varphi}\,\varepsilon$ and $\mathrm{d}_t\boldsymbol{P} = -\boldsymbol{P}\,\boldsymbol{\varphi}\,\boldsymbol{\varphi}^{\mathsf{T}}\,\boldsymbol{P}$, with $\varepsilon \equiv y - \boldsymbol{\varphi}^{\mathsf{T}}\hat{\boldsymbol{\theta}}$. The parameter estimate $\hat{\boldsymbol{\theta}}$ changes because of non-zero innovations ε, the difference between the prediction $\boldsymbol{\varphi}^{\mathsf{T}}\hat{\boldsymbol{\theta}}$ and the observation y.

c. Include a forgetting factor λ' by defining $\chi^2(\boldsymbol{\theta}) = \int_0^t \mathrm{d}t'\,\mathrm{e}^{-\lambda'(t-t')}\left[y(t') - \boldsymbol{\varphi}^{\mathsf{T}}(t')\,\boldsymbol{\theta}\right]^2$. Show that the only change to the RLS equations is to take $\dot{\boldsymbol{P}} = \lambda'\boldsymbol{P} - \boldsymbol{P}\,\boldsymbol{\varphi}\,\boldsymbol{\varphi}^{\mathsf{T}}\,\boldsymbol{P}$.

d. Derive the discrete RLS equations, Eq. (10.26). Hint: Apply the *Sherman–Morrison* matrix-inversion formula, Eq. (A.15), to $d_t\boldsymbol{P}^{-1}$.

Solution.

a. Differentiating with respect to the vector $\boldsymbol{\theta}$ gives

$$\frac{\partial \chi^2}{\partial \boldsymbol{\theta}} = \int_0^t \mathrm{d}t' \left[y(t') - \boldsymbol{\varphi}^{\mathsf{T}}(t')\,\hat{\boldsymbol{\theta}} \right] \boldsymbol{\varphi}^{\mathsf{T}}(t') = \mathbf{0}^{\mathsf{T}} .$$

We take the transpose and recognize that the estimate $\hat{\boldsymbol{\theta}}$ is really a function of time because it will change as the length of data collected over the interval $(0, t)$ increases. Then

$$\left[\int_0^t \mathrm{d}t'\, \boldsymbol{\varphi}(t')\,\boldsymbol{\varphi}^{\mathsf{T}}(t') \right] \hat{\boldsymbol{\theta}}(t) \equiv \boldsymbol{P}^{-1}(t)\,\hat{\boldsymbol{\theta}}(t) = \int_0^t \mathrm{d}t'\, \boldsymbol{\varphi}(t')\, y(t') ,$$

which, on multiplying by $\boldsymbol{P}(t)$, gives the result for $\hat{\boldsymbol{\theta}}$.

b. We first establish the identity for $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{P}^{-1}$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{I} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\boldsymbol{P}\,\boldsymbol{P}^{-1}\right) = \left(\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{P}\right)\boldsymbol{P}^{-1} + \boldsymbol{P}\left(\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{P}^{-1}\right) = \mathbf{0} ,$$

so that

$$\dot{\boldsymbol{P}} = -\boldsymbol{P}\left(\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{P}^{-1}\right)\boldsymbol{P} .$$

In our problem, $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{P}^{-1} = \boldsymbol{\varphi}\,\boldsymbol{\varphi}^{\mathsf{T}}$, so that

$$\dot{\boldsymbol{P}} = -\boldsymbol{P}\boldsymbol{\varphi}\,\boldsymbol{\varphi}^{\mathsf{T}}\boldsymbol{P} .$$

Then, differentiating the equation for $\hat{\boldsymbol{\theta}}$ gives

$$\begin{aligned}
\frac{\mathrm{d}\hat{\boldsymbol{\theta}}}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}\left[\boldsymbol{P} \int_0^t \mathrm{d}t'\, \boldsymbol{\varphi}(t')\, y(t') \right] \\
&= \dot{\boldsymbol{P}} \int_0^t \mathrm{d}t'\, \boldsymbol{\varphi}(t')\, y(t') + \boldsymbol{P}\,\boldsymbol{\varphi}(t)\, y(t) \\
&= -\boldsymbol{P}\boldsymbol{\varphi}\,\boldsymbol{\varphi}^{\mathsf{T}}\boldsymbol{P} \int_0^t \mathrm{d}t'\, \boldsymbol{\varphi}(t')\, y(t') + \boldsymbol{P}\,\boldsymbol{\varphi}(t)\, y(t) \\
&= -\boldsymbol{P}\boldsymbol{\varphi}\,\boldsymbol{\varphi}^{\mathsf{T}}\hat{\boldsymbol{\theta}} + \boldsymbol{P}\,\boldsymbol{\varphi}(t)\, y(t) \\
&= \boldsymbol{P}\boldsymbol{\varphi}\left(y - \boldsymbol{\varphi}^{\mathsf{T}}\hat{\boldsymbol{\theta}} \right) \\
&= \boldsymbol{P}\boldsymbol{\varphi}\,\varepsilon , \qquad \varepsilon \equiv y - \boldsymbol{\varphi}^{\mathsf{T}}\hat{\boldsymbol{\theta}} .
\end{aligned}$$

c. We quickly repeat Parts (a) and (b), modified to include the forgetting factor $\mathrm{e}^{-\lambda' t}$. Differentiating with respect to the vector $\boldsymbol{\theta}$ gives

$$\frac{\partial \chi^2}{\partial \boldsymbol{\theta}} = \int_0^t \mathrm{d}t'\, \mathrm{e}^{-\lambda'(t-t')} \left[y(t') - \boldsymbol{\varphi}^{\mathsf{T}}(t')\,\hat{\boldsymbol{\theta}} \right] \boldsymbol{\varphi}^{\mathsf{T}}(t') = \mathbf{0}^{\mathsf{T}} .$$

Taking the transpose then gives

$$\left[\int_0^t \mathrm{d}t'\, \mathrm{e}^{-\lambda'(t-t')}\, \boldsymbol{\varphi}(t')\,\boldsymbol{\varphi}^{\mathsf{T}}(t') \right] \hat{\boldsymbol{\theta}}(t) \equiv \boldsymbol{P}^{-1}(t)\,\hat{\boldsymbol{\theta}}(t) = \int_0^t \mathrm{d}t'\, \mathrm{e}^{-\lambda'(t-t')}\, \boldsymbol{\varphi}(t')\, y(t') .$$

We thus have

$$\begin{aligned}\boldsymbol{P}^{-1}(t) &= \int_0^t \mathrm{d}t'\, \mathrm{e}^{-\lambda'(t-t')}\ \boldsymbol{\varphi}(t')\,\boldsymbol{\varphi}^{\mathsf{T}}(t') \\ &= \mathrm{e}^{-\lambda' t} \int_0^t \mathrm{d}t'\, \mathrm{e}^{\lambda' t'}\ \boldsymbol{\varphi}(t')\,\boldsymbol{\varphi}^{\mathsf{T}}(t')\,.\end{aligned}$$

Differentiating with respect to time gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{P}^{-1}(t) = -\lambda \boldsymbol{P}^{-1}(t) + \boldsymbol{\varphi}(t)\,\boldsymbol{\varphi}^{\mathsf{T}}(t)\,.$$

Then,

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{P}}{\mathrm{d}t} &= -\boldsymbol{P}\left(\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{P}^{-1}\right)\boldsymbol{P} \\ &= -\boldsymbol{P}\left(-\lambda \boldsymbol{P}^{-1} + \boldsymbol{\varphi}\,\boldsymbol{\varphi}^{\mathsf{T}}\right)\boldsymbol{P} \\ &= \lambda \boldsymbol{P} - \boldsymbol{P}\boldsymbol{\varphi}\,\boldsymbol{\varphi}^{\mathsf{T}}\boldsymbol{P}\,.\end{aligned}$$

If there is no input ($\boldsymbol{\varphi} = \boldsymbol{0}$), then $\dot{\boldsymbol{P}} = \lambda \boldsymbol{P}$, and $\boldsymbol{P}(t)$ diverges exponentially.

d. From Eq. (A.15), the Sherman-Morrison formula is

$$\left(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}\right)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}}{1 + \boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{u}}\,,$$

We apply this to $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{P}^{-1} = \boldsymbol{\varphi}\,\boldsymbol{\varphi}^{\mathsf{T}}$, with $\boldsymbol{A} \to \boldsymbol{P}^{-1}$, $\boldsymbol{u} \to \boldsymbol{\varphi}$ and $\boldsymbol{v}^{\mathsf{T}} \to \boldsymbol{\varphi}^{\mathsf{T}}$. First,

$$\boldsymbol{P}_{k+1}^{-1} = \boldsymbol{P}_k^{-1} + T_\mathrm{s}\left(\boldsymbol{\varphi}_{k+1}\,\boldsymbol{\varphi}_{k+1}^{\mathsf{T}}\right)\,.$$

If we now scale time by $t \to t/T_\mathrm{s}$, we can write

$$\boldsymbol{P}_{k+1} = \left(\boldsymbol{P}_k^{-1} + \boldsymbol{\varphi}_{k+1}\,\boldsymbol{\varphi}_{k+1}^{\mathsf{T}}\right)^{-1}$$

Then the Sherman-Morrison formula gives

$$\boldsymbol{P}_{k+1} = \boldsymbol{P}_k - \frac{\boldsymbol{P}_k\boldsymbol{\varphi}_{k+1}\,\boldsymbol{\varphi}_{k+1}^{\mathsf{T}}\boldsymbol{P}_k}{1 + \boldsymbol{\varphi}_{k+1}^{\mathsf{T}}\boldsymbol{P}_k\boldsymbol{\varphi}_{k+1}}\,.$$

Next, we discretize $\frac{\mathrm{d}\hat{\boldsymbol{\theta}}}{\mathrm{d}t} = \boldsymbol{P}\boldsymbol{\varphi}\varepsilon$:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{k+1} &= \hat{\boldsymbol{\theta}}_k + \boldsymbol{P}_{k+1}\boldsymbol{\varphi}_{k+1}\varepsilon_{k+1} \\ &= \hat{\boldsymbol{\theta}}_k + \left(\boldsymbol{P}_k - \frac{\boldsymbol{P}_k\boldsymbol{\varphi}_{k+1}\,\boldsymbol{\varphi}_{k+1}^{\mathsf{T}}\boldsymbol{P}_k}{1 + \boldsymbol{\varphi}_{k+1}^{\mathsf{T}}\boldsymbol{P}_k\boldsymbol{\varphi}_{k+1}}\right)\boldsymbol{\varphi}_{k+1}\varepsilon_{k+1} \\ &= \hat{\boldsymbol{\theta}}_k + \left(\frac{\boldsymbol{P}_k\boldsymbol{\varphi}_{k+1}(1 + \cancel{\boldsymbol{\varphi}_{k+1}^{\mathsf{T}}\boldsymbol{P}_k\boldsymbol{\varphi}_{k+1}}) - \cancel{\boldsymbol{P}_k\boldsymbol{\varphi}_{k+1}\boldsymbol{\varphi}_{k+1}^{\mathsf{T}}\boldsymbol{P}_k\boldsymbol{\varphi}_{k+1}}}{1 + \boldsymbol{\varphi}_{k+1}^{\mathsf{T}}\boldsymbol{P}_k\boldsymbol{\varphi}_{k+1}}\right)\varepsilon_{k+1} \\ &= \hat{\boldsymbol{\theta}}_k + \boldsymbol{L}_{k+1}\varepsilon_{k+1}\,,\end{aligned}$$

where we define, using the notation of Chapter 8, the Kalman observer gain

$$\boldsymbol{L}_{k+1} \equiv \frac{\boldsymbol{P}_k\boldsymbol{\varphi}_{k+1}}{1 + \boldsymbol{\varphi}_{k+1}^{\mathsf{T}}\boldsymbol{P}_k\boldsymbol{\varphi}_{k+1}}\,.$$

Notice that

$$\begin{aligned}\boldsymbol{P}_{k+1} &= \boldsymbol{P}_k - \frac{\boldsymbol{P}_k \boldsymbol{\varphi}_{k+1}\, \boldsymbol{\varphi}_{k+1}^{\mathsf{T}} \boldsymbol{P}_k}{1 + \boldsymbol{\varphi}_{k+1}^{\mathsf{T}} \boldsymbol{P}_k \boldsymbol{\varphi}_{k+1}} \\ &= \boldsymbol{P}_k - \boldsymbol{L}_{k+1} \boldsymbol{\varphi}_{k+1} \boldsymbol{P}_k \\ &= \left(\boldsymbol{I} - \boldsymbol{L}_{k+1} \boldsymbol{\varphi}_{k+1}^{\mathsf{T}}\right) \boldsymbol{P}_k \,.\end{aligned}$$

One subtlety is that the innovations $\varepsilon_{k+1} = y_{k+1} - \boldsymbol{\varphi}_{k+1}^{\mathsf{T}} \hat{\boldsymbol{\theta}}_k$. That is, the innovations use $\hat{\boldsymbol{\theta}}_k$ and not $\hat{\boldsymbol{\theta}}_{k+1}$. In fact, the issue is mainly one of notation. Our $\boldsymbol{\varphi}_{k+1}^{\mathsf{T}}$ depend on quantities at time k. Essentially, we are just using the forward Euler discretization, where all terms on the right-hand side are evaluated at time k. Physically, the innovation is the prediction at time $k+1$ based on the information available at time k, before the observation at $k+1$.
Finally, in unscaled units, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{P}^{-1} = \frac{\boldsymbol{\varphi}\, \boldsymbol{\varphi}^{\mathsf{T}}}{\xi_c^2} \,,$$

or

$$\boldsymbol{P}_{k+1}^{-1} = \boldsymbol{P}_k^{-1} + T_{\mathrm{s}} \frac{\boldsymbol{\varphi}_{k+1}\, \boldsymbol{\varphi}_{k+1}^{\mathsf{T}}}{\xi_c^2} = \boldsymbol{P}_k^{-1} + \frac{\boldsymbol{\varphi}_{k+1}\, \boldsymbol{\varphi}_{k+1}^{\mathsf{T}}}{\xi^2} \,, \,.$$

where $\xi^2 \equiv \xi_c^2 / T_{\mathrm{s}}$ is the discrete-time variance over T_{s}. The Sherman-Morrison formula then gives

$$\boldsymbol{P}_{k+1} = \boldsymbol{P}_k - \frac{\boldsymbol{P}_k \boldsymbol{\varphi}_{k+1}\, \boldsymbol{\varphi}_{k+1}^{\mathsf{T}} \boldsymbol{P}_k}{\xi^2 + \boldsymbol{\varphi}_{k+1}^{\mathsf{T}} \boldsymbol{P}_k \boldsymbol{\varphi}_{k+1}} \,.$$

Similarly,

$$\boldsymbol{L}_{k+1} \equiv \frac{\boldsymbol{P}_k \boldsymbol{\varphi}_{k+1}}{\xi^2 + \boldsymbol{\varphi}_{k+1}^{\mathsf{T}} \boldsymbol{P}_k \boldsymbol{\varphi}_{k+1}} \,.$$

10.11 Persistent excitation. Input signals that are not persistent can bias parameter estimates. Estimate the parameters of the FIR filter $y_k = b_0 u_k + b_1 u_{k-1} + \xi_k$, where $\xi_k \sim \mathcal{N}(0, 1)$. Using non-recursive least squares, find the asymptotic parameter estimates and associated covariance matrix for time steps $N \to \infty$ for a step input (what goes wrong?) and for a random input, $u_k \sim \mathcal{N}(0, 1)$. See Åström and Murray (2008).

Solution.

To match up with ordinary least-squares analysis in the form $y_k = \boldsymbol{\varphi}_k^{\mathsf{T}} \boldsymbol{\theta} + \varepsilon_k$, we write

$$\boldsymbol{\varphi} = \begin{pmatrix} u_k \\ u_{k-1} \end{pmatrix}, \qquad \boldsymbol{\theta} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}.$$

Then

$$\hat{\boldsymbol{\theta}} = \boldsymbol{P} \boldsymbol{\Phi}^{\mathsf{T}} \boldsymbol{Y} = \boldsymbol{P} \begin{pmatrix} \sum u_k\, y_k \\ \sum u_{k-1}\, y_k \end{pmatrix}, \qquad \boldsymbol{P}^{-1} = \begin{pmatrix} \sum u_k^2 & \sum u_k\, u_{k-1} \\ \sum u_k\, u_{k-1} & \sum u_{k-1}^2 \end{pmatrix}.$$

All sums are from 1 to N.

a. Step input ($u_k = 1$). Then

$$\boldsymbol{P}^{-1} = N\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix},$$

which is degenerate, implying infinite covariance, which means that we do not expect the parameters to converge to their correct values in practice. Another way to understand the problem is that the recursion relation becomes

$$y_k = (b_0 + b_1) + \xi_k\,,$$

so that there is no way to identify b_0 or b_1 individually. Only their sum is determined. (Actually, the initial conditions slightly break the degeneracy.)

b. Random input: Using $\langle u_k\rangle = 0$ and $\langle u_k^2\rangle = 1$, we have

$$\boldsymbol{P}^{-1} = \begin{pmatrix}\sum u_k^2 & \sum u_k\,u_{k-1}\\ \sum u_k\,u_{k-1} & \sum u_{k-1}^2\end{pmatrix} = N\begin{pmatrix}\langle u_k^2\rangle & \langle u_k\,u_{k-1}\rangle\\ \langle u_k\,u_{k-1}\rangle & \langle u_{k-1}^2\rangle\end{pmatrix} = N\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}.$$

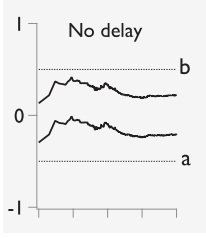
Similarly,

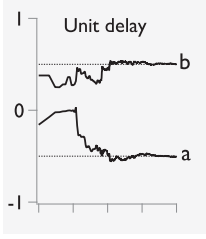
$$\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{Y} = \begin{pmatrix}\sum u_k\,y_k\\ \sum u_{k-1}\,y_k\end{pmatrix} = N\begin{pmatrix}\langle u_k\,(b_0u_k + b_1u_{k-1} + \xi_k)\rangle\\ \langle u_{k-1}\,(b_0u_k + b_1u_{k-1} + \xi_k)\rangle\end{pmatrix} = N\begin{pmatrix}b_0\\ b_1\end{pmatrix}.$$

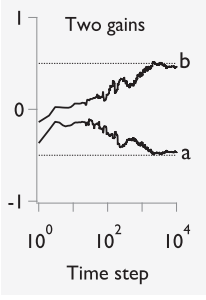
Thus,

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix}b_0\\ b_1\end{pmatrix}.$$

The conclusion is that, for this example, a random Gaussian input is persistent but not a step input.







10.12 Identification in closed-loop systems. Investigate $y_{k+1} = -ay_k + bu_k + \xi_k$, from Example 10.7, with $a = -\frac{1}{2}$ and $b = \frac{1}{2}$ and $1 \le k \le N$. Compare two feedback laws: $u_k = -\kappa y_k$ (no delay) and $u_k = -\kappa y_{k-1}$ (unit delay).

a. Show that the closed-loop system is stable for $-1 < \kappa < 3$ for feedback without delay and for $-1 < \kappa < 2$ for feedback with delay. (As usual, delay limits gain.)

b. Show that the inverse covariance matrix $\boldsymbol{P}^{-1} = \boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi} \to N\begin{pmatrix}\langle y^2\rangle & -\langle y\,u\rangle\\ -\langle y\,u\rangle & \langle u^2\rangle\end{pmatrix}$.

c. For $u_k = -\kappa y_k$, show that $\boldsymbol{P}^{-1}$ is degenerate at large k and hence that the least-squares estimate will not converge.

d. For $u_k = -\kappa y_{k-1}$, show that $\boldsymbol{P}^{-1}$ is invertible if the closed-loop dynamics is stable.

e. For $u_k = -\kappa_1 y_k$ or $-\kappa_2 y_k$, with 50% probability, show that $\boldsymbol{P}^{-1}$ is invertible.

f. Simulate the system and reproduce the graphs at left, showing lack of identifiability for feedback with no delay, convergence with a delay and also with no delay but two randomly alternating gains.

Solution.

a. The roots of the discrete transfer function must satisfy $|z| < 1$.

i. For feedback without delay, $y_{k+1} = -(a + b\kappa)y_k + \xi_k$. The Z-transform is

$$(z + a + b\kappa)y = \xi \quad \rightarrow \quad G_d(z) = \frac{y(z)}{\xi(z)} = \frac{1}{z + a + b\kappa}\,,$$

whose poles are at $z = -(a + b\kappa) = \frac{1}{2}(1 - \kappa)$, implying $-1 < \kappa < 3$.

ii. For feedback with unit delay, $y_{k+1} = -ay_k - b\kappa y_{k-1} + \xi_k$. The transfer function is

$$G_d(z) = \frac{z}{z^2 + az + b\kappa}\,,$$

whose poles are at

$$z = \tfrac{1}{2}\left(-a \pm \sqrt{a^2 - 4b\kappa}\right) \quad \rightarrow \quad z = \tfrac{1}{4}\left(1 \pm \sqrt{1 - 8\kappa}\right).$$

For $\kappa = -1$, we confirm $z = +1, -\frac{1}{2}$, which is one limit. The other is found by allowing complex roots, $z = \frac{1}{4}\left(1 \pm \mathrm{i}\sqrt{8\kappa - 1}\right)$, which implies $|z| = \frac{1}{16}(1+8\kappa-1) = \kappa/2$. Thus, the upper gain is $\kappa = 2$ and we require $-1 < \kappa < 2$.

b. We have $\boldsymbol{\varphi}_k^\mathsf{T} = (-y_k\ u_k)$ and, hence,

$$\boldsymbol{P}^{-1} = \sum_{k=1}^{N} \boldsymbol{\varphi}_k\,\boldsymbol{\varphi}_k^\mathsf{T} = \begin{pmatrix} \sum y_k^2 & -\sum y_k\,u_k \\ -\sum y_k\,u_k & \sum u_k^2 \end{pmatrix} \rightarrow N\begin{pmatrix} \langle y^2\rangle & -\langle y\,u\rangle \\ -\langle y\,u\rangle & \langle u^2\rangle \end{pmatrix}.$$

The last identity holds for feedback laws u_k that lead to time-invariant dynamics.

c. For $u_k = -\kappa y_k$, we have $\langle y\,u\rangle = -\kappa\langle y^2\rangle$ and $\langle u^2\rangle = \kappa^2\langle y^2\rangle$, so that

$$\boldsymbol{P}^{-1} \rightarrow N\langle y^2\rangle\begin{pmatrix} 1 & \kappa \\ \kappa & \kappa^2 \end{pmatrix},$$

which is clearly degenerate (det=0).

d. For $u_k = -\kappa y_{k-1}$, the cross-correlation $N\langle y\,u\rangle$ is

$$N\langle y\,u\rangle = \sum_k y_k\,u_k = -\kappa y_k\,y_{k-1} = -N\kappa\langle y\,y_{-1}\rangle\,.$$

We can evaluate $\langle yy_{-1}\rangle$ by multiplying the closed-loop dynamical equations by y_k, summing over k, and using the time invariance of the dynamics. Thus,

$$\sum_k y_{k+1}\,y_k = -a\sum y_k^2 - b\kappa\sum_k y_k y_{k-1} + \cancel{\sum_k \xi_k y_k}\,,$$

which implies

$$\langle y\,y_{-1}\rangle = -a\left\langle y^2\right\rangle - b\kappa\,\langle y\,y_{-1}\rangle\ .$$

Solving gives $\langle y\,y_{-1}\rangle = -\frac{a}{1+b\kappa}\langle y^2\rangle$, and the inverse covariance matrix is

$$\boldsymbol{P}^{-1} \rightarrow N\left\langle y^2\right\rangle\begin{pmatrix} 1 & \frac{a\kappa}{1+b\kappa} \\ \frac{a\kappa}{1+b\kappa} & \kappa^2 \end{pmatrix},$$

whose determinant is

$$N^2 \left\langle y^2 \right\rangle^2 \kappa^2 \left[1 - \left(\frac{a}{1+b\kappa} \right)^2 \right] \to N^2 \left\langle y^2 \right\rangle^2 \kappa^2 \left[1 - \left(\frac{1}{2+\kappa} \right)^2 \right] .$$

The latter expression is for $a = -\frac{1}{2}$ and $b = \frac{1}{2}$. The determinant thus vanishes only for $\kappa = -1$, which is also the stability limit. Since it will be close to zero as $\kappa \to -1$, we expect greater and greater fluctuations in estimation in this limit. By contrast, the variance stays finite for all $\kappa > 0$, even as the upper stability limit ($\kappa = 2$) is approached.

e. For $u_k = -\kappa_1 y_k$ or $-\kappa_2 y_k$, with 50% probability, we can write

$$\langle y\, u \rangle = -\frac{\left\langle y^2 \right\rangle}{2} (\kappa_1 + \kappa_2) .$$

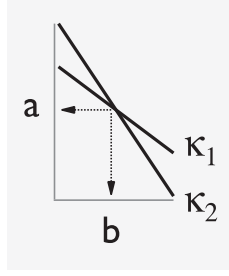
On the other hand,

$$\left\langle u^2 \right\rangle = \frac{\left\langle y^2 \right\rangle}{2} \left(\kappa_1^2 + \kappa_2^2 \right) .$$

Thus,

$$\boldsymbol{P}^{-1} \to \frac{N}{2} \left\langle y^2 \right\rangle \begin{pmatrix} 2 & \kappa_1 + \kappa_2 \\ \kappa_1 + \kappa_2 & \kappa_1^2 + \kappa_2^2 \end{pmatrix} ,$$

and the determinant is $(\frac{N}{2}\langle y^2 \rangle)^2 (\kappa_1 - \kappa_2)^2$, which vanishes only when $\kappa_1 = \kappa_2$. A geometrical interpretation is that when $\kappa = \kappa_1$, the linear combination $a + b\kappa_1$ is fixed and equals, say, c_1. The locus of points (a, b) satisfying $a + b\kappa_1 = c_1$ determines a line of slope $-\kappa_1$ in the a-b plane. Similarly, when $\kappa = \kappa_2$, another line, of slope $-\kappa_2$ is also determined. The intersection of the two lines gives a unique point in the a-b plane. See below.



10.13 Colored-noise. For $x_k = \theta x_{k-1} + \xi_k + a\xi_{k-1}$, with $\langle \xi_k \xi_\ell \rangle = \delta_{k\ell}$,

a. Find $\langle x^2 \rangle$ and $\langle x\, x_{-1} \rangle$ by following the suggestions in Example 10.8.
b. Simulate the system with $\theta = 0.5$ and $a = -0.5$ and analyze by RLS on the raw and filtered signals. Make plots such as the one shown in Example 10.9.
c. For unknown a, implement the extended RLS scheme from Example 10.10. Show numerically that $\hat{\theta} \to \theta$ and $\hat{a} \to a$ unless $a = -\theta$. Why does that case not work?

Solution.

a. Causality implies that $\langle x_{k-1}\,\xi_k\rangle = 0$: the past state does not affect the present noise. Then, with $\langle \xi_k^2\rangle = 1$, the covariances are as follows:

$$\begin{aligned}
\times\ \xi_k: &\qquad \langle x\xi\rangle = 0 + 1 + 0 = 1\\
\times\ \xi_{k-1}: &\qquad \langle x\xi_{-1}\rangle = \theta(1) + 0 + a(1) = \theta + a\\
\times\ x_k: &\qquad \left\langle x^2\right\rangle = \theta\,\langle x\,x_{-1}\rangle + 1 + a(\theta + a)\\
\times\ x_{k-1}: &\qquad \langle x\,x_{-1}\rangle = \theta\left\langle x^2\right\rangle + a
\end{aligned}$$

There are two coupled equations for $\langle x^2\rangle$ and $\langle x\,x_{-1}\rangle$:

$$\begin{pmatrix} 1 & -\theta \\ -\theta & 1 \end{pmatrix}\begin{pmatrix} \langle x^2\rangle \\ \langle x\,x_{-1}\rangle \end{pmatrix} = \begin{pmatrix} 1 + a(\theta + a) \\ a \end{pmatrix}$$

$$\begin{aligned}
\begin{pmatrix} \langle x^2\rangle \\ \langle x\,x_{-1}\rangle \end{pmatrix} &= \frac{1}{1-\theta^2}\begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix}\begin{pmatrix} 1 + a(\theta + a) \\ a \end{pmatrix}\\
&= \frac{1}{1-\theta^2}\begin{pmatrix} 1 + 2a\theta + a^2 \\ (\theta + a)(1 + \theta a) \end{pmatrix}.
\end{aligned}$$

b. The filtered signal for x is determined by $x_k^{\mathrm{f}} + a x_{k-1}^{\mathrm{f}} = x_k$. We substitute this into $x_k = \theta x_{k-1} + \xi_k + a\xi_{k-1}$, which gives

$$x_k^{\mathrm{f}} + a x_{k-1}^{\mathrm{f}} = \theta\left(x_{k-1}^{\mathrm{f}} + a x_{k-2}^{\mathrm{f}}\right) + \xi_k + a\xi_{k-1}\,.$$

Then we can isolate at times k and $k-1$:

$$\begin{aligned}
x_k^{\mathrm{f}} &= \theta x_{k-1}^{\mathrm{f}} + \xi_k\\
a x_{k-1}^{\mathrm{f}} &= a\left(\theta x_{k-2}^{\mathrm{f}} + \xi_{k-1}\right).
\end{aligned}$$

which must both hold and are in fact the same equation, just shifted by one time step. This new equation has a white-noise source.

c. Code is similar to previous section, except that no filtering is needed. We insert the estimate for ξ_{k-1} into the $\boldsymbol{\varphi}$ vector. When $a = -\theta$, we have that $x_k - \theta x_{k-1} = \xi_k - \theta\xi_{k-1}$, so that $x_k = \xi_k$. Thus, the inverse covariance matrix

$$\boldsymbol{P}_{\boldsymbol{k}}^{-1} = \sum_k \begin{pmatrix} x_{k-1}^2 & x_{k-1}\xi_{k-1} \\ x_{k-1}\xi_{k-1} & \xi_{k-1}^2 \end{pmatrix} \to N\left\langle x^2\right\rangle\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

is degenerate and not invertible. In the spirit of Problem 10.11, if we replace the θx_{k-1} term with θu_{k-1}, we can say that $u_k = x_k$ is not a persistent input. On the other hand, changing it slightly—e.g., to $u_k = x_{k-1}$ solves the problem.

10.14 Brownian particle with noisy observations. We seek to trap an overdamped particle diffusing in a liquid and measure its mobility and diffusion coefficients. In 1d, the position $x_k = x_{k-1} + \mu u_k + \nu_k$. The measurement $y_k = x_k + \xi_k$, reflecting the finite resolution of the microscope. The mobility is μ, the input u_k. The Gaussian noise terms have $\langle \nu_k\,\nu_\ell\rangle = \nu^2\delta_{k\ell}$ and $\langle \xi_k\xi_\ell\rangle = \xi_0^2\delta_{k\ell}$. The variance $\nu^2 = 2DT_{\mathrm{s}}$, with D the diffusion coefficient and T_{s} the sampling time.

a. For instantaneous observation-based feedback, $u_k = -\alpha_0 y_k$, eliminate x to find an equation for y and the noise. Find the correlation functions $\langle y^2 \rangle$ and $\langle y\, y_{-1} \rangle$ and use them to show that the least-squares mobility estimate $\hat{\mu} = \mu\left[1 + \frac{(2-\alpha)\xi^2}{1+2\alpha\xi^2}\right]$, where $\alpha \equiv \mu\alpha_0$ and $\xi^2 = \xi_0^2/\nu^2$. The estimate is biased for $\xi^2 \neq 0$.

b. For $u_k = -\alpha_0 y_{k-1}$, show that $\hat{\mu} = \mu$ for all ξ. Why does delaying the feedback eliminate the bias? (You need not solve for $\langle y^2 \rangle$, $\langle y\, y_{-1} \rangle$, and $\langle y\, y_{-2} \rangle$ to find $\hat{\mu}$.)

c. Use RLS to estimate μ recursively. Check that $\hat{\mu}$ converges to a biased value (calculate it) for no-delay feedback and to the correct value for unit delay. Include a recursive estimate of ν^2, based on Eq. (10.27). Why is $\hat{\nu}^2$ biased when $\hat{\mu} \neq \mu$?

d. For $\xi_k = 0$ and known μ, show numerically that choosing $\alpha \approx 0.47$ minimizes the position variance, with $\langle y^2 \rangle_{\min} \approx 2.4\,\nu^2$. For unknown μ, combine the RLS and simulation codes. Then try to adaptively control the particle variance by choosing the gain $(\alpha_0)_k = 0.47/\hat{\mu}_k$. Why does this algorithm fail? Propose a simple fix and then compare the observed variance with $\langle y^2 \rangle_{\min}$.

Solution.

a. Substituting $x_k = y_k - \xi_k$ and calculating covariances, we have

$$y_k = (1-\alpha)y_{k-1} + \nu_k + \xi_k - \xi_{k-1}$$

$$\begin{aligned}
\times\ \nu_k: &\qquad \langle y\nu \rangle = 0 + \nu^2 + 0 - 0 = \nu^2 \\
\times\ \xi_k: &\qquad \langle y\xi \rangle = 0 + 0 + \xi_0^2 - 0 = \xi_0^2 \\
\times\ \xi_{k-1}: &\qquad \langle y\xi_{-1} \rangle = (1-\alpha)\xi_0^2 + 0 + 0 - \xi_0^2 = -\alpha\xi_0^2 \\
\times\ y_k: &\qquad \left\langle y^2 \right\rangle = (1-\alpha)\langle y\, y_{-1} \rangle + \nu^2 + \xi_0^2 - (-\alpha\xi_0^2) \\
\times\ y_{k-1}: &\qquad \langle y\, y_{-1} \rangle = (1-\alpha)\left\langle y^2 \right\rangle + 0 + 0 - \xi_0^2,
\end{aligned}$$

The coupled equations for $\left\langle y^2 \right\rangle$ and $\langle y\, y_{-1} \rangle$ are

$$\begin{pmatrix} 1 & -(1-\alpha) \\ -(1-\alpha) & 1 \end{pmatrix} \begin{pmatrix} \left\langle y^2 \right\rangle \\ \langle y\, y_{-1} \rangle \end{pmatrix} = \begin{pmatrix} \nu^2 + (1+\alpha)\xi_0^2 \\ -\xi_0^2 \end{pmatrix}$$

$$\begin{aligned}
\begin{pmatrix} \left\langle y^2 \right\rangle \\ \langle y\, y_{-1} \rangle \end{pmatrix} &= \frac{1}{(2-\alpha)\alpha} \begin{pmatrix} 1 & 1-\alpha \\ 1-\alpha & 1 \end{pmatrix} \begin{pmatrix} \nu^2 + (1+\alpha)\xi_0^2 \\ -\xi_0^2 \end{pmatrix} \\
&= \frac{1}{(2-\alpha)\alpha} \begin{pmatrix} \nu^2 + 2\alpha\xi_0^2 \\ (1-\alpha)\nu^2 - \alpha^2\xi_0^2 \end{pmatrix}.
\end{aligned}$$

Finally, the least-squares estimate for μ is

$$\begin{aligned}
\hat{\mu} &= \left(\frac{1}{\alpha_0}\right) \frac{\langle (y_k - y_{k-1})(-y_{k-1}) \rangle}{\left\langle y_{k-1}^2 \right\rangle} \\
&= \left(\frac{1}{\alpha_0}\right) \frac{\left\langle y^2 \right\rangle - \langle y\, y_{-1} \rangle}{\langle y^2 \rangle}
\end{aligned}$$

$$= \left(\frac{1}{\alpha_0}\right) \frac{\alpha \left\langle y^2 \right\rangle + \xi_0^2}{\langle y^2 \rangle}$$

$$= \mu + \frac{\xi_0^2}{\alpha_0 \langle y^2 \rangle} = \mu + \frac{\xi_0^2 (2-\alpha)\alpha}{\alpha_0 (\nu^2 + 2\alpha\xi_0^2)} = \mu \left[1 + \frac{(2-\alpha)\xi^2}{1 + 2\alpha\xi^2}\right].$$

which is biased ($\hat{\mu} \neq \mu$) when $\xi^2 > 0$.

b. For delayed feedback $u_k = -\alpha_0 y_{k-1}$, the covariances are

$$y_k = y_{k-1} - \alpha y_{k-2} + \nu_k + \xi_k - \xi_{k-1}$$

$$\begin{aligned}
\times\ \nu_k: &\quad \langle y\nu \rangle = 0 + 0 + \nu^2 + 0 - 0 = \nu^2 \\
\times\ \xi_k: &\quad \langle y\xi \rangle = 0 + 0 + 0 + \xi_0^2 - 0 = \xi_0^2 \\
\times\ \xi_{k-1}: &\quad \langle y\xi_{-1} \rangle = \xi_0^2 + 0 + 0 + 0 - \xi_0^2 = 0 \\
\times\ y_k: &\quad \left\langle y^2 \right\rangle = \langle y\, y_{-1} \rangle - \alpha \langle y\, y_{-2} \rangle + \nu^2 + \xi_0^2 - 0 \\
\times\ y_{k-1}: &\quad \langle y\, y_{-1} \rangle = \left\langle y^2 \right\rangle - \alpha \langle y\, y_{-1} \rangle + 0 + 0 - \xi_0^2 \\
\times\ y_{k-2}: &\quad \langle y\, y_{-2} \rangle = \langle y\, y_{-1} \rangle - \alpha \left\langle y^2 \right\rangle + 0 + 0 - 0,
\end{aligned}$$

The coupled equations for $\left\langle y^2 \right\rangle$, $\langle y\, y_{-1} \rangle$, and $\langle y\, y_{-2} \rangle$ are

$$\begin{pmatrix} 1 & -1 & \alpha \\ -1 & 1+\alpha & 0 \\ \alpha & -1 & 1 \end{pmatrix} \begin{pmatrix} \left\langle y^2 \right\rangle \\ \langle y\, y_{-1} \rangle \\ \langle y\, y_{-2} \rangle \end{pmatrix} = \begin{pmatrix} \nu^2 + \xi_0^2 \\ -\xi_0^2 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} \left\langle y^2 \right\rangle \\ \langle y\, y_{-1} \rangle \\ \langle y\, y_{-2} \rangle \end{pmatrix} = \frac{1}{\alpha(1-\alpha)(2+\alpha)} \begin{pmatrix} \nu^2(1+\alpha) + 2\alpha\xi_0^2 \\ \nu^2 + \alpha^2\xi_0^2 \\ \nu^2(\alpha^2 + \alpha - 1) + \alpha^2\xi_0^2 \end{pmatrix}$$

$$= \frac{\nu^2}{\alpha(1-\alpha)(2+\alpha)} \begin{pmatrix} 1 + \alpha + 2\alpha\xi^2 \\ 1 + \alpha^2\xi^2 \\ \alpha^2 + \alpha - 1 + \alpha^2\xi^2 \end{pmatrix}.$$

The least-squares estimate for μ is

$$\hat{\mu} = \left(\frac{1}{\alpha_0}\right) \frac{\langle (y_k - y_{k-1})(-y_{k-2}) \rangle}{\left\langle y_{k-2}^2 \right\rangle}$$

$$= \left(\frac{1}{\alpha_0}\right) \frac{\langle y\, y_{-1} \rangle - \langle y\, y_{-2} \rangle}{\langle y^2 \rangle}.$$

But we can substitute directly $\langle y\, y_{-2} \rangle = \langle y\, y_{-1} \rangle - \alpha \left\langle y^2 \right\rangle$ to get

$$\hat{\mu} = \frac{\alpha \left\langle y^2 \right\rangle}{\alpha_0 \langle y^2 \rangle} = \mu.$$

Notice that this result is independent of the noise-correlation terms, and we did not need to solve for $\langle y^2 \rangle$, etc., to find $\hat{\mu}$ and to show that the estimate is

unbiased. Intuitively, the feedback has been delayed long enough that it does not correlate with any of the noise terms. Thus, the $\langle y\, y_{-2}\rangle$ equation does not involve any noise-correlation terms.

c. To estimate ν^2, we can calculate the innovations $\varepsilon_k \equiv y_k - y_{k-1} - \hat{\mu}\, u_{k-1}$. If $\hat{\mu} = \mu$, then $\varepsilon_k = \nu_k + \xi_k - \xi_{k-1}$, implying that $\hat{\nu}^2 = \langle e^2\rangle - 2\xi_0^2$. From Eq. (10.27),

$$\hat{v}_k = \begin{cases} 0 & k = 0 \\ \hat{v}_{k-1} + \varepsilon_k^2 & k = \{1, 2\} \\ \left(\frac{k-2}{k-1}\right)\hat{v}_{k-1} + \left(\frac{1}{k-1}\right)\varepsilon_k^2 & k \geq 3 \end{cases} .$$

But when $\hat{\mu}$ is biased, we cannot correctly subtract the deterministic parts of the motion to get the stochastic residuals. For feedback without delay, we can calculate the bias in $\hat{v}$ directly:

$$\begin{aligned} \varepsilon_k &= y_k - y_{k-1} + \hat{\alpha}\, y_{k-1} \\ &= y_k - y_{k-1} + \alpha\, y_{k-1} + \Delta\alpha\, y_{k-1} \\ &= \nu_k + \xi_k - \xi_{k-1} + \Delta\alpha\, y_{k-1}\,, \end{aligned}$$

where $\Delta\alpha \equiv \hat{\alpha} - \alpha$. Then

$$\begin{aligned} \left\langle e^2 \right\rangle &= \nu^2 + 2\xi_0^2 + (\Delta\alpha)^2 \left\langle y^2 \right\rangle - 2(\Delta\alpha)\langle y\xi\rangle \\ &= \nu^2 + 2\xi_0^2 + (\Delta\alpha)^2 \left\langle y^2 \right\rangle - 2(\Delta\alpha)\xi_0^2\,. \end{aligned}$$

In Part (a), we derived that $\Delta\alpha = \xi_0^2 / \left\langle y^2 \right\rangle$. Thus,

$$\begin{aligned} \hat{\nu}^2 &= \left\langle e^2 \right\rangle - 2\xi_0^2 = \nu^2 + (\Delta\alpha)\xi_0^2 - 2(\Delta\alpha)\xi_0^2 \\ &= \nu^2 - (\Delta\alpha)\xi_0^2\,. \end{aligned}$$

Thus, the bias in $\hat{\mu}$ (or $\hat{\alpha}$) leads to a bias for $\hat{\nu}^2$, as well.

d. For $\xi_k = 0$ and known μ, the variance is

$$\left\langle y^2 \right\rangle = \nu^2 \frac{1+\alpha}{\alpha(1-\alpha)(2+\alpha)}\,.$$

We minimize the variance by setting $\partial_\alpha \left\langle y^2 \right\rangle = 0$. This gives the cubic equation

$$\alpha^3 + 2\alpha^2 + \alpha - 1 = 0\,,$$

whose real solution is $\alpha^* = 0.465571 \approx 0.47$. The corresponding minimum variance is $\left\langle y^2 \right\rangle^* = 2.38898\,\nu^2 \approx 2.4\,\nu^2$.

For unknown μ, if we impose $(\alpha_0)_k = 0.47/\hat{\mu}_k$, then y_k diverges because the initial estimates of μ are bad enough that the $(\alpha_0)_k$ correspond to unstable dynamics. (Recall that the dynamics are stable for $0 < \alpha < 1$ and that $\alpha \equiv \alpha_0\mu$.) A simple solution is to use a fixed, "safe" gain for the first K time steps and then to use the value based on the adaptive estimate. Empirically, $K = 10$ works well for $0.1 < \mu < 10$.

Once divergences are prevented, the estimate $\hat{\mu}$ converges to within 10% of μ in typically < 100 time steps. The variance curve $\langle y^2 \rangle(\alpha)$ is broad enough about the minimum that there is negligible performance difference between the case with known μ and that with unknown μ. Of course, by using an optimal condition (minimum variance) as our goal, we minimize the sensitivity to deviations between the estimated parameters and their true values.

10.15 Colored noise subtleties. From the Wiener–Khintchine theorem, colored noise with temporal correlations has a non-flat power spectrum. A white-noise sequence ξ_k, with $\langle \xi_k \xi_\ell \rangle = \delta_{k\ell}$, that is filtered by a rational transfer function $C(z)$ has an output power spectrum density $\phi(\omega) = C(\mathrm{e}^{\mathrm{i}\omega T_s}) C(\mathrm{e}^{-\mathrm{i}\omega T_s})$. Conversely, under reasonable conditions, a measured power spectrum can be approximated by the output of a linear system H driven by ideal white noise (Åström, 2006a). Use this representation to show that if a noise source characterized by $C(z^{-1}, a)$ in the z-domain has zeros at a which is outside the unit circle in the complex z-domain, then the power spectrum of $C(z^{-1}, a^{-1})$ matches that of the original function. Illustrate for $C(z^{-1}) = 1 - az^{-1}$.

Solution.

This problem is longer to state than to solve! The power spectrum is, with $z = \mathrm{e}^{\mathrm{i}\omega T_s}$, equal to $C(z)\, C(z^{-1})$. Now, if $C(z)$ has a zero at $z = a$, it is clear that $C'(z) \equiv C(z^{-1})$ has a zero at $z = 1/a$. If $|a| > 1$, then $|a^{-1}| < 1$. For the example of $C(z) = 1 - az^{-1}$, $C(z) = 1 - az$ has a zero at $1/a$, as claimed.

Also, the power spectrum of C' is $C(z^{-1})\, C(z)$, which obviously matches the power spectrum of C. In effect, every stable response function has an unstable counterpart with identical magnitude plot vs. frequency response (but different phase response).

10.16 Estimate mobility of a trapped particle by correlation methods. We use the correlation method to study the case of a diffusing particle with unknown mobility, from Example 10.12. Use steady-state statistics (i.e., assume that you analyze a long time series where any initial or final conditions can be neglected).

a. Set up the problem (e.g., define all estimators, etc.), including the effects of measurement noise ($y_k = x_k + \xi_k$, with $\langle \xi_k \xi_\ell \rangle = \xi^2 \delta_{k\ell}$).
b. State carefully the simplifications that occur if $\xi^2 = 0$.
c. Derive Eq. (10.47) for the innovation correlations, assuming $\xi^2 = 0$.

Solution.

a. The equations for the physical system are

$$x_{k+1} = x_k + \mu u_k + \nu_k\,, \qquad u_k = -\alpha \hat{x}_k\,, \quad ,$$

with $\langle \nu_k \nu_\ell \rangle = \nu^2 \delta_{k\ell}$ and $\langle \xi_k \xi_\ell \rangle = \xi^2 \delta_{k\ell}$. With $\hat{y}_{k+1} = \hat{x}^-_{k+1}$, the estimators are

$$\hat{x}^-_{k+1} = \hat{x}_k + \mu' u_k\,, \qquad \hat{x}_{k+1} = \hat{x}^-_{k+1} + L(y_{k+1} - \hat{y}_{k+1})\,.$$

b. When the measurement noise $\xi^2 = 0$, then $y_k = x_k$. From Chapter 8, we also know that the steady-state Kalman gain L will tend to 0. That is, with no observational noise, the best estimate we can make is to use just the measurement and not at all the prediction. Then

$$\hat{y}_k = \hat{x}_k = \hat{x}_k^-$$

and the prediction $\hat{x}_{k+1}^-$ then becomes

$$\hat{x}_{k+1}^- = x_k + \mu' u_k \,.$$

The innovations are

$$\varepsilon_k = y_k - \hat{y}_k^- = x_k - \hat{x}_k = e_k \,.$$

That is, the innovations are the same as the state error in this simple example. More generally, they are not, and are even vectors of different dimension, since the dimension of $\boldsymbol{\varepsilon}$ equals the number of simultaneous measurements and the dimension of $\boldsymbol{e}$ is the number of state-vector elements.

c. Using the simplifications from (b), we write the recurrence relation for the innovations as

$$\begin{aligned}
\varepsilon_{k+1} &= x_{k+1} - \hat{x}_k^- \\
&= (x_k + \mu u_k + \nu_k) - (x_k + \mu' u_k) \\
&= (\mu - \mu') u_k + \nu_k \\
&= +(\Delta\mu)\,\alpha x_k + \nu_k \,.
\end{aligned}$$

We can now calculate the correlations:

$$\begin{aligned}
\langle \varepsilon_{k+\ell}\, \varepsilon_k \rangle &= \langle [(\Delta\mu)\,\alpha x_{k+\ell} + \nu_{k+\ell}]\,[(\Delta\mu)\,\alpha x_k + \nu_k] \rangle \\
&= \cancelto{\text{small}}{(\Delta\mu)^2\alpha^2 \langle x_{k+\ell}\, x_k \rangle} + (\Delta\mu)\alpha \langle x_{k+\ell}\, \nu_k \rangle + \cancelto{0}{(\Delta\mu)\alpha \langle x_k\, \nu_{k+\ell} \rangle} + \cancelto{0}{\langle \nu_{k+\ell}\, \nu_k \rangle} \\
&\approx (\Delta\mu)\alpha \,\langle x_{k+\ell}\, \nu_k \rangle \,.
\end{aligned}$$

We evaluate successively the correlations $\langle x_{k+\ell}\, \nu_k \rangle$:

$$\begin{aligned}
\langle x_{k+1}\, \nu_k \rangle &= (1 - \alpha\mu) \cancelto{0}{\langle x_k\, \nu_k \rangle} + \langle \nu_k^2 \rangle = \nu^2 \\
\langle x_{k+2}\, \nu_k \rangle &= (1 - \alpha\mu) \langle x_{k+1}\, \nu_k \rangle + \cancelto{0}{\langle \nu_{k+1}\, \nu_k \rangle} = (1 - \alpha\mu)\nu^2 \\
\langle x_{k+3}\, \nu_k \rangle &= (1 - \alpha\mu) \langle x_{k+2}\, \nu_k \rangle + \cancelto{0}{\langle \nu_{k+2}\, \nu_k \rangle} = (1 - \alpha\mu)^2 \nu^2 \\
&\vdots \\
\langle x_{k+\ell}\, \nu_k \rangle &= \cdots = (1 - \alpha\mu)^{\ell - 1}\, \nu^2 \,.
\end{aligned}$$

Thus,

$$\langle \varepsilon_{k+\ell}\, \varepsilon_k \rangle = (\Delta\mu)\,\alpha\nu^2 (1 - \alpha\mu)^{\ell - 1} + O(\Delta\mu^2) \,,$$

which vanishes when the correct mobility is used to make predictions.

10.17 LMS and a variant. Explore the system from Example 10.13.

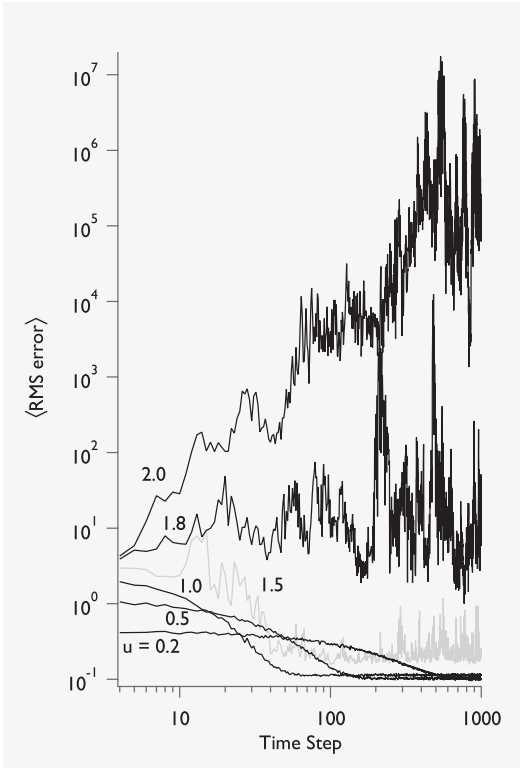
a. Implement the standard LMS adaptive filter, and reproduce the associated plots.
b. Let $u_k \sim \mathcal{N}(0, u^2)$. Keep $\gamma = 0.1$. What happens as the input amplitude u increases and why? Plot RMS convergence error vs time, for representative values of u.
c. The normalized LMS (n-LMS) algorithm is $\hat{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_k + \gamma(\frac{\boldsymbol{\varphi}_k}{\alpha+\boldsymbol{\varphi}_k^T\boldsymbol{\varphi}_k})(y_k - \boldsymbol{\varphi}_k^T\hat{\boldsymbol{\theta}}_k)$. Justify the algorithm when $\alpha = 0$. Why add the small parameter α?
d. Explore n-LMS convergence. Try changing the input level.

Solution.

a. See book website for code.
b. For $\gamma = 0.1$, as the amplitude u is increased, the error decay time decreases until about $u = 1$. Then it first becomes unstable to noise (less averaging in the 1000 trials done here). Then at about $u = 1.8$, it becomes unstable.

   This value is lower than the prediction from the stability-limit given in the text. That limit would suggest $u_{\max} = 1/\sqrt{2\gamma} \approx 2.2$. The difference results from using the variance of the u_k series to approximate the actual values of the 4 previous values u_k, u_{k-1}, u_{k-2}, u_{k-3}. The fluctuations help build up the instability in the b coefficients.

   See graphs below, labeled with the value of u.



c. "Bottom-up" argument: In the text, we argued that the stability limit is

$$0 < \gamma < \frac{2}{\boldsymbol{\varphi}_k^\mathsf{T} \boldsymbol{\varphi}_k} .$$

Here, we simply rescale $\gamma \to \frac{\gamma}{\boldsymbol{\varphi}_k^\mathsf{T} \boldsymbol{\varphi}_k}$. We then add the α term to prevent the possibility that noise on small inputs leads to a γ that is too large.

A "top-down" argument starts from Eq. (10.49). To derive the ordinary LMS algorithm, we made the approximation

$$\frac{\boldsymbol{P}_k}{1 + \boldsymbol{\varphi}_k^\mathsf{T} \boldsymbol{P}_k \boldsymbol{\varphi}_k} \approx \gamma ,$$

From the batch LS algorithm, we know that

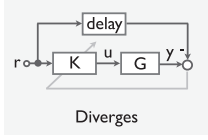
$$\boldsymbol{P}_k^{-1} = \sum_{i=1}^{k} \boldsymbol{\varphi}_i \, \boldsymbol{\varphi}_i^\mathsf{T} .$$

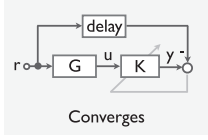
We now assume that the $\boldsymbol{\varphi}_k^\mathsf{T} \boldsymbol{P}_k \boldsymbol{\varphi}_k$ term in the denominator leads to a roughly constant numerical factor $\gamma = 1/(1 + \boldsymbol{\varphi}_k^\mathsf{T} \boldsymbol{P}_k \boldsymbol{\varphi}_k)$. In the numerator, we approximate the covariance with just the last term in the sum, $\boldsymbol{P}_k^{-1} \approx \boldsymbol{\varphi}_k \, \boldsymbol{\varphi}_k^\mathsf{T} \approx (\boldsymbol{\varphi}_k^\mathsf{T} \boldsymbol{\varphi}_k) \, \boldsymbol{I}$, we have

$$\boldsymbol{P}_k^{-1} \approx \boldsymbol{\varphi}_k \, \boldsymbol{\varphi}_k^\mathsf{T} \quad \to \quad \frac{\boldsymbol{P}_k}{1 + \boldsymbol{\varphi}_k^\mathsf{T} \boldsymbol{P}_k \boldsymbol{\varphi}_k} \approx \gamma \boldsymbol{P}_k \approx \gamma \left(\boldsymbol{\varphi}_k^\mathsf{T} \boldsymbol{\varphi}_k \right)^{-1} \boldsymbol{I} .$$

The approximations turn out to be reasonable if the eigenvalues of the full $\boldsymbol{P}_k$ are not too unequal.

d. Now, since numerator and denominator are $O(u^2)$, we can use arbitrarily large values of the input. For small values of input, the behavior will revert back to the ordinary LMS algorithm, with an effective learning rate of γ/α.

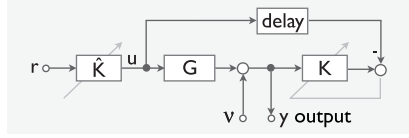




10.18 Feedforward control using LMS. In the simplest architecture for adaptive inverse feedforward control, shown at top left, the LMS algorithm does not converge.

a. The reversed configuration, shown bottom left, does converge, as suggested by Example 10.14. Implement and make a version of the figures in that example.

b. Consider the more elaborate scheme depicted below. Why should it work?



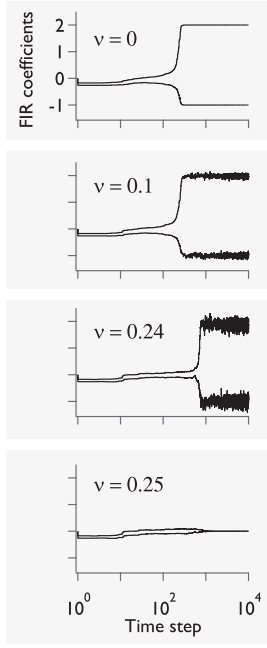
c. Simulate the above example, showing that you can control the system.

d. Add a disturbance $v \sim \mathcal{N}(0, v^2)$ and show that the above LMS scheme converges for small disturbances but not for large ones. An even more elaborate

scheme – copy the system and find the controller offline – is needed to both follow a reference and compensate for a disturbance. See Widrow and Walach (1996).

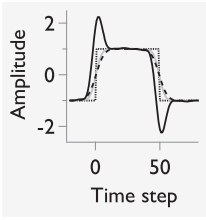
Solution.

a. See website for code.
b. In the absence of the disturbance ν, the right-hand part of the scheme is the same as in part (a). The difference is that we put the reference through the current estimate of the filter and use that filtered signal as the input to the system-controller combination of (a). Now we both use the LMS algorithm directly on the controller output and we run the input through the controller before it enters the system.
c. See website for code.
d. The appearance of the controller K and its copy $\hat{K}$ accentuates the nonlinearity. Below, we see that there is a subcritical bifurcation as a function of noise strength ν. We have noisy convergence up to $\nu = 0.24$ and collapse to zero for the filter coefficients for $\nu = 0.25$. The threshold levels are stochastic. We illustrate below the convergence (or lack thereof) for different values of ν.



10.19 Feedforward with preview. Adaptive feedforward filters can have zero-phase-lag output. Of course, to obey causality, you need to know the reference signal in advance. If the desired waveform is periodic, then you do. We use the feedforward architecture from the previous problem, without disturbances. First, shift the desired reference N steps into the future. Then track this reference with a delay of N steps.

a. To understand the timing, let the G dynamics be a simple delay of Δ. Let the reference be a square wave. Your code should find that the required feedforward filter is zero, except for the "advanced weight," which is one.
b. Next, filter the reference to avoid aliasing. The figure below uses a binomial smoothing filter with $n + 1 = 51$ coefficients (or "taps"), with weights $2^{-n}\binom{n}{k}$. The filter is symmetric (acausal), to follow the reference with no phase shift. Use a square wave (dotted line) as input to generate the filtered reference (dashed line).
c. Input the filtered reference into an LMS algorithm, using the timing from (a). Reproduce the figure at left, for $y_k = (1 - a)y_{k-1} + au_k$, with $a = 0.1$ and a 21-tap LMS filter. The learning rate $\gamma = 0.05$, and $\Delta = 10$. The input (heavy line) leads the output so that the phase-shifted output is centered on the reference. Filtering the reference with more taps reduces the required amplitude for u.



Solution.

The coding is straightforward in principle, but the numerics easily blow up. Using smaller γ can help, as can smaller number of coefficients for the LMS filter. See adaptive filtering books, e.g. Widrow and Walach (1996), for tricks to make the numerics more robust.

10.20 Learning an unknown charge. One step in the reinforcement-learning example of Section 10.4.2 is to parameterize the knowledge of the unknown charge gained from observations x_k and inputs u_k. In particular, given the diffusive dynamics of Eq. (10.60), use Bayes' theorem to prove the update law $\theta_{k+1} = \theta_k + (x_{k+1} - x_k)u_k$ stated in Eq. (10.63). Hint: show that you can write $\frac{1}{2}(1 + \beta_k) = \frac{\exp(b\theta_k)}{\exp(b\theta_k)+\exp(-b\theta_k)}$. You might want to start by showing the formula works for $k = 1$, given that $\theta_0 = 0$.

Solution.

Using Bayes' theorem, we can write, with $x^k \equiv \{x_k, x_{k-1}, \ldots, x_0\}$ and $P_k(b) \equiv P(b|x^k)$,

$$P(b|x^{k+1}, u_k) = \frac{p(x_{k+1}|x_k, u_k, b)\, P_k(b)}{p(x_{k+1}|x_k, u_k)}$$

$$= \frac{p(x_{k+1}|x_k, u_k, b)\, P_k(b|x^k)}{p(x_{k+1}|x_k, u_k, b = 1)\, P_k(b) + p(x_{k+1}|x_k, u_k, b = -1)\, P_k(-b)} .$$

The likelihood function

$$p(x_{k+1}|x_k, u_k, b) \propto \exp\left[-\tfrac{1}{2}(x_{k+1} - x_k - bu_k)^2\right] .$$

As suggested, we parametrize $P_k(b)$ by

$$P_k(b) = \frac{\exp(b\theta_k)}{\exp(b\theta_k) + \exp(-b\theta_k)} .$$

We also follow the hint and compute first the $k = 1$ case, given $k = 0$ and given $P(b = \pm 1|x_0) = \frac{1}{2}$. Then,

$$P(b|x_1, x_0, u_0) = \frac{\frac{1}{2}\exp\left[-\frac{1}{2}(x_1 - x_0 - bu_0)^2\right]}{\frac{1}{2}\exp\left[-\frac{1}{2}(x_1 - x_0 - bu_0)^2\right] + \frac{1}{2}\exp\left[-\frac{1}{2}(x_1 - x_0 + bu_0)^2\right]}$$

$$= \frac{\exp(b\Delta\theta_0)}{\exp(b\Delta\theta_0) + \exp(-b\Delta\theta_0)} ,$$

where $\Delta\theta_0 \equiv (x_1 - x_0)u_0 = \theta_1$. We have cancelled factors of $\frac{1}{2}$ and $\exp[-\frac{1}{2}((x_1 - x_0)^2 + u_0^2)]$. This proves the formula for $k = 1$.

For general k, the proof is similar:

$$P(b|x^{k+1}, u_k) = \frac{\exp\left[-\frac{1}{2}(x_{k+1} - x_k - bu_k)^2\right] P_k(b)}{\exp\left[-\frac{1}{2}(x_{k+1} - x_k - bu_k)^2\right] P_k(b) + \exp\left[-\frac{1}{2}(x_{k+1} - x_k + bu_k)^2\right] P_k(-b)}$$

$$= \frac{\exp(b\Delta\theta_k)\, P_k(b)}{\exp(b\Delta\theta_k)\, P_k(b) + \exp(-b\Delta\theta_k)\, P_k(-b)} .$$

But

$$P_k(\pm b) = \frac{\exp(\pm b\theta_k)}{\exp(b\theta_k) + \exp(-b\theta_k)} .$$

Substituting $P_k(\pm b)$ and cancelling the denominator then gives

$$P(b|x^{k+1}, u_k) = \frac{\exp(b\Delta\theta_k)\, \exp(b\theta_k)}{\exp(b\Delta\theta_k)\, \exp(b\theta_k) + \exp(-b\Delta\theta_k)\, \exp(-b\theta_k)}$$

$$= \frac{\exp(b\theta_{k+1})}{\exp(b\theta_{k+1}) + \exp(-b\theta_{k+1})} ,$$

where

$$\theta_{k+1} = \theta_k + \Delta\theta_{k+1} = \theta_k + (x_{k+1} - x_k)u_k .$$

This proves the requested update formula for θ_k.

10.21 Control of a diffusing particle with unknown sign of charge.

a. Derive the single-stage optimization results in Eqs. (10.64) and (10.66) by averaging over both b and ν_1 and using the equation of motion for x_2.

b. Show that $J^* = 2 + \min_{u_0}[(1+R)u_0^2 - \frac{f_+ + f_-}{2(1+R)}]$, with $f_\pm = \langle(\nu_0 \pm u_0)^2 \tanh^2(\nu_0 \pm u_0)u_0\rangle_{\nu_0}$.

c. For large R, assume that u_0 is small. Taylor expand to show $f \approx 3u_0^2 - 4u_0^4$. Deduce Eq. (10.68) and the expression for u_0^*.

d. For small R, assume $u_0 \gg 1$ and show that $(u_0^2)^* \sim -\frac{1}{2}\ln R$, dropping constants and logarithmic corrections. Hint: show, for $x \gg 1$, that $\tanh^2 x \sim 1 - 4\,\mathrm{e}^{-2|x|}$.

e. For $\beta_0 \neq 0$ (partial knowledge of the sign of the charge) and $x_0 \neq 0$ (biased initial position), show that the bifurcation is biased by the analog of an external field equal to $2\beta_0 x_0$. Why must β_0 and x_0 *both* be non-zero to have a finite field?

Solution.

a. The one-stage cost-to-go function is

$$\begin{aligned}
J_1(x_1,\beta_1,u_1) &= \left\langle x_2^2 + Ru_1^2 \right\rangle_{b,\nu} \\
&= \left\langle (x_1 + bu_1 + \nu_1)^2 + Ru_1^2 \right\rangle_{b,\nu} \\
&= \left\langle (x_1 + bu_1 + \nu_1)^2 \right\rangle_{b,\nu} + Ru_1^2 \\
&= \left(\frac{1+\beta_1}{2}\right)\left\langle (x_1 + u_1 + \nu_1)^2 \right\rangle_{\nu} + \left(\frac{1-\beta_1}{2}\right)\left\langle (x_1 - u_1 + \nu_1)^2 \right\rangle_{\nu} + Ru_1^2 \\
&= \left(\frac{1+\beta_1}{2}\right)(x_1 + u_1)^2 + \left(\frac{1-\beta_1}{2}\right)(x_1 - u_1)^2 + Ru_1^2 + 1 \\
&= (1+R)u_1^2 + x_1^2 + 2\beta_1 x_1 u_1 + 1\,.
\end{aligned}$$

Then

$$\frac{\partial J_1}{\partial u_1} = 2(1+R)u_1 + 2\beta_1 x_1 = 0 \qquad \Longrightarrow \qquad u_1^* = -\frac{\beta_1 x_1}{1+R}$$

Substituting u_1 into $J_1(x_1,\beta_1,u_1)$ gives

$$\begin{aligned}
J_1(x_1,\beta_1) &= J_1(x_1,\beta_1,u_1^*) \\
&= (1+R)(u_1^*)^2 + x_1^2 + 2\beta_1 x_1 u_1^* + 1 \\
&= (1+R)\left(\frac{\beta_1 x_1}{1+R}\right)^2 + x_1^2 - 2\beta_1 x_1 \left(\frac{\beta_1 x_1}{1+R}\right) + 1 \\
&= \frac{1+R-\beta_1^2}{1+R} x_1^2 + 1\,.
\end{aligned}$$

b. Starting from the logic given in the text or, more formally, from the Bellman equations, the two-stage cost-to-go function is given by

$$J^* = \min_{u_0}\left[Ru_0^2 + \langle J_1(x_1,\beta_1)\rangle_{x_1,\beta_1,b,\nu_0}\right].$$

In principle $J^* = J^*(x_0,\beta_0)$, but the problem specifies $x_0 = \beta_0 = 0$. We use our result on the one-stage problem to proceed. Defining $J_0(u_0)$ by J^* = minimum of $J_0(u_0)$ over u_0, we have

$$\begin{aligned}
J_0(u_0) &= \left[Ru_0^2 + \langle J_1(x_1,\beta_1)\rangle_{x_1,\beta_1,b,\nu_0}\right] \\
&= Ru_0^2 + \left\langle \frac{1+R-\beta_1^2}{1+R} x_1^2 \right\rangle_{b,\nu_0} + 1 \\
&= Ru_0^2 + \left\langle x_1^2 \right\rangle_{b,\nu_0} - \frac{1}{1+R}\left\langle x_1^2 \beta_1^2 \right\rangle_{b,\nu_0} + 1
\end{aligned}$$

We use $\langle x_1^2\rangle = \langle (bu_0 + \nu_0)^2\rangle = u_0^2 + 1$ and also $\beta_1 = \tanh\theta_1$, with $\theta_1 = x_1 u_0 = (bu_0 + \nu_0)u_0$. Then,

$$J_0(u_0) = 2 + (1+R)u_0^2 - \frac{f(u_0)}{1+R}\,,$$

with

$$\begin{aligned} f(u_0) &= \left\langle x_1^2 \beta_1^2 \right\rangle_{b,\nu_0} \\ &= \left\langle (bu_0 + \nu_0)^2 \, \tanh^2(bu_0 + \nu_0) u_0 \right\rangle_{b,\nu_0} \\ &= \frac{1}{2} \left\langle (u_0 + \nu_0)^2 \, \tanh^2(u_0 + \nu_0) u_0 + (-u_0 + \nu_0)^2 \, \tanh^2(-u_0 + \nu_0) u_0 \right\rangle_{\nu_0} . \end{aligned}$$

In the last line, we carry out the average over b, recalling that at stage 0, we have $\beta_0 = 0$, so that we just have to use a factor of $\frac{1}{2}$ for each case.

c. For small R, we assume we can Taylor expand about $u_0 = 0$, using $\tanh x \approx x - \frac{1}{3}x^3$ and thus $\tanh^2 x \approx x^2 - \frac{2}{3}x^4$. Before averaging, we have

$$f(u_0, \nu_0) = \nu_0^4 \, u_0^2 + \left(6\nu_0^2 - \frac{2}{3}\nu_0^6 \right) u_0^4 + O(u_0^6) .$$

Recalling that $\langle \nu_0^2 \rangle = 1$, $\langle \nu_0^4 \rangle = 3$, and $\langle \nu_0^6 \rangle = 15$, we find $f(u_0) = 3u_0^2 - 4u_0^4$. Then substituting into the expression for J_0 gives

$$\begin{aligned} J_0(u_0) &= 2 + (1 + R)u_0^2 - \frac{f(u_0)}{1 + R} \\ &= 2 + \left((1 + R) - \frac{3}{1 + R} \right) u_0^2 + \left(\frac{4}{1 + R} \right) u_0^4 . \end{aligned}$$

We find the value of u_0^* by solving

$$\frac{\partial J_0}{\partial u_0^2} = \left((1 + R) - \frac{3}{1 + R} \right) + \left(\frac{8}{1 + R} \right) \left(u_0^2 \right)^* = 0 ,$$

which gives

$$\left(u_0^2 \right)^* = \frac{\frac{3}{1+R} - (1 + R)}{\frac{8}{1+R}} = \frac{2 - 2R - R^2}{8} .$$

The roots u_0^* are real for $R < R^* = \sqrt{3} - 1$. Otherwise, the minimum is at $u^* = 0$.

d. In the small-R limit, we assume $u_0^2 \gg 1$. We use the asymptotic expansion

$$\tanh^2 x = \left(\frac{e^x - e^{-x}}{e^x + e^{-x}} \right)^2 = \left(\frac{1 - e^{-2x}}{1 + e^{-2x}} \right)^2 \sim 1 - 4\, e^{-2x} .$$

Repeating for $-x$ leads to

$$\tanh^2 x \sim 1 - 4\, e^{-2|x|} .$$

Then, assuming that $u_0 \gg \nu_0$, so that $\tanh(u_0 + \nu_0)u_0 \approx \tanh u_0^2$, we have

$$\begin{aligned} f(u_0) &= \frac{1}{2} \left\langle (u_0 + \nu_0)^2 \, \tanh^2(u_0 + \nu_0) u_0 + (-u_0 + \nu_0)^2 \, \tanh^2(-u_0 + \nu_0) u_0 \right\rangle_{\nu_0} \\ &\sim \frac{1}{2} \left\langle (u_0 + \nu_0)^2 \, \tanh^2 u_0^2 + (-u_0 + \nu_0)^2 \, \tanh^2 u_0^2 \right\rangle_{\nu_0} \\ &\sim \left(1 - 4\, e^{-2u_0^2} \right) \left(u_0^2 + 1 \right) . \end{aligned}$$

Then, since $R \ll 1$ and $u_0 \gg 1$, we have

$$\begin{aligned} J_0(u_0) &= 2 + (1+R)u_0^2 - \frac{f(u_0)}{1+R} \\ &\sim 2 + \left(1 + R - \frac{1}{1+R}\right)u_0^2 - \frac{1}{1+R} + \frac{4}{1+R}\,\mathrm{e}^{-2u_0^2}\left(u_0^2+1\right) \\ &\sim 1 + R + 2Ru_0^2 + 4u_0^2\,\mathrm{e}^{-2u_0^2}\,. \end{aligned}$$

Taking the derivative with respect to u_0^2 gives

$$\frac{\partial J_0}{\partial(u_0^2)} = 2R + (4 - 8u_0^2)\,\mathrm{e}^{-2u_0^2} \approx 2R - 8u_0^2\,\mathrm{e}^{-2u_0^2} = 0\,.$$

Then

$$\frac{R}{4} = \left(u_0^2\right)^*\,\mathrm{e}^{-2(u_0^2)^*}\,,$$

implying

$$\ln\frac{4}{R} = 2(u_0^2)^* + \text{logarithmic corrections}\,,$$

Thus,

$$u_0^* \sim \pm\sqrt{-\tfrac{1}{2}\ln R}\,,$$

The approximation is relatively crude, as we drop the $\ln 2$ inside the square root as well as the logarithmic corrections. But these have only a small effect on the final result.

e. Now we start with the particle at x_0 and an initial knowledge of its charge β_0. We re-evaluate $J_0(u_0)$:

$$J_0(u_0) = Ru_0^2 + \left\langle x_1^2\right\rangle_{b,\nu_0} - \frac{1}{1+R}\left\langle x_1^2\beta_1^2\right\rangle_{b,\nu_0} + 1$$

Now, the average of x_1^2 is given by the same expression we derived for stage 2 when picking u_1.

$$\left\langle x_1^2\right\rangle_{b,\nu} = \left\langle (x_0 + bu_0 + \nu_0)^2\right\rangle_{b,\nu} = x_0^2 + u_0^2 + 2\beta_0 x_0 u_0 + 1$$

The average of $x_1^2\beta_1^2$ is

$$\begin{aligned} f(u_0, x_0, \beta_0) &= \left\langle x_1^2\beta_1^2\right\rangle_{b,\nu_0} \\ &= \frac{1}{2}\Big\langle (x_0 + u_0 + \nu_0)^2\,\tanh^2(u_0+\nu_0)u_0 \\ &\qquad + (x_0 - u_0 + \nu_0)^2\,\tanh^2(-u_0+\nu_0)u_0\Big\rangle_{\nu_0}\,. \end{aligned}$$

By inspection, we can see that the lowest order terms are $O(u_0^2)$. Thus, the average will give higher-order corrections to the field $O(u_0^3)$. In fact, a calculation using symbolic algebra gives

$$f(u_0, x_0, \beta_0) = \left(3 + x_0^2\right) u_0^2 + 6 x_0 \beta_0 u_0^3 - \left(4 + x_0^2\right) u_0^4 .$$

Putting together the averages, we find the modified "free energy" to $O(u_0^4)$:

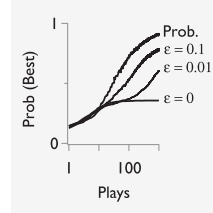
$$J_0(u_0) = \left(2 + x_0^2\right) + (2\beta_0 x_0)\, u_0 + \left(1 + R - \frac{3 + x_0^2}{1 + R}\right) u_0^2 - \left(\frac{6 x_0 \beta_0}{1 + R}\right) u_0^3 + \left(\frac{4 + x_0^2}{1 + R}\right) u_0^4 .$$

The linear term in u_0 implies an effective field $2\beta_0 x_0$. (Actually, this is all the question asks for. We did not need to calculate the higher-order terms.) The field leads to an imperfect (symmetry-broken) bifurcation. The field-free critical point is now at $R^* = \sqrt{3 + x_0^2} - 1$. Thus, starting at x_0 increases the value of R where it becomes profitable to use control: since we need to move the particle, it can be worth it to exert control even if it is relatively costly.
The question also asks why both β_0 and x_0 must be non-zero. Let us consider what happens if only one of the two is zero:

- $x_0 = 0, \beta_0 \neq 0$: Imagine, for simplicity, that $\beta_0 = 1$. If we start and want to end up at the same point, it does not matter which direction we push the particle.
- $\beta_0 = 0, x_0 \neq 0$: We know which way we want to move. But if we try to push the particle in a particular direction, there is a 50-50 chance it will move in the opposite direction. Thus, again, it does not matter which direction we choose.

On the other hand, if we know which direction we need to push the particle ($x_0 \neq 0$) and if we know something about the sign of the charge ($\beta_0 \neq 0$), *then* there will be a preferred direction (sign) for u_0. This is the meaning of the external field $2\beta_0 x_0$.

10.22 Multi-armed bandits. Consider n slot machines ("bandits"), each paying a reward $J_i \sim \mathcal{N}(Q^{(i)}, 1)$, with average payout $Q^{(i)} \sim \mathcal{N}(0, 1)$: On average, some bandits pay out, while others take money. Given an infinite number of trials, we could determine the payout of each bandit precisely and then play the best ever after. With a finite number of trials, we trade off finding the best bandit with having time to play it. Here, we play 10 bandits 1000 times, repeating for 2000 Monte Carlo trials. We explore different strategies, with the goal of reproducing the figure at right. The estimate of the average payout $\hat{Q}_k^{(i)}$ of bandit i at time k averages the payout over the $n_k^{(i)}$ times it has been observed at time step k. The estimated variance of bandit i at time k is $\hat{\sigma}_i{}^2 = 1/(\text{times played})$. The largest $\hat{Q}_k^{(i)}$ defines the "best bandit" at time k.



a. ϵ-*Greedy*. At each time step, pick the best bandit with probability $1 - \epsilon$ or another at random with probability ϵ. Three curves are shown here ($\epsilon = 0$,

the greedy strategy, and $\epsilon = 0.01$ and 0.1). Notice that the $\epsilon = 0.01$ curve learns more slowly but will eventually surpass the $\epsilon = 0.1$ curve. Why? To maximize the fraction of times we play the best bandit at $k = 1000$, what is the best value of ϵ?

b. *Probabilistic*. At each time step and for each bandit, draw the random number $p_i \sim \mathcal{N}(\hat{Q}_i, \hat{\sigma}_i)$. Play the bandit with the highest p_i. Why might this be an effective strategy? (The rule differs from the typical Boltzmann rule but works better here.)

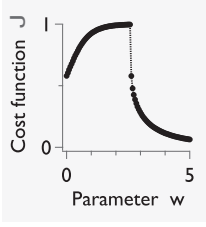
The optimal solution (Gittins) is complicated. Can you improve our heuristic one?

Solution.

See website for code.

a. For the ϵ-Greedy strategy, the best value of ϵ (for maximizing the probability of choosing the truly best bandit at time step $k = 1000$) appears to be roughly $\epsilon = 0.07$.

b. In the probabilistic strategy, if we neglected the standard deviation, we would simply be playing each bandit with a probability proportional to its estimated payout. Drawing from a distribution means that we have a greater chance to explore more poorly known alternatives. We can check that omitting the term (or even changing the proportionality constant in front of the variance), makes the behavior worse.



10.23 Training a recurrent neural network can be hard. Consider a toy RNN, $\dot{x} = -x + \tanh(wx - 1)$, with constant input ($u = -1$) and output $y(t) = x(t)$. Train it by tuning the parameter w, with the goal of minimizing the value of the steady state $x_{ss} \equiv x(t \to \infty)$. Show that the cost function $J(w) = \min_x \left(x_{ss}^2\right)$ has the form at left. Thus, since reasonable cost functions can have discontinuities, training algorithms that perturb coefficients locally can have problems (Doya, 1992).

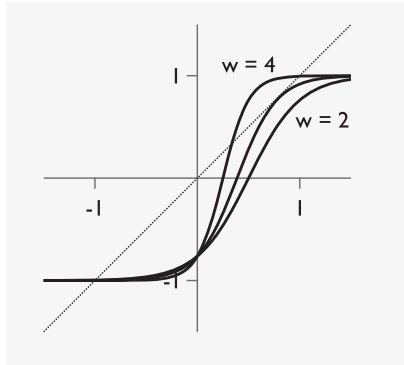
Solution.

The fixed points are given by solving $\dot{x} = 0$, which leads to the algebraic equation,

$$x = \tanh(wx - 1)\,.$$

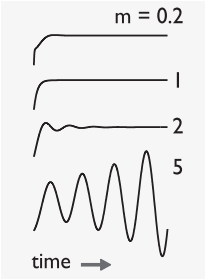
The graph below plots the left- and right-hand sides, for $w = 2$, 2.56, and 4. The fixed points are given by the intersection of the tanh curve with the dotted line $y = x$. We see that there is a saddle-node bifurcation at $w \approx 2.56$, where the system goes from one to three fixed points. Since the goal is to have the smallest steady state, there is a discontinuity in the associated cost function because of this bifurcation. Such behavior is typical when training the internal network connections of a recurrent neural network.

As mentioned in the main text, the reservoir computing scheme skirts this difficulty by using fixed, random connections and training only the output layer, $\boldsymbol{W}^{\text{out}}$, which is a convex problem with a simple, easy-to-find solution.



11 Nonlinear Control

Problems



11.1 Gain scheduling for a linear system. Consider an oscillator with mass m and transfer function $G(s) = (ms^2 + s + 1)^{-1}$.

a. For $m = 1$, find a PID controller $K(s) = K_\mathrm{p} + K_\mathrm{i}/s + K_\mathrm{d}s$ such that the closed-loop transfer function between reference and output $T(s) = \frac{1}{1+s}$.
b. Given $K(s)$ from (a), find the closed-loop transfer function $T_m(s)$ for arbitrary m.
c. Calculate the step response numerically for $m = 0.2$, 1, 2, and 5. (Compare at left.) Note that the response is more robust for smaller than larger masses.
d. Show that the closed-loop response goes unstable at $m = 4$.
e. Given the mass m, design a "gain scheduled" controller $k_\mathrm{m}(s)$ such that $T(s) = \frac{1}{1+s}$. This controller produces a nice step response for all (known) masses.

Solution.

a. We have $T = \frac{KG}{1+KG}$. Solving for K gives

$$\begin{aligned} K(s) &= G(s)^{-1} \frac{T(s)}{1 - T(s)} \\ &= G^{-1} \frac{1}{T^{-1} - 1} \\ &= \left(s^2 + s + 1\right) \frac{1}{(s+1) - 1} \\ &= \frac{s^2 + s + 1}{s}, \end{aligned}$$

which is a PID controller with $K_\mathrm{p} = K_\mathrm{i} = K_\mathrm{d} = 1$.

b. If the mass is now m but the controller is for $m = 1$, the closed-loop transfer function will be $T(s) = \frac{KG}{1+KG}$, with

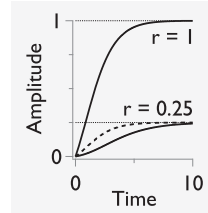
$$G(s) = \frac{1}{ms^2 + s + 1}, \qquad K(s) = \frac{s^2 + s + 1}{s},$$

Simplifying the algebra leads to

$$T(s) = \frac{s^2 + s + 1}{ms^3 + 2s^2 + 2s + 1}.$$

c. The step response is based on $T(s)$ as found in the previous part. Remember that the controller is fixed to be $K(s)$.
d. For $m = 4$, the closed-loop denominator is $4s^3 + 2s^2 + 2s + 1$. We verify that this factors as $(2s + 1)(2s^2 + 1)$, which implies a damped pole at $s = -\frac{1}{2}$, and two poles $s = \pm\, \mathrm{i}\omega$ on the imaginary axis (marginal stability), at $\omega = \frac{1}{\sqrt{2}}$.
e. To make a loop transfer function $L(s) = 1/s$, choose $K(s) = (ms^2 + s + 1)/s$, which corresponds to keeping $K_\mathrm{p} = K_\mathrm{i} = 1$ and choosing $K_\mathrm{d} = m$. As a practical note, noise will make the system will be harder to control at large m, since the derivative term will become large and since that term amplifies the effects of noise.

11.2 Gain scheduling for a nonlinear system. Analyzing a nonlinear system locally can improve control.



a. Integrate Eq. (11.1) numerically to reproduce the step plots shown. The PI gains are $K_\mathrm{p} = K_\mathrm{i} = 0.5$. Hint: differentiate $u(t)$ and integrate the system $[y(t), u(t)]$.
b. Derive the linear system by expanding about the steady-state solution at r.
c. Show that choosing $K_\mathrm{p} = K_\mathrm{i} = 0.5/\sqrt{r}$ makes the transfer function of the resulting linear system between r and y equal to $\frac{1}{(1+s)^2}$.
d. Redo the numerical integration of the nonlinear system for this choice of gains and show that the response time is now independent of r (see dashed line at right).

Solution.

a. The coupled system of equations is

$$\dot{y} = -y + u^2\,, \qquad \dot{u} = -K_\mathrm{p}\dot{y} + K_\mathrm{i}(r - y)\,.$$

These can be integrated numerically by any convenient routine. Note that some programs will require you to solve the equations explicitly for $\dot{y}$ and $\dot{u}$.
b. We linearize using $y = y_0 + y_1(t)$, $u = u_0 + u_1(t)$, and $r = r_0 + r_1(t)$. The steady-state solution for set point r_0 is $y_0 = r_0$ and $u_0 = \sqrt{r_0}$. The deviations obey

$$\begin{pmatrix} 1 & 0 \\ K_\mathrm{p} & 1 \end{pmatrix} \frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} y_1 \\ u_1 \end{pmatrix} = \begin{pmatrix} -1 & 2\sqrt{r_0} \\ -K_\mathrm{i} & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ u_1 \end{pmatrix} + \begin{pmatrix} 0 \\ K_\mathrm{i} \end{pmatrix} r_1(t)\,.$$

Some programs can directly handle such a *descriptor* state-space description. Others will want you to solve explicitly for $\dot{y}_1$ and $\dot{u}_1$. We can do this easily by inverting the left-hand matrix and multiplying through. This gives

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} y_1 \\ u_1 \end{pmatrix} = \underbrace{\begin{pmatrix} -1 & 2\sqrt{r_0} \\ -K_\mathrm{i} + K_\mathrm{p} & -2K_\mathrm{p}\sqrt{r_0} \end{pmatrix}}_{\boldsymbol{A}} \begin{pmatrix} y_1 \\ u_1 \end{pmatrix} + \underbrace{\begin{pmatrix} 0 \\ K_\mathrm{i} \end{pmatrix}}_{\boldsymbol{B}} r_1(t)\,.$$

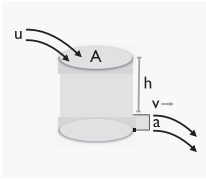
The corresponding transfer function from r_1 to y_1 is found either using control software or by $\boldsymbol{G} = \boldsymbol{C}(s\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{B}$, with $\boldsymbol{C} = \begin{pmatrix} 1 & 0 \end{pmatrix}$:

$$G(s) = \frac{2K_\mathrm{i}\sqrt{r_0}}{2K_\mathrm{i}\sqrt{r_0} + (1 + 2K_\mathrm{p}\sqrt{r_0})s + s^2}\,.$$

Notice that $G(s = 0) = 1$, implying that steady-state perturbations go to $y_1 = r_1$. In addition, we see explicitly that choosing $K_\mathrm{p} = K_\mathrm{i} = \frac{1}{2\sqrt{r_0}}$ makes the transfer function equal to $G = \frac{1}{(1+s)^2}$.

c. Repeating the numerics for the $r = 0.25$ case with the new values of K_p and K_i gives the curve at left. Notice that its time scale is the same. In fact, if you rescale the amplitudes, you will see that the response is essentially identical.

11.3 Level control in a fluid tank.



a. Using Bernouilli's Law, show that fluid exits a tank at a velocity $v^2 = 2gh$, where g is the gravitational acceleration and h is the fluid level above the outlet orifice. Conclude that the level of a tank with cross section A and outlet area a obeys $A\dot{h} = -a\sqrt{2gh} + u(t)$, where $u(t)$ is the added fluid flow volume/time. See left.

b. Design a simple nonlinear controller based on a sensor that measures $h(t)$ and controls $u(t)$. The controller should linearize the dynamics and converge to a desired reference height h_0. Adapted from Slotine and Li (1991).

Solution.

a. Bernouilli's Law neglects any frictional losses and integrates the inviscid Navier-Stokes (Euler) equations for fluids:

$$P + \frac{1}{2}\rho v^2 + \rho g h = \text{const}\,.$$

In a tank with a hole in the side, the pressure at the outlet ($z = 0$) is just the atmospheric pressure, since the exiting fluid is in contact with the atmosphere. Likewise, the pressure at the top of the tank ($z = h$) is also 0, and the tank cross section is big enough that we can consider the fluid as stationary there. Thus, evaluating at $z = h$ and $z = 0$ gives

$$\cancel{P}^{0} + \frac{1}{2}\rho\cancel{v^2}^{0} + \rho g h = \cancel{P}^{0} + \frac{1}{2}\rho v^2 + \cancel{\rho g h}^{0},$$

which implies that the velocity at the outlet orifice obeys $v^2 = 2gh$.

The rate of change of volume in the draining tank is $-A\dot{h}$, which must equal the volume per time of fluid exiting: av. Putting these together and adding the input volume / time gives

$$-A\dot{h} + u(t) = a\sqrt{2gh}\,,$$

which is equivalent to the requested result.

b. To get the desired result, we should "cancel" the nonlinear term and create a PI controller. Thus, we write

$$u(t) = a\sqrt{2gh(t)} + A\left\{K_{\mathrm{p}}[h_0 - h(t)] + K_{\mathrm{i}}\int_0^t \mathrm{d}t'\,[h_0 - h(t')]\right\},$$

which is equivalent to

$$\ddot{h} + K_{\mathrm{p}}\dot{h} + K_{\mathrm{i}}h = K_{\mathrm{i}}h_0\,, \quad h(t=0) = h(0)\,, \quad \dot{h}(0) = K_{\mathrm{p}}[h_0 - h(0)]\,.$$

Choosing K_{p} and K_{i} allows you to choose the pole positions, as in Example 11.4. (But note that the pendulum example uses PD control, and here we use PI control. Think about why.)

11.4 Robustness and error cancellation. Consider the first-order system $\dot{x} = ax^2 + u$.

a. Choose a state-based control $u(t)$ that eliminates the nonlinearity and imposes linear dynamics with a pole at $s = -1$. Assume the state $x(t)$ is observable.
b. Assume now that you mistakenly estimate the parameter a as $\hat{a}$ and choose the feedback accordingly. Analyze the global stability of the closed-loop system.
c. Show that adding $-bx^3$ to the feedback can make $x = 0$ globally stable. Find the smallest value of b that stabilizes the origin.

Solution.

a. We choose $u = -ax^2 - x$, which gives closed-loop dynamics $\dot{x} = -x$, which has a pole at $s = -1$.
b. Now we choose $u = -\hat{a}x^2 - x$, which gives closed-loop dynamics

$$\dot{x} = -x + \epsilon x^2\,,$$

where $\epsilon \equiv \hat{a} - a$. To analyze the global stability, we rewrite the equation in terms of a Lyapunov function. For a one-dimensional state vector $x(t)$, this is always possible.

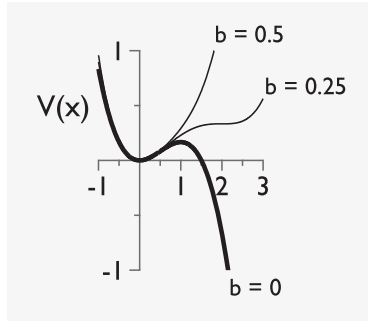
$$\dot{x} = -\frac{\mathrm{d}V}{\mathrm{d}x} = -\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{2}x^2 - \frac{\epsilon}{3}x^3\right).$$

From the plot of $V(x)$ given below for $\epsilon = 1$, it is clear, for positive ϵ, that there is a maximum amplitude $x_{\max}$ for perturbations or initial conditions. The system is stable only if $x(t) < x_{\max}$ for all t. We can find this value by looking at $-\partial_x V(x) = -x + \epsilon x^2 = 0$, which implies fixed points at $x = \{0, \epsilon^{-1}\}$. Thus,

$$x_{\max} = \frac{1}{\epsilon}\,,$$

which implies that the more accurate the estimate of a (the smaller ϵ), the greater the domain of stability.

For negative ϵ (estimated $\hat{a} < a$), the system is stable for $x > x_{\min} = \left(-\epsilon^{-1}\right)$.



c. Next, consider $u = -\hat{a}x^2 - x - bx^3$, which gives closed-loop dynamics

$$\dot{x} = -x + \epsilon x^2 - bx^3 .$$

The Lyapunov function is

$$V(x) = \frac{1}{2}x^2 - \frac{\epsilon}{3}x^3 + \frac{b}{4}x^4 .$$

In order to have absolute stability for the point $x = 0$, we must have $V'(x) < 0$ for $x > 0$ and $V'(x) > 0$ for $x < 0$. The derivative is

$$\frac{\mathrm{d}V}{\mathrm{d}x} = x - \epsilon x^2 + bx^3 = x\left(1 - \epsilon x + bx^2\right) .$$

The condition $V'(x) = 0$ implies that $1 - \epsilon x + bx^2 = 0$. Solving the quadratic equation gives

$$x = \frac{1}{2b}\left(\epsilon \pm \sqrt{\epsilon^2 - 4b}\right) .$$

Absolute stability means that there is no x (except for $x = 0$) that makes $V'(x) = 0$. Thus, we set

$$b > \frac{\epsilon^2}{4} .$$

In other words, we have to estimate the error $\epsilon = \hat{a} - a$ and set b accordingly. Then we impose a feedback

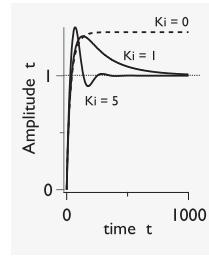
$$u = -\hat{a}x^2 - x - bx^3 .$$

Notice how each term has a function:

- $-\hat{a}x^2$ tries to cancel the nonlinearity;
- $-x$ imposes the desired linear dynamics and time constant for small perturbations;
- $-bx^3$ makes the feedback robust to larger perturbations.

11.5 Nonlinear integral control. In Chapter 3, we saw that integral control can stabilize a set point without offset. The same trick can work for nonlinear systems, too. Consider, as in Problem 11.4, the system $\dot{x} = ax^2 + u$. Assume that $a > 0$ but

that its value is not known. Let the goal now be to stabilize $x(t)$ about the point $x_0 \neq 0$.



a. For proportional control $u = K_p(x_0 - x)$, find the equilibrium point x^* and determine the conditions for linear stability of the fixed point.
b. Add an integral control term, $K_i \int dt'\,(x_0 - x)$. Show that the set point x_0 is now a fixed-point solution. Determine its linear stability as a function of $\{a, x_0, K_p, K_i\}$. Does integral control improve the stability?
c. Reproduce the step responses at right for $K_p = 5$ and $K_i = \{0, 1, 5\}$.
d. Can integral control stabilize an arbitrary set point for a general nonlinear system?

Solution.

a. For proportional control, the fixed points are given by

$$ax^2 - K_p x + K_p x_0 = 0\,.$$

The solutions are

$$x^* = \frac{K_p}{2a}\left[1 \pm \sqrt{1 - \frac{4a}{K_p}x_0}\right].$$

We see that for real fixed points to exist, $K_p \geq 4ax_0$. In the limit $K_p \gg 4ax_0$, we have $x^* \approx K_p/a, x_0$.

The linear stability is given by $x(t) = x^* + x_1(t)$, with $\dot{x}_1 = f'(x^*)\,x_1$. Evaluating the derivative at the fixed point gives $f'(x^*) = 2ax^* - K_p = \pm\sqrt{1 - \frac{4ax_0}{K_p}}$, which is stable for $K_p \geq 4ax_0$.

b. For PI control, we have $u = K_p(x_0 - x) + K_i \int dt'\,(x_0 - x)$, or

$$\dot{x} = ax^2 + K_p(x_0 - x) + K_i w$$
$$\dot{w} = x_0 - x\,.$$

The fixed points are now $x^* = x_0$ and $w^* = -ax_0^2/K_i$. Linearizing about $\{x^*, w^*\}$ gives

$$\boldsymbol{A} = \begin{pmatrix} 2ax_0 - K_p & K_i \\ -1 & 0 \end{pmatrix}.$$

The eigenvalues obey a characteristic equation, $\lambda^2 - 2(ax_0 - K_p/2)\lambda + K_i = 0$, which gives eigenvalues

$$\lambda = (ax_0 - K_p/2) \pm \sqrt{(ax_0 - K_p/2)^2 - K_i}\,,$$

which implies we need $K_p > 2ax_0$ for linear stability and that increasing K_i will make the system more oscillatory, albeit with faster response times.

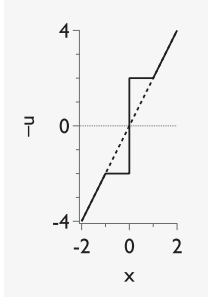
c. You just need to integrate the ODEs. See the book website for Mathematica code.

d. In general, we have $\dot{x} = f(x, u, w)$ and $\dot{w} = x_0 - x$. The second equation will drive $x \to x_0$ and the auxiliary variable $w(t)$ to a stationary value, w^*. Can the equation

$$f(x_0, u, w^*) = 0$$

be solved by picking a control law $u(t)$? If yes, then you also need to check for linear stability about the fixed-point (x_0, w^*). (And if it is linearly stable, you can ask whether there is a finite domain of stability.)

11.6 Stabilization in finite time. Nonlinear control can stabilize a system against perturbations in finite time, whereas linear feedback leads to exponential relaxation that takes an infinite amount of time to decay. As a simple example, consider $\dot{x} = x + u$, a one-dimensional system with unstable fixed point at $x = 0$. Let the feedback law be $u(t) = -k\ \mathrm{sign}[x(t)]$ for $|x| < 1$ and $-k\,x(t)$ for $|x| \geq 1$. See left, where $k = 2$.



a. Show that a perturbation $x(0) = x_0$ returns in finite time to $x = 0$, for $k > 1$.
b. Compare the control effort, $\int_0^\infty \mathrm{d}t\, u^2(t)$, for this system against that for $u = -kx$.
c. Show that you can replace the destabilizing $+x$ term with a nonlinear $f(x)$ satisfying $|f(x)| < \ell|x|$, for some $\ell > 0$ (Lipschitz condition), and $x = 0$ continues to be stable. Hint: show that $V = \frac{1}{2}x^2$ is a Lyapunov function, using an inequality.

See Sun et al. (2017) for hints and applications to the control of complex networks.

Solution.

a. The closed-loop equation of motion is

$$\dot{x} = x + u\,, \qquad u(t) = \begin{cases} -k\ \mathrm{sign}[x(t)] & |x| < 1 \\ -k\,x(t) & |x| \geq 1 \end{cases}.$$

In principle, there are four cases to consider: $x_0 > 0$ and $x_0 < 0$, each subdivided into $|x_0| \leq 1$ and $|x_0| > 1$. If we let $x \to -x$, we see that the equations keep the same form and thus need only consider the magnitude of $x_0 > 0$.
If $x_0 > 1$, then the equation of motion becomes, assuming $k > 1$,

$$\dot{x} = x - kx\,, \qquad \Longrightarrow \qquad x(t) = x_0\ \mathrm{e}^{-(k-1)t}\,.$$

At time $t = t^* = \ln|x_0|\,/(k-1)$, the state reaches one: $|x(t^*)| = 1$, and thereafter, the feedback algorithm changes, and we can look at the $x < 1$ closed-loop equation,

$$\dot{x} = x - k\,, \qquad \Longrightarrow \qquad x(t - t^*) = k - (k-1)\,\mathrm{e}^{t-t^*}\,.$$

Solving for the time for $x(t)$ to reach 0 then gives,

$$t^{**} = \begin{cases} \frac{\ln|x_0|}{k-1} + \ln\left(\frac{k}{k-1}\right) & |x_0| > 1 \\ \ln\left(\frac{k}{k-|x_0|}\right) & |x_0| < 1 \end{cases}$$

The state $x(t^{**}) = 0$ at $t^{**} = \frac{\ln|x_0|}{k-1} + \ln\frac{k}{k-1}$ for $x_0 > 1$. For $0 < x_0 < 1$, the corresponding expression is $t^{**} = \ln(\frac{k}{k-x_0})$ for $x_0 > 1$. Thus, whatever the value of x_0, the system reaches $x(t) = 0$ in finite time, t^{**}. Collecting these
For reference, for $k = 2$, the expressions simplify to

$$t^{**} = \begin{cases} \ln|x_0| + \ln 2 & |x_0| > 1 \\ \ln\left(\frac{2}{2-|x_0|}\right) & |x_0| < 1 \end{cases}$$

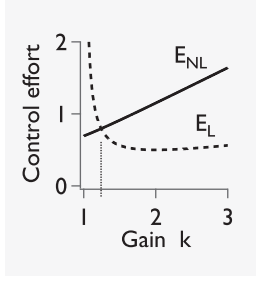
b. Let us first consider the case $0 < x_0 < 1$. Then, from the previous section, the control signal is $u = -k$ for a time $t^{**} = \ln k/(k - x_0)$. The corresponding control effort is

$$E_{\mathrm{NL}} = k^2 \ln\left(\frac{k}{k - x_0}\right).$$

We compare this to the linear case, where $x(t) = x_0\,\mathrm{e}^{-(k-1)t}$ for $0 < t < \infty$. The control effort is then

$$E_{\mathrm{L}} = k^2 x_0^2 \int_0^\infty \mathrm{d}t\,\mathrm{e}^{-2(k-1)t} = \frac{k^2 x_0^2}{2(k-1)}.$$

Below, we plot E_{NL} and E_{L} vs. $k > 1$ for $x_0 = \frac{1}{2}$. For gains $k \gtrsim k_{\mathrm{c}} = 1$, the nonlinear control is cheaper; for $k > k^* \approx 1.25$, the linear control is cheaper (but slower). In general, as we have seen often previously, there is a tradeoff between the speed of response and control effort.



c. For the original equations, let us verify that $V = \frac{1}{2}x^2$ is a Lyapunov function. For $x_0 > 1$, we start with the upper branch:

$$\dot{V} = x\dot{x} = x^2(1 - k) = -(k - 1)x^2 < 0\,, \quad \text{for } x > 1\,.$$

Then, the lower branch, with $0 < x_0 < 1$:

$$\dot{V} = x\dot{x} = x(x - k) < 0\,, \quad \text{for } 0 < x < 1\,.$$

For the inequality, we note that $x - k < 0$ for $0 < x < 1$, since we are also given that $k > 1$.
Now we modify the closed-loop equation of motion to

$$\dot{x} = f(x) + u(x)\,, \quad \text{with } \|f(x)\| < \ell\|x\|\,,$$

for some positive bound ℓ. Then

$$\dot{V} = x\dot{x} = x\,[f(x) + u(x)] \le \ell x^2 + x\,u(x)\,.$$

So, in the upper branch, $\dot{V} \le (\ell - k)x^2 \le 0$ for $k > \ell$. And, for $0 < x < 1$, we have $\dot{V} \le \ell x^2 - kx \le 0$ when $0 < x < 1$. The point is that $f(x)$ is less destabilizing than ℓx and can be stabilized by linear feedback $k > \ell$ for $x > 1$.

11.7 Relative degree. Consider a linear system with transfer function $G(s)$,

$$G(s) = \frac{b_0 s^k + b_1 s^{k-1} + \cdots + b_k}{s^n + a_1 s^{n-1} + \cdots + a_n} \equiv \frac{b(s)}{a(s)}\,.$$

Show that the input $u(t)$ first appears after $n - k$ differentiations of the output $y(t)$. The relative degree is thus the difference between the numerator and denominator orders.

Solution.

Let us write a state-space equivalent of the linear system. From Eq. (2.59), we have,

$$\dot{\boldsymbol{x}} = \underbrace{\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_2 & -a_1 \end{pmatrix}}_{\boldsymbol{A}} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} + \underbrace{\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}}_{\boldsymbol{B}} u$$

$$y = \underbrace{\begin{pmatrix} b_k & b_{k-1} & \cdots & b_1 & b_0 & 0 & \cdots & 0 \end{pmatrix}}_{\boldsymbol{C}} \boldsymbol{x}\,.$$

To check the relative degree, we need to take derivatives of the output $y(t)$. We have,

$$\dot{y} = \boldsymbol{CA}\boldsymbol{x} + \boldsymbol{CB}u\,.$$

If $k = n - 1$, then the row vector $\boldsymbol{C}$ has no zeros and $\boldsymbol{CB} = b_{n-1} \neq 0$. We would conclude the system has relative degree $n - 1$.

If $\boldsymbol{CB} = 0$, then we differentiate y again:

$$\begin{aligned} \ddot{y} &= \boldsymbol{CA}\dot{\boldsymbol{x}} + \cancelto{0}{\boldsymbol{CB}\dot{u}} \\ &= \boldsymbol{CA}^2\boldsymbol{x} + \boldsymbol{CAB}u\,. \end{aligned}$$

Now we look at $\boldsymbol{CAB}$. Because of the structure of $\boldsymbol{A}$, with ones above the diagonal, it just shifts the elements of $\boldsymbol{C}$ one element to the right. That is,

$$\boldsymbol{CA} = \begin{pmatrix} 0 & b_k & b_{k-1} & \cdots & b_1 & b_0 & 0 & \cdots & 0\,. \end{pmatrix}$$

Before, there were $n-k-1$ zeros on the right. Now there are $n-k-2$. Thus, if the relative degree is $n-2$, then $\boldsymbol{CAB} = b_0$. Otherwise, we repeat until b_0 finally appears. This will be after $n-k$ differentiations, which proves the claim.

11.8 Internal dynamics. More on the simple example from Section 11.1.3.

a. Show that choosing $u = -x_2^3 - x_1$ stabilizes the $\binom{0}{0}$ solution to $\frac{\mathrm{d}}{\mathrm{d}t}\binom{x_1}{x_2} = \binom{x_2^3+u}{u}$.
b. Now alter $\dot{x}_2 = u$ to $\dot{x}_2 = -u$. Show that the internal dynamics is unstable.

Solution.

a. The control law $u = -x_2^3 - x_1$ leads to $\dot{x}_1 = -x_1$, whose solution, $x_1(t) = x_1(0)\,\mathrm{e}^{-t}$, converges to zero for all initial conditions $x_1(0)$.

The second dynamical equation,

$$\dot{x}_2 = u = -x_2^3 - x_1\,,$$

is a nonlinear differential equation. If the $x_1(t)$ term were absent, then $x_2 = 0$ would be a globally stable fixed point. Indeed, for $x_2(0) \equiv x_0$,

$$x_2(t) = \frac{1}{\sqrt{2t + \frac{1}{x_0^2}}}\,.$$

The $x_1(t) = x_1(0)\,\mathrm{e}^{-t}$ term appears as an external driving that causes $x_2(t)$ to change and makes the ODE for x_2 non-autonomous. We begin by considering short- and long-time asymptotic limits:

- At short times, $t \ll 1$: then $\dot{x}_2 \approx +x_1(0)$, so that $x_2(t) \sim x_1(0)\,t$.
- At long times, $t \gg 1$: the $x_1(t)$ term decreases exponentially, whereas the unperturbed $x_2(t)$ $t^{-1/2}$, a much slower decay. Thus, after sufficient time, $\dot{x}_2 \approx -x_2^3$, independent of the initial condition, and $x_2(t) \sim \frac{1}{\sqrt{2t}}$. The overall sign will depend on the initial condition. Thus, although the $x_1(t)$ term can be initially destabilizing, we still expect asymptotic stability.

Another approach is to bound $x_1(t)$ by $x_1(0)$. Let us take the case, $x_2(0) > 0$ and $x_1(0) = -\alpha^3$, with $\alpha > 0$. Then

$$\dot{x}_2 = -x_2^3 + \alpha^3$$

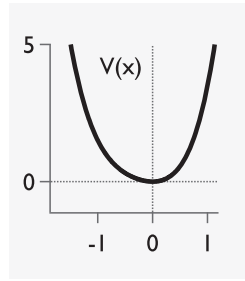
has a fixed point $x_2 = \alpha$. With $x_2 \equiv x + \alpha$, we have

$$\dot{x} = -x^3 - 3\alpha x^2 - 3\alpha^2 x\,.$$

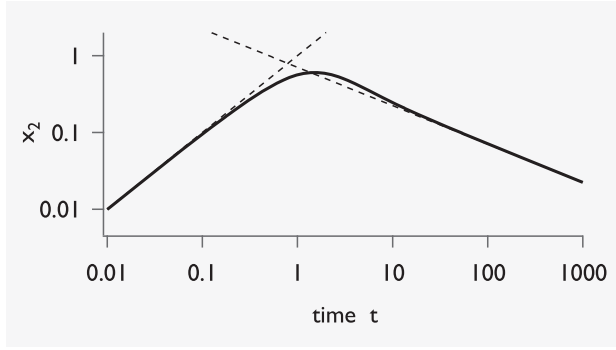
The fixed point is determined from $x(x^2 + 3\alpha x + 3\alpha^2) = 0$, which has solutions

$$x = 0\,,\, \frac{\alpha}{2}\left(-3 \pm \mathrm{i}\,\sqrt{3}\right)\,.$$

The only real root is $x = 0$. If we interpret the right-hand side as $-\frac{\mathrm{d}V}{\mathrm{d}x}$, with $V(x)$ a Lyapunov function, we can easily see that $x = 0$ will be a globally stable solution. Below, we show $V(x)$ for $\alpha = 1$.



We then plot numerically, below the full solution for $x_2(t)$ assuming $x_2(0) = 0$ and $x_1(0) = 1$. In the plot, the solid line is the full $x_2(t)$ solution, while the dashed lines give long-time ($x_2 \sim \frac{1}{\sqrt{2t}}$) and short-time ($x_2 \sim t$) approximations.



b. If we consider the equations,

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_2^3 + u \\ -u \end{pmatrix}, \quad y = x_1\,,$$

then

$$\dot{x}_2 = -u = x_2^3 + x_1(t)\,,$$

which clearly blows up. The reference solution to $\dot{x}_2 = x_2^3$ is

$$x_2(t) = \frac{1}{\sqrt{-2t + \frac{1}{x_0^2}}}\,,$$

which blows up at time $t = 1/(2y_0^2)$.

Thus, the internal dynamics are stable in Part (a) and unstable in Part (b).

11.9 Feedback linearization of a two-dimensional system. Complete Example 11.5 by deriving the equations of motion in the new coordinate system $\boldsymbol{z} = \boldsymbol{T}(\boldsymbol{x})$. If the output is given as $y = x_1$, what is the relative degree?

Solution.

Recall the equations of motion,

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} f_1(\boldsymbol{x}) \\ f_2(\boldsymbol{x}) \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u\,.$$

Using the input-state linearization technique of Section 11.1.4, we found that

$$T_1 = x_1\,, \qquad T_2 = f_1(\boldsymbol{x})\,, \qquad u = -f_2 + \left(\frac{\partial f_1}{\partial x_2}\right)^{-1}\left[v - f_1\frac{\partial f_1}{\partial x_1}\right].$$

Thus, $\dot{z}_1 = \dot{x_1} = f_1 = z_2$. Next,

$$\begin{aligned} \dot{z}_2 = \dot{f}_1 &= \frac{\partial f_1}{\partial x_1}\dot{x}_1 + \frac{\partial f_1}{\partial x_2}\dot{x}_2 \\ &= \frac{\partial f_1}{\partial x_1}f_1 + \frac{\partial f_1}{\partial x_2}(f_2 + u) = v\,. \end{aligned}$$

Thus, collecting these equations, we have

$$\begin{aligned} \dot{z}_1 &= z_2 \\ \dot{z}_2 &= v\,. \end{aligned}$$

We can then pick v as desired. For example, we can choose a feedback law to stabilize the fixed point $z_1 = z_2 = 0$.

The relative degree can be found by differentiating the output until the input appears. That is just the calculation done above, which shows that the relative degree = 2.

11.10 Linearizable, but not reachable. Consider $\dot{x_1} = x_2^2$ and $\dot{x}_2 = u(t)$.

a. Using the results from Example 11.5, find a nonlinear change of coordinates that makes this system linear and controllable.
b. Show that, nonetheless, you cannot reach all states starting from the origin, $\binom{0}{0}$. Qualitatively, what goes wrong?
c. Find the set of states that can be reached after a time T.

Solution.

a. To put the dynamics in the form of Example 11.5, we write

$$\boldsymbol{f} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = \begin{pmatrix} x_2^2 \\ 0 \end{pmatrix}, \qquad \boldsymbol{g} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

We then choose $z_1 = T_1 = x_1$ and $z_2 = T_2 = f_1 = x_2^2$. Then

$$\beta^{-1} = \frac{\partial T_2}{\partial x_2} = 2x_2\,, \qquad \alpha = 0\,,$$

and we write $\dot{z}_2 = 2x_2\,u = v$, or $u = v/(2x_2)$. The linear system is then

$$\underbrace{\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}}_{\dot{z}} = \underbrace{\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}}_{A}\underbrace{\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}}_{z} + \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{B} v\,.$$

Noting that

$$\boldsymbol{AB} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \text{and that} \quad \boldsymbol{W}_{\mathrm{c}} = (\boldsymbol{B},\, \boldsymbol{AB}) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

has full rank, we conclude that the new linear system is controllable.

b. We can clearly choose a $u(t)$ to make $x_2(t = T)$ an arbitrary value. But since $\dot{x}_1 = x_2^2 \geq 0$, we can never decrease $x_1(t)$ below its original value. Qualitatively, even though our control can access a set that has the dimension of the full state space (2 here), that is not sufficient to say that it can access the whole of that state space. Something akin to ordinary functional mappings is occurring here. Essentially the mapping of controls to the tangent space is not onto, in the same way that $y = x^2$ maps the whole real line to the positive line $y \geq 0$.

c. The previous section shows that half the state space is inaccessible. Here, we show that we can reach any point in the rest of state space. Let the desired endpoint be $\binom{a}{b}$, to be reached at time T Let us assume that a and b are greater than zero. The case $a < 0$ is handled similarly.
Consider a control

$$u(t) = \begin{cases} +1 & 0 < t < \tau \\ -1 & \tau < t < T \end{cases}, \quad \Longrightarrow \quad x_2(t) = \begin{cases} t & 0 < t < \tau \\ 2\tau - t & \tau < t < T \end{cases}.$$

Note that $x_2(T) = 2\tau - T = b$. Then, using $x_1(t) = \int_0^t \mathrm{d}t'\, x_2(t')^2$, we have,

$$x_1(T) = \frac{\tau^3}{3} + \frac{1}{3}\left[\tau^3 - (2\tau - T)^3\right] = \frac{2\tau^3}{3} - \frac{b^3}{3} = a\,.$$

This shows that we can choose τ and T to match a and b, iff $b > 0$.

11.11 Involutive or not? Consider the vector fields, $\boldsymbol{f} = \begin{pmatrix} 1 \\ 0 \\ x_2 \end{pmatrix}$, $\boldsymbol{g} = \begin{pmatrix} 0 \\ -1 \\ \pm x_1 \end{pmatrix}$. Show that choosing $+x_1$ gives an involutive pair but choosing $-x_1$ does not.

Solution.

The Lie bracket $[\boldsymbol{f}, \boldsymbol{g}]$ is given by

$$[\boldsymbol{f}, \boldsymbol{g}] = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{f} - \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\boldsymbol{g} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \pm 1 & 0 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \\ x_2 \end{pmatrix} - \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} 0 \\ -1 \\ \pm x_1 \end{pmatrix}$$

$$= \begin{pmatrix} 0 \\ 0 \\ \pm 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix}_{+} \text{ or } \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}_{-}.$$

The $+x_1$ case thus leads to three vector fields that we can put together in a matrix $\boldsymbol{W}_+ = \{\boldsymbol{f}, \boldsymbol{g}_+, [\boldsymbol{f}, \boldsymbol{g}_+]\}$, which we write explicitly as,

$$\boldsymbol{W}_+ = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ x_2 & x_1 & 2 \end{pmatrix}.$$

Since $\det \boldsymbol{W}_+ = -2 \neq 0$, for all $\boldsymbol{x} \in M = \mathbb{R}^3$, the three vectors span the full three-dimensional tangent space for all $\boldsymbol{x} \in M$.

The $-x_1$ case leads to

$$\boldsymbol{W}_- = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ x_2 & -x_1 & 0 \end{pmatrix}.$$

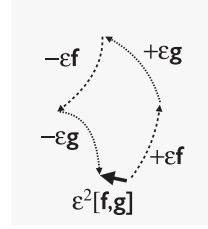
Now, $\det \boldsymbol{W}_- = 0$, which shows that $\boldsymbol{f}$ and $\boldsymbol{g}$ are not involutive.

This exercise is due to Roger Brockett, Harvard University.

11.12 Lie brackets. Let $[\boldsymbol{f}, \boldsymbol{g}] \equiv \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}} \boldsymbol{f} - \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \boldsymbol{g}$, with $\boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{g}(\boldsymbol{x})$ vector fields and $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$ and $\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}$ Jacobian matrices. Let $h(\boldsymbol{x})$ be a real-valued function, and recall that the Lie derivative $L_f h \equiv \frac{\partial h}{\partial \boldsymbol{x}} \boldsymbol{f}$. Then show the following:

a. *Geometrical interpretation*: Consider the following sequence of vector-field flows:

$$\dot{\boldsymbol{x}} = \underbrace{+\boldsymbol{f}(\boldsymbol{x})}_{0<t<\epsilon}, \quad \underbrace{+\boldsymbol{g}(\boldsymbol{x})}_{\epsilon<t<2\epsilon}, \quad \underbrace{-\boldsymbol{f}(\boldsymbol{x})}_{2\epsilon<t<3\epsilon}, \quad \underbrace{-\boldsymbol{g}(\boldsymbol{x})}_{3\epsilon<t<4\epsilon}.$$



By Taylor-expanding the state $\boldsymbol{x}(t)$, show that the state fails to return to the origin by an amount $\boldsymbol{x}(4\epsilon) - \boldsymbol{x}(0) = \epsilon^2[\boldsymbol{f}, \boldsymbol{g}] + O(\epsilon^3)$. (Caution: messy algebra.)

b. *Bilinearity*: $[\alpha_1 \boldsymbol{f}_1 + \alpha_2 \boldsymbol{f}_2\,,\, \boldsymbol{g}] = \alpha_1\,[\boldsymbol{f}_1\,,\, \boldsymbol{g}] + \alpha_2\,[\boldsymbol{f}_2\,,\, \boldsymbol{g}]$.

c. *Skew commutivity*: $[\boldsymbol{f}\,,\, \boldsymbol{g}] = -[\boldsymbol{g}\,,\, \boldsymbol{f}]$.

d. *Jacobi identity*: $L_{[\boldsymbol{f},\boldsymbol{g}]}\, h(\boldsymbol{x}) = L_f L_g\, h(\boldsymbol{x}) - L_g L_f\, h(\boldsymbol{x})$.

Solution.

a. For simplicity, let $\boldsymbol{x}(t = 0) = \boldsymbol{0}$. Then, an $O(\epsilon^2)$ Taylor expansion gives

$$\boldsymbol{x}(\epsilon) = \epsilon \dot{\boldsymbol{x}}(0) + \frac{\epsilon^2}{2} \ddot{\boldsymbol{x}}(0) = \epsilon \boldsymbol{f}(\boldsymbol{0}) + \frac{\epsilon^2}{2} \left.\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \boldsymbol{f}\right|_{\boldsymbol{0}}.$$

Then we evolve this state from ϵ to 2ϵ with $\dot{\boldsymbol{x}} = +\boldsymbol{g}(\boldsymbol{x})$:

$$\begin{aligned} \boldsymbol{x}(2\epsilon) &= \boldsymbol{x}(\epsilon) + \epsilon \dot{\boldsymbol{x}}(\epsilon) + \frac{\epsilon^2}{2} \ddot{\boldsymbol{x}}(\epsilon) \\ &= \boldsymbol{x}(\epsilon) + \epsilon \boldsymbol{g}[\boldsymbol{x}(\epsilon)] + \frac{\epsilon^2}{2} \left.\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}} \boldsymbol{g}\right|_{\boldsymbol{x}(\epsilon)}. \end{aligned}$$

We note that, to the lowest required order in ϵ,

$$\boldsymbol{g}[\boldsymbol{x}(\epsilon)] = \boldsymbol{g}(\boldsymbol{0}) + \epsilon \left.\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}} \boldsymbol{f}\right|_{\boldsymbol{0}} \quad \text{and} \quad \left.\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}} \boldsymbol{g}\right|_{\boldsymbol{x}(\epsilon)} = \left.\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}} \boldsymbol{g}\right|_{\boldsymbol{0}}.$$

Then, substituting into the expression for $\boldsymbol{x}(2\epsilon)$ gives, to $O(\epsilon^2)$,

$$\begin{aligned} \boldsymbol{x}(2\epsilon) &= \epsilon \boldsymbol{f}(\boldsymbol{0}) + \frac{\epsilon^2}{2} \left.\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \boldsymbol{f}\right|_{\boldsymbol{0}} + \epsilon \left(\boldsymbol{g}(\boldsymbol{0}) + \epsilon \left.\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}} \boldsymbol{f}\right|_{\boldsymbol{0}} \right) + \frac{\epsilon^2}{2} \left.\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}} \boldsymbol{g}\right|_{\boldsymbol{0}} \\ &= \epsilon\,[\boldsymbol{f}(\boldsymbol{0}) + \boldsymbol{g}(\boldsymbol{0})] + \frac{\epsilon^2}{2} \left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \boldsymbol{f} + \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}} \boldsymbol{g} + 2 \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}} \boldsymbol{f} \right)_{\boldsymbol{0}}. \end{aligned}$$

Next, we evolve from time 2ϵ to 3ϵ using $\dot{\boldsymbol{x}} = -\boldsymbol{f}(\boldsymbol{x})$, which gives,

$$\begin{aligned}\boldsymbol{x}(3\epsilon) &= \boldsymbol{x}(2\epsilon) + \epsilon\dot{\boldsymbol{x}}(2\epsilon) + \frac{\epsilon^2}{2}\ddot{\boldsymbol{x}}(2\epsilon) \\ &= \boldsymbol{x}(2\epsilon) - \epsilon\boldsymbol{f}[\boldsymbol{x}(2\epsilon)] + \frac{\epsilon^2}{2}\left.\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\boldsymbol{f}\right|_{\boldsymbol{x}(2\epsilon)} .\end{aligned}$$

We have

$$\boldsymbol{f}[\boldsymbol{x}(2\epsilon)] = \boldsymbol{f}(\boldsymbol{0}) + \epsilon\left[\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\boldsymbol{f} + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\boldsymbol{g}\right]_{\boldsymbol{0}}, \quad \text{and} \quad \left.\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\boldsymbol{f}\right|_{\boldsymbol{x}(2\epsilon)} = \left.\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\boldsymbol{f}\right|_{\boldsymbol{0}} .$$

Substituting into the expansion then gives,

$$\begin{aligned}\boldsymbol{x}(3\epsilon) &= \underbrace{\epsilon\left[\cancel{\boldsymbol{f}(\boldsymbol{0})} + \boldsymbol{g}(\boldsymbol{0})\right] + \frac{\epsilon^2}{2}\left(\cancel{\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\boldsymbol{f}} + \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{g} + 2\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{f}\right)_{\boldsymbol{0}}}_{\boldsymbol{x}(2\epsilon)} \\ &\quad - \underbrace{\epsilon\left[\cancel{\boldsymbol{f}(\boldsymbol{0})} + \epsilon\left[\cancel{\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\boldsymbol{f}} + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\boldsymbol{g}\right]_{\boldsymbol{0}}\right]}_{\boldsymbol{f}[\boldsymbol{x}(2\epsilon)]} + \frac{\epsilon^2}{2}\left.\cancel{\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\boldsymbol{f}}\right|_{\boldsymbol{0}} \\ &= \epsilon\boldsymbol{g}(\boldsymbol{0}) + \frac{\epsilon^2}{2}\left(\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{g} + 2\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{f} - 2\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\boldsymbol{g}\right)_{\boldsymbol{0}} \\ &= \epsilon\boldsymbol{g}(\boldsymbol{0}) + \frac{\epsilon^2}{2}\left.\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{g}\right|_{\boldsymbol{0}} + \epsilon^2\left(\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{f} - \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\boldsymbol{g}\right)_{\boldsymbol{0}} \\ &= \epsilon\boldsymbol{g}(\boldsymbol{0}) + \frac{\epsilon^2}{2}\left.\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{g}\right|_{\boldsymbol{0}} + \epsilon^2[\boldsymbol{f}, \boldsymbol{g}]_{\boldsymbol{0}} .\end{aligned}$$

Finally, we evolve from 3ϵ to 4ϵ with $\dot{\boldsymbol{x}} = -\boldsymbol{g}(\boldsymbol{x})$:

$$\begin{aligned}\boldsymbol{x}(4\epsilon) &= \boldsymbol{x}(3\epsilon) + \epsilon\dot{\boldsymbol{x}}(3\epsilon) + \frac{\epsilon^2}{2}\ddot{\boldsymbol{x}}(3\epsilon) \\ &= \boldsymbol{x}(3\epsilon) - \epsilon\boldsymbol{g}[\boldsymbol{x}(3\epsilon)] + \frac{\epsilon^2}{2}\left.\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{g}\right|_{\boldsymbol{x}(3\epsilon)} .\end{aligned}$$

We have

$$\boldsymbol{g}[\boldsymbol{x}(3\epsilon)] = \boldsymbol{g}(\boldsymbol{0}) + \epsilon\left.\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{g}\right|_{\boldsymbol{0}}, \quad \text{and} \quad \left.\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{g}\right|_{\boldsymbol{x}(3\epsilon)} = \left.\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{g}\right|_{\boldsymbol{0}} ,$$

so that

$$\begin{aligned}\boldsymbol{x}(4\epsilon) &= \underbrace{\cancel{\epsilon\boldsymbol{g}(\boldsymbol{0})} + \frac{\epsilon^2}{2}\left.\cancel{\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{g}}\right|_{\boldsymbol{0}} + \epsilon^2[\boldsymbol{f}, \boldsymbol{g}]_{\boldsymbol{0}}}_{\boldsymbol{x}(3\epsilon)} \underbrace{- \epsilon\left(\cancel{\boldsymbol{g}(\boldsymbol{0})} + \epsilon\left.\cancel{\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{g}}\right|_{\boldsymbol{0}}\right)}_{\boldsymbol{g}[\boldsymbol{x}(3\epsilon)]} + \frac{\epsilon^2}{2}\left.\cancel{\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{g}}\right|_{\boldsymbol{0}} \\ &= \epsilon^2[\boldsymbol{f}, \boldsymbol{g}]_{\boldsymbol{0}} .\end{aligned}$$

b. Bilinearity follows from the definition and from the linearity of the derivative:

$$[\alpha_1 \boldsymbol{f}_1 + \alpha_2 \boldsymbol{f}_2\,,\ \boldsymbol{g}] = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\,(\alpha_1 \boldsymbol{f}_1 + \alpha_2 \boldsymbol{f}_2) - \frac{\partial}{\partial \boldsymbol{x}}(\alpha_1 \boldsymbol{f}_1 + \alpha_2 \boldsymbol{f}_2)\,\boldsymbol{g} = \alpha_1\,[\boldsymbol{f}_1\,,\ \boldsymbol{g}] + \alpha_2\,[\boldsymbol{f}_2\,,\ \boldsymbol{g}]\,.$$

c. Skewness is a trivial property of the definition:

$$[\boldsymbol{g}, \boldsymbol{f}] = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\boldsymbol{g} - \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{f} = -\left(\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{f} - \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\boldsymbol{g}\right) = -[\boldsymbol{f}, \boldsymbol{g}]\,.$$

d. To establish the Jacobi identity, we work back from the right-hand side to the left. With $L_{\boldsymbol{g}}\, h = \frac{\partial h}{\partial \boldsymbol{x}}\boldsymbol{g}$, we have

$$\begin{aligned}
L_{\boldsymbol{f}} L_{\boldsymbol{g}}\, h(\boldsymbol{x}) - L_{\boldsymbol{g}} L_{\boldsymbol{f}}\, h(\boldsymbol{x}) &= \frac{\partial}{\partial \boldsymbol{x}}\left(\frac{\partial h}{\partial \boldsymbol{x}}\boldsymbol{g}\right)\boldsymbol{f} - \frac{\partial}{\partial \boldsymbol{x}}\left(\frac{\partial h}{\partial \boldsymbol{x}}\boldsymbol{f}\right)\boldsymbol{g} \\
&= \cancel{\frac{\partial^2 h}{\partial \boldsymbol{x}^2}\boldsymbol{g}\,\boldsymbol{f}} + \frac{\partial h}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{f} - \cancel{\frac{\partial^2 h}{\partial \boldsymbol{x}^2}\boldsymbol{f}\,\boldsymbol{g}} - \frac{\partial h}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\boldsymbol{g} \\
&= \frac{\partial h}{\partial \boldsymbol{x}}\left(\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{f} - \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\boldsymbol{g}\right) \\
&= \frac{\partial h}{\partial \boldsymbol{x}}[\boldsymbol{f}\,,\ \boldsymbol{g}] \\
&= L_{[\boldsymbol{f},\boldsymbol{g}]}\, h(\boldsymbol{x})\,.
\end{aligned}$$

11.13 Parking a unicycle. The equations of motion are $\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \\ \theta \end{pmatrix} u_1(t)\begin{pmatrix} \cos\theta \\ \sin\theta \\ 0 \end{pmatrix} + u_2(t)\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ for the unicycle, where u_1 controls the *drive* and u_2 the *spin*. Show that the Lie bracket, [*drive*, *spin*] = *slip*, where *slip* is $\perp$ to *drive*.

Solution.

To calculate the Lie bracket [*drive*, *spin*], we simplify notation by defining

$$\mathit{drive} \equiv \boldsymbol{f} = \begin{pmatrix} \cos\theta \\ \sin\theta \\ 0 \end{pmatrix}, \qquad \mathit{spin} \equiv \boldsymbol{g} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Then $[\boldsymbol{f}, \boldsymbol{g}]$ is given by

$$\begin{aligned}
[\boldsymbol{f}, \boldsymbol{g}](\boldsymbol{x}) &= \cancel{\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}}^{\,0} f - \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} g \\
&= -\begin{pmatrix} 0 & 0 & -\sin\theta \\ 0 & 0 & \cos\theta \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \\
&= \begin{pmatrix} \sin\theta \\ -\cos\theta \\ 0 \end{pmatrix}.
\end{aligned}$$

The result is a new vector that we call *slip*. It clearly corresponds to a displacement perpendicular to the unicycle's drive axis. Note that $drive \cdot slip = 0$.

11.14 Feedback linearization. Let us show that Eq. (11.38) gives necessary and sufficient conditions for a control affine system to be feedback linearizable.

a. Use the Jacobi identity (Problem 11.12d) to show, for a smooth function $h(\boldsymbol{x})$ and smooth vector fields $\boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{g}(\boldsymbol{x})$ and arbitrary positive integer k, that

$$L_{\boldsymbol{g}}h = L_{\boldsymbol{g}}L_{\boldsymbol{f}}h = \cdots = L_{\boldsymbol{g}}L_{\boldsymbol{f}}^{k}h = 0 \quad \Longleftrightarrow \quad L_{\boldsymbol{g}}h = L_{\mathrm{ad}_{\boldsymbol{f}}\boldsymbol{g}}h = \cdots = L_{\mathrm{ad}_{\boldsymbol{f}}^{k}\boldsymbol{g}}h = 0\,.$$

b. Using the above result, show that Eqs. (11.28) and (11.29) imply,

$$\frac{\partial T_1}{\partial \boldsymbol{x}}\,\mathrm{ad}_{\boldsymbol{f}}^{k}\boldsymbol{g} = 0 \text{ for } k = 0, 1, 2, \ldots, n-2 \quad \text{and} \quad \frac{\partial T_1}{\partial \boldsymbol{x}}\,\mathrm{ad}_{\boldsymbol{f}}^{n-1}\boldsymbol{g} \neq 0\,.$$

c. Conclude that the vector fields $\boldsymbol{g}$, $\mathrm{ad}_{\boldsymbol{f}}\boldsymbol{g}$, …, $\mathrm{ad}_{\boldsymbol{f}}^{n-1}\boldsymbol{g}$ are linearly independent.

d. Using the Frobenius theorem, conclude that the vector fields are involutive.

Solution.

a. For $k = 0$, the statement is trivial, as both sides of the relation state that $L_{\boldsymbol{g}}h = 0$. For $k = 1$, we use the Jacobi relation to write,

$$L_{\mathrm{ad}_{\boldsymbol{f}}\boldsymbol{g}}h = L_{[\boldsymbol{f},\boldsymbol{g}]}\,h = L_{\boldsymbol{f}}\cancelto{0}{L_{\boldsymbol{g}}h} - \cancelto{0}{L_{\boldsymbol{g}}L_{\boldsymbol{f}}h} = 0\,.$$

For $k = 2$, we use the Jacobi relation twice to write

$$\begin{aligned} L_{\mathrm{ad}_{\boldsymbol{f}}^{2}\boldsymbol{g}}h &= L_{[\boldsymbol{f},[\boldsymbol{f},\boldsymbol{g}]]}\,h = L_{\boldsymbol{f}}\cancelto{0}{L_{[\boldsymbol{f},\boldsymbol{g}]}h} - L_{[\boldsymbol{f},\boldsymbol{g}]}L_{\boldsymbol{f}}h \\ &= -\left(L_{\boldsymbol{f}}L_{\boldsymbol{g}} - L_{\boldsymbol{g}}L_{\boldsymbol{f}}\right)L_{\boldsymbol{f}}h = -L_{\boldsymbol{f}}\cancelto{0}{\left(L_{\boldsymbol{g}}L_{\boldsymbol{f}}h\right)} + L_{\boldsymbol{g}}\cancelto{0}{\left(L_{\boldsymbol{f}}^{2}h\right)} = 0\,. \end{aligned}$$

We continue to establish the relation for all k. We can also reverse the argument to establish equivalence between the two sets of relationships.

b. The second relation in Eq. (11.29) states, for $k = 0, 1, \ldots, n-2$, that

$$L_{\boldsymbol{g}}L_{\boldsymbol{f}}^{k}T_1 = 0\,.$$

Part (a), with $h \to T_1$, then implies that

$$L_{\boldsymbol{g}}T_1 = L_{\mathrm{ad}_{\boldsymbol{f}}\boldsymbol{g}}T_1 = \cdots = L_{\mathrm{ad}_{\boldsymbol{f}}^{k}\boldsymbol{g}}T_1 = 0\,.$$

Using the definition of the Lie derivative then gives

$$\frac{\partial T_1}{\partial \boldsymbol{x}}\,\boldsymbol{g} = \frac{\partial T_1}{\partial \boldsymbol{x}}\,\mathrm{ad}_{\boldsymbol{f}}\boldsymbol{g} = \cdots = \frac{\partial T_1}{\partial \boldsymbol{x}}\,\mathrm{ad}_{\boldsymbol{f}}^{k}\boldsymbol{g} = 0\,.$$

For $k = n-1$, we start from Eq. (11.29), which states that $L_{\boldsymbol{g}}T_n \neq 0$. Using the recursive equations $L_{\boldsymbol{f}}^{k-1}T_1 = T_k$, we have

$$L_{\boldsymbol{g}}T_n = L_{\boldsymbol{g}}\left(L_{\boldsymbol{f}}T_{n-1}\right) = L_{\boldsymbol{g}}\left(L_{\boldsymbol{f}}^{2}T_{n-2}\right) = \cdots = L_{\boldsymbol{g}}\left(L_{\boldsymbol{f}}^{n-1}T_1\right) \neq 0\,.$$

The result of Part (a) then implies that $L_{\mathrm{ad}_f^{n-1}\boldsymbol{g}} T_1 \neq 0$ or, equivalently,

$$\frac{\partial T_1}{\partial \boldsymbol{x}}\,\mathrm{ad}_f^k \boldsymbol{g} \neq 0\,.$$

c. If $\boldsymbol{g}$, $\mathrm{ad}_f \boldsymbol{g}$, …, $\mathrm{ad}_f^{n-1}\boldsymbol{g}$ are linearly dependent, then there exist $\ell - 1$ functions $\alpha_1(\boldsymbol{x})$, …, $\alpha_{\ell-1}(\boldsymbol{x})$, with $\ell \leq n-1$, such that

$$\mathrm{ad}_f^{\ell}\boldsymbol{g} = \sum_{k=0}^{\ell-1} \alpha_k(\boldsymbol{x})\,\mathrm{ad}_f^k \boldsymbol{g}\,.$$

Now, relabel $\ell \to n-1$, and then apply the row vector ∇T_1 to the left on both sides. The first step is permissible, since the labels are arbitrary. This gives,

$$\frac{\partial T_1}{\partial \boldsymbol{x}}\,\mathrm{ad}_f^{\ell}\boldsymbol{g} = \sum_{k=n-\ell-1}^{n-2} \alpha_k(\boldsymbol{x}) \frac{\partial T_1}{\partial \boldsymbol{x}}\,\mathrm{ad}_f^k \boldsymbol{g}\,.$$

But from Part (b), the left-hand side is non-zero, while the right-hand side is zero, a contradiction. Hence the vector fields must be linearly independent.

d. From the Frobenius theorem, if the $n-1$ vector fields $\boldsymbol{g}$, $\mathrm{ad}_f \boldsymbol{g}$, …, $\mathrm{ad}_f^{n-2}\boldsymbol{g}$ satisfy

$$\frac{\partial T_1}{\partial \boldsymbol{x}}\,\mathrm{ad}_f^k \boldsymbol{g} = 0\,,$$

then they are involutive.

This sketch of the proof of the conditions for input-state linearization follows that of Slotine and Li (1991), who also show that the above conditions are sufficient for linearizability of the nonlinear, control-affine system, $\dot{\boldsymbol{x}} = \boldsymbol{f} + \boldsymbol{g}\,u(t)$.

11.15 Linear systems from a nonlinear point of view. By specializing the n-dimensional control-affine system of the form $\dot{\boldsymbol{x}} = \boldsymbol{f} + \boldsymbol{g}\,u(t) \;\to\; \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}u$, show that

$$\boldsymbol{W}_{\mathrm{c}} = \begin{pmatrix} \boldsymbol{g} & \mathrm{ad}_f \boldsymbol{g} & \cdots & \mathrm{ad}_f^{n-1}\boldsymbol{g} \end{pmatrix} \quad \to \quad \begin{pmatrix} \boldsymbol{B} & \boldsymbol{A}\boldsymbol{B} & \boldsymbol{A}^2\boldsymbol{B} & \cdots & \boldsymbol{A}^{n-1}\boldsymbol{B} \end{pmatrix}.$$

In both cases, the matrix $\boldsymbol{W}_{\mathrm{c}}$ must have rank n. But the nonlinear condition refers to feedback linearizability, while the linear condition also implies controllability.

Solution.

We have $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$ and $\boldsymbol{g} = \boldsymbol{B}$. Then, the Lie bracket

$$[\boldsymbol{f}, \boldsymbol{g}] = [\boldsymbol{A}\boldsymbol{x}, \boldsymbol{B}] = \cancelto{0}{\frac{\partial \boldsymbol{B}}{\partial \boldsymbol{x}}}(\boldsymbol{A}\boldsymbol{x}) - \frac{\partial(\boldsymbol{A}\boldsymbol{x})}{\partial \boldsymbol{x}}\boldsymbol{B} = -\boldsymbol{A}\boldsymbol{B}\,.$$

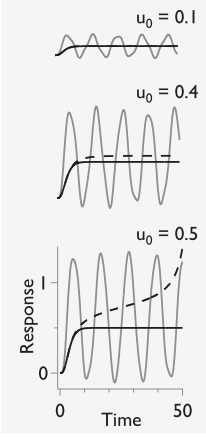
Similarly,

$$[\boldsymbol{f}, [\boldsymbol{f}, \boldsymbol{g}]] = [\boldsymbol{A}\boldsymbol{x}, -\boldsymbol{A}\boldsymbol{B}] = -\cancelto{0}{\frac{\partial(\boldsymbol{A}\boldsymbol{B})}{\partial \boldsymbol{x}}}(\boldsymbol{A}\boldsymbol{x}) + \frac{\partial(\boldsymbol{A}\boldsymbol{x})}{\partial \boldsymbol{x}}\boldsymbol{A}\boldsymbol{B} = +\boldsymbol{A}^2\boldsymbol{B}\,.$$

Further Lie brackets give $(-1)^k \boldsymbol{A}^k\boldsymbol{B}$. This reproduces $\boldsymbol{W}_{\mathrm{c}}$, except that odd columns have a minus sign. But this does not affect the rank of $\boldsymbol{W}_{\mathrm{c}}$, implying

that both cases have the condition of rank n. But only the linear case carries the further implication of controllability.

11.16 Flexible link. Consider controlling a pendulum via an elastic, "soft" torque such as that provided by a motor. The system may be used to model muscles, both real and artificial. In robotics jargon, it is known as a *single-link flexible joint*. With scaled units and simplified choices for spring constant and inertial moments, the equations of motion are $\ddot{\theta}_1 + \sin\theta_1 + (\theta_1 - \theta_2) = 0$ and $\ddot{\theta}_2 + (\theta_2 - \theta_1) = u(t)$, where θ_1 is the angle of the pendulum with respect to its down equilibrium, θ_2 is the angle of the input to the coupling, and u is the torque applied to the coupling.



a. Write these equations in control-affine form, $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})\,u(t)$.

b. Calculate the linearizability matrix $\boldsymbol{W}_c = (\boldsymbol{g} \;\; \mathrm{ad}_f\,\boldsymbol{g} \;\; \mathrm{ad}_f^2\,\boldsymbol{g} \;\; \mathrm{ad}_f^3\,\boldsymbol{g})$, and verify that it has full rank. We thus can linearize the system exactly.

c. Find the change of variable $\boldsymbol{z} = \boldsymbol{T}(\boldsymbol{x})$ and control that linearizes the system in the form of Eq. (11.20). Hint: See Problem 11.14b. In particular, show that

$$u = -\Big(\sin x_1(x_2^2 + \cos x_1 + 1) + (x_1 - x_3)(2 + \cos x_1)\Big) + v(t)\,.$$

d. Find the inverse transformation $\boldsymbol{x} = \boldsymbol{T}^{-1}(\boldsymbol{z})$ and use it to confirm explicitly that the dynamics are linear in the new variables: $\dot{z}_1 = z_2$, $\dot{z}_2 = z_3$, $\dot{z}_3 = z_4$, $\dot{z}_4 = v$.

e. Reproduce the plots at left, for step inputs of amplitude 0.1, 0.4, and 0.5. The light oscillating trace shows the undamped, open-loop response. The heavy dark line uses the exact feedback linearization, assuming that the linear part of the control is set to have 4 poles at -1. You should find a gain (row) vector $\boldsymbol{K} = (1\ 4\ 6\ 4)$. The dashed curve is based on the linearization of the system about $\theta_1 = \dot{\theta}_1 = \theta_2 = \dot{\theta}_2 = 0$. Its gain vector is designed so that the *linear* approximation has poles at -1. You should find $\boldsymbol{K}' = (-6 \;\; -4\ 3\ 4)$. The approximate design matches the exact linearization very well for $u_0 = 0.1$ but goes unstable for large input torque input, near $u_0 = 0.5$.

Solution.

a. In control-affine form,

$$\boldsymbol{f} = \begin{pmatrix} x_2 \\ -\sin x_1 - x_1 + x_3 \\ x_4 \\ x_1 - x_3 \end{pmatrix}, \qquad \boldsymbol{g} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

b. We calculate the Lie brackets as follows:

$$\mathrm{ad}_f\,\boldsymbol{g} = [\boldsymbol{f}, \boldsymbol{g}] = \cancelto{0}{\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}}\boldsymbol{f} - \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\boldsymbol{g} = -\begin{pmatrix} 0 & 1 & 0 & 0 \\ -\cos x_1 - 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & -1 & 0 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -1 \\ 0 \end{pmatrix}.$$

Next,

$$\mathrm{ad}_f^2\, \boldsymbol{g} = [\boldsymbol{f}, \mathrm{ad}_f\, \boldsymbol{g}] = -\begin{pmatrix} 0 & 1 & 0 & 0 \\ -\cos x_1 - 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & -1 & 0 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \\ -1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \end{pmatrix}.$$

Last,

$$\mathrm{ad}_f^3\, \boldsymbol{g} = [\boldsymbol{f}, \mathrm{ad}_f^2\, \boldsymbol{g}] = -\begin{pmatrix} 0 & 1 & 0 & 0 \\ -\cos x_1 - 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & -1 & 0 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \end{pmatrix}.$$

Putting these all together gives

$$\boldsymbol{W}_\mathrm{c} = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 1 & 0 & -1 & 0 \end{pmatrix},$$

which has rank=4 and det=1. Full rank implies that we can carry out feedback linearization.

c. Following the hint, we use (with $n = 4$),

$$\frac{\partial T_1}{\partial \boldsymbol{x}}\, \mathrm{ad}_f^k \boldsymbol{g} = 0 \text{ for } k = \{0, 1, 2\} \quad \text{and} \quad \frac{\partial T_1}{\partial \boldsymbol{x}}\, \mathrm{ad}_f^3 \boldsymbol{g} \neq 0\,,$$

substituting the results from Part b. Then,

$$\frac{\partial T_1}{\partial \boldsymbol{x}}\, \boldsymbol{g} = \frac{\partial T_1}{\partial x_4} = 0$$

$$\frac{\partial T_1}{\partial \boldsymbol{x}}\, \mathrm{ad}_f \boldsymbol{g} = -\frac{\partial T_1}{\partial x_3} = 0$$

$$\frac{\partial T_1}{\partial \boldsymbol{x}}\, \mathrm{ad}_f^2 \boldsymbol{g} = \frac{\partial T_1}{\partial x_2} - \cancelto{0}{\frac{\partial T_1}{\partial x_4}} = 0$$

$$\frac{\partial T_1}{\partial \boldsymbol{x}}\, \mathrm{ad}_f^3 \boldsymbol{g} = -\frac{\partial T_1}{\partial x_1} + \cancelto{0}{\frac{\partial T_1}{\partial x_3}} \neq 0\,.$$

Putting all these together, we conclude that $T_1(\boldsymbol{x}) = T_1(x_1)$ only. As in the text, we try the simplest choice, $T_1(x_1) = x_1$. We then use Eq. (11.28) to find T_2, T_3, and T_4. Thus,

$$T_2 = L_f T_1 = \frac{\partial T_1}{\partial \boldsymbol{x}}\, \boldsymbol{f} = \frac{\partial T_1}{\partial x_1}\, f_1 = f_1 = x_2\,.$$

Similarly,

$$T_3 = L_f T_2 = \frac{\partial T_2}{\partial \boldsymbol{x}}\, \boldsymbol{f} = \frac{\partial T_2}{\partial x_2}\, f_2 = f_2 = -\sin x_1 - x_1 + x_3\,.$$

Finally,

$$T_4 = L_f T_3 = \frac{\partial T_3}{\partial \boldsymbol{x}} \boldsymbol{f} = \begin{pmatrix} -(1+\cos x_1) & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_2 \\ -\sin x_1 - x_1 + x_3 \\ x_4 \\ x_1 - x_3 \end{pmatrix}$$

$$= -x_2 \cos x_1 + x_4 - x_2 \,.$$

Collecting the change of variables, we have

$$\boldsymbol{T}(\boldsymbol{x}) = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ -\sin x_1 - x_1 + x_3 \\ -x_2 \cos x_1 + x_4 - x_2 \end{pmatrix} .$$

The function $\beta(\boldsymbol{x})$ is given by

$$\beta^{-1} = \frac{\partial T_4}{\partial \boldsymbol{x}} \boldsymbol{g} = \left(\frac{\partial T_4}{\partial x_4} \right) = 1 \,,$$

and the function $\alpha(\boldsymbol{x})$ is given, from Eq. (11.21) by

$$\alpha\beta^{-1} = \alpha = -\frac{\partial T_4}{\partial \boldsymbol{x}} \boldsymbol{f} = -\begin{pmatrix} x_2 \sin x_1 & -\cos x_1 - 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_2 \\ -\sin x_1 - x_1 + x_3 \\ x_4 \\ x_1 - x_3 \end{pmatrix}$$

$$= -x_2^2 \sin x_1 + (\cos x_1 + 1)(-\sin x_1 - x_1 + x_3) - x_1 + x_3$$
$$= -\sin x_1 (x_2^2 + \cos x_1 + 1) - (x_1 - x_3)(2 + \cos x_1)$$

Thus, we find

$$u = -\left(\sin x_1 (x_2^2 + \cos x_1 + 1) + (x_1 - x_3)(2 + \cos x_1) \right) + v(t) \,.$$

d. By substituting, we easily see that the inverse transformation is given by

$$\boldsymbol{x} = \boldsymbol{T}^{-1}(\boldsymbol{z}) = \begin{pmatrix} z_1 \\ z_2 \\ z_1 + \sin z_1 + z_3 \\ z_2 + z_2 \cos z_1 + z_4 \end{pmatrix} .$$

Notice that $\boldsymbol{T}(\boldsymbol{x})$ and $\boldsymbol{T}^{-1}(\boldsymbol{z})$ are both continuous and everywhere, meaning that the coordinate transformation is a global diffeomorphism.

Let's substitute into $\boldsymbol{f}(\boldsymbol{x})$:

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{pmatrix} x_2 \\ -\sin x_1 - x_1 + x_3 \\ x_4 \\ x_1 - x_3 \end{pmatrix} = \begin{pmatrix} z_2 \\ -\sin z_1 - z_1 + \underbrace{z_1 + \sin z_1 + z_3}_{x_3} \\ \underbrace{z_2 + z_2 \cos z_1 + z_4}_{x_4} \\ z_1 - \underbrace{(z_1 + \sin z_1 + z_3)}_{x_3} \end{pmatrix} = \begin{pmatrix} z_2 \\ z_3 \\ z_2 + z_2 \cos z_1 + z_4 \\ -\sin z_1 - z_3 \end{pmatrix}.$$

Differentiating the inverse transform gives,

$$\dot{\boldsymbol{x}} = \begin{pmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \frac{\mathrm{d}}{\mathrm{d}t}(z_1 + \sin z_1 + z_3) \\ \frac{\mathrm{d}}{\mathrm{d}t}(z_2 + z_2 \cos z_1 + z_4) \end{pmatrix} = \begin{pmatrix} \dot{z}_1 \\ \dot{z}_2 \\ z_2 + (\cos z_1) z_2 + \dot{z}_3 \\ z_3 + z_3 \cos z_1 - z_2^2 \sin z_1 + \dot{z}_4 \end{pmatrix}.$$

Matching the first two equations, we see that we already have $\dot{z}_1 = z_2$ and $\dot{z}_2 = z_3$. The third equation is

$$\dot{z}_3 = -z_2 - z_2 \cos z_1 + \underbrace{z_2 + z_2 \cos z_1 + z_4}_{f_3} = z_4 \,.$$

Finally, the fourth equation is,

$$\dot{z}_4 = -z_3(1 + \cos z_1) + z_2^2 \sin z_1 + \underbrace{-\sin z_1 - z_3}_{f_4} + u \,.$$

We can transform it into the desired $\dot{z}_4 = v(t)$ by setting

$$u = z_3(2 + \cos z_1) + (1 - z_2^2) \sin z_1 + v(t) \,.$$

Back in the original coordinates, we have

$$\begin{aligned} u &= \underbrace{(-\sin x_1 - x_1 + x_3)}_{z_3}(2 + \cos x_1) + (1 - x_2^2) \sin x_1 + v \\ &= -\Big(\sin x_1 (x_2^2 + \cos x_1 + 1) + (x_1 - x_3)(2 + \cos x_1)\Big) + v \,, \end{aligned}$$

which is just the expression for u that we found above.

e. To do the numerics, you first need to simulate the open-loop system. Then simulate the closed-loop system based on the exact linearization. The control is modified in two ways: first, we use $u = \alpha + \beta v$; second, we substitute $v = -\boldsymbol{K}\boldsymbol{z}$, where we have to express the latter back in terms of the x components. Expicitly, this is

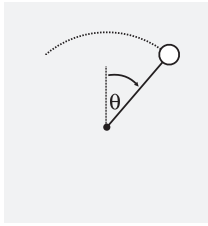
$$\begin{aligned} v &= -K_1 z_1 - K_2 z_2 - K_3 z_3 - K_4 z_4 \\ &= -K_1 x_1 - K_2 x_2 - K_3(-\sin x_1 - x_1 + x_3) - K_4(-x_2 \cos x_1 + x_4 - x_2) \,, \end{aligned}$$

which is included in the above expression for u. Finally, the linear approximation has its own gain vector, $\boldsymbol{K}'$. We again simulate the response using the nonlinear dynamics, but this time the control is just $u = -\boldsymbol{K}'\boldsymbol{x}$.

For small linearities, the approximation-based linearization works pretty well, with the main difference being in the value of the steady state (the difference between θ and $\sin\theta$). But for larger values, it means the difference between stability and instability!

Finally, although we wanted to work out everything explicitly in this solution, it is useful to note that programs such as *Mathematica* can do these calculations with built-in functions, in just a few lines of code.

11.17 Global Stabilization of the nonlinear pendulum. For $\ddot{\theta} - \sin\theta = u$ and for the "up" position ($\theta = \dot{\theta} = 0$), consider the function $V = 2a\sin^2\left(\frac{\theta}{2}\right) + \frac{1}{2}\dot{\theta}^2$.



a. For the uncontrolled system ($u = 0$), show that V is not a Lyapunov function.
b. Show that choosing $u = -K_p\sin\theta - K_d\dot{\theta}\cos^2\theta$ can make V a Lyapunov function that stabilizes the top position. In particular, show that you need to choose $a > 0$, $K_p = a + 1 > 1$, and $K_d > 0$. What is the physical significance of these conditions?
c. What if there is friction in the pendulum, so that $\ddot{\theta} + \lambda\dot{\theta} - \sin\theta = u$?
d. Example 11.8 notes that $\dot{V}$ is negative semi-definite, not negative definite. Yet the up solution is stable to almost all perturbations. Consider a new term in the control algorithm $u \to u - \epsilon\theta$, with θ defined to be in the range of $(-\pi, \pi)$. There is a discontinuity of amplitude 2ϵ when crossing the down position, $\theta = \pm\pi$. Is such a term helpful? Is the modified V still a Lyapunov function?
e. Plot $\theta(t)$, $V(t)$, and $u(t)$ for a perturbation of the form $\theta(0) = 0$, $\dot{\theta}(0) = 2$, and $a = K_d = 1$. Show that the controller recovers from a "kick" of almost 80°.

Solution.

a. We compute the time derivative for general u:

$$\dot{V} = 4a\sin\left(\frac{\theta}{2}\right)\cos\left(\frac{\theta}{2}\right)\left(\frac{1}{2}\right)\dot{\theta} + \dot{\theta}\ddot{\theta} = \dot{\theta}\,(a\sin\theta + \underbrace{\sin\theta + u}_{\ddot{\theta}})\,.$$

For $u = 0$, this gives

$$\dot{V} = (a+1)\,\dot{\theta}\,\sin\theta\,,$$

which can be both positive and negative, even when $a > 0$. Thus, V is not a Lyapunov function if $u = 0$.

b. For $u = -K_p\sin\theta - K_d\dot{\theta}\cos^2\theta$, with $K_p = a + 1$ and $K_d > 0$, we have

$$\begin{aligned}\dot{V} &= \dot{\theta}[(a+1)\sin\theta + u] \\ &= -K_d\,\dot{\theta}^2\,\cos^2\theta \le 0\,.\end{aligned}$$

This expression is negative semi-definite for $K_d > 0$, and V is now a Lyapunov function.

Physically, the condition $a > 0$ gives the Lyapunov function a minimum at $\theta = 0$ (up) and maximum at $\theta = \pi$ (down). For negative a, these are reversed, and we would stabilize the down position. The condition $K_p > 1$ implies

that we can supply enough torque to overcome the effects of gravity. That is, the actuator can directly swing up the pendulum to its up equilibrium. Example 7.2 discusses how to swing up a pendulum in the more-challenging *underactuated* case, where the maximum torque that can be supplied is less than the gravitational torque.

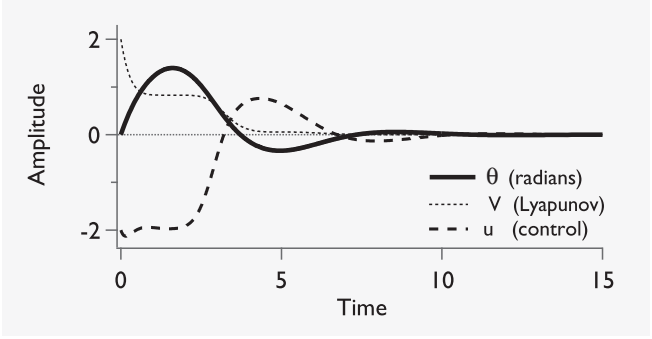
c. If there is a friction term $\lambda\dot{\theta}$ in the equation of motion, then $\dot{V} \rightarrow \dot{V} - \lambda\dot{\theta}^2$. In other words, friction helps stabilize the motion. We can use the previous controller and the response will be even faster.

d. The "down" solution ($\theta = \pi$, $\dot{\theta} = 0$ in our coordinates) is an unstable equilibrium. There is a set of measure zero of perturbations that will end up in these solutions. Notice that, for the down solution $u = 0$, meaning that if the down solution is reached, then the system is "stuck" there. Of course, in practice, there would never be such a perturbation, and the slightest deviation from the ideal would be enough to push the system out of the unstable down position. But we can easily imagine modifying the control law to avoid even this remote possibility. What we need is a control law that is *discontinuous* at $\theta = \pi$. Why discontinuous? Angles that are just less than π should be pushed "down" to zero. Angles just on the other side should be represented as near $-\pi$ and thus pushed "up" to zero. A term in the control law of the form

$$u \rightarrow u - \epsilon\theta$$

will work for small $\epsilon > 0$. We can replace θ by any monotonic function of θ, too. Here, we restrict the angle θ to be within the range $(-\pi, \pi)$, so that there is the required discontinuity at $\pm\pi$. If we adopt such a strategy, then we cannot have a Lyapunov function, which, by definition, must be differentiable. Again, we emphasize that this discussion is mostly to make a somewhat academic point. In practical applications, the Lyapunov design works well.

e. Below, we integrate the equations of motion for the closed-loop system, extracting at the same time V and u. The initial condition is $\theta(0) = 0$ and $\dot{\theta}(0) = 2$, and we use $a = K_d = 1$. Thus, we knock the pendulum nearly 80° away from the vertical, and it recovers easily. (Note that we set $\epsilon = 0$, too.)



It is instructive to look at the Lyapunov function and the feedback law for small deviations about the up equilibrium. Linearizing at $\theta = 0$ then gives an

approximate Lyapunov function,

$$V \approx 2a\left(\frac{\theta^2}{4}\right) + \frac{1}{2}\dot{\theta}^2 \approx \frac{1}{2}\left(a\theta^2 + \dot{\theta}^2\right).$$

For the controller,

$$u \approx -K_\mathrm{p}\theta - K_\mathrm{d}\dot{\theta},$$

which is just standard PD control.

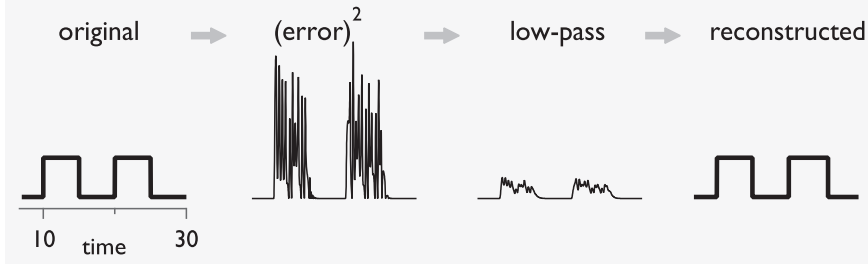
11.18 Sending "secret" messages by synchronized chaos. Section 10.5.1 introduced the Lorenz equations for state variables $\boldsymbol{x}^\mathsf{T} = (x, y, z)$. Here, think of this system as the *transmitter*, and set up a *receiver* with state variables $\boldsymbol{x}_\mathrm{r}^\mathsf{T} = (x_\mathrm{r}, y_\mathrm{r}, z_\mathrm{r})$ driven by $x(t)$ from the original system (but not in the $\dot{x}_\mathrm{r}$ equation). The two systems are

$$\begin{aligned} \dot{x} &= \sigma(y - x) & \dot{x}_\mathrm{r} &= \sigma(y_\mathrm{r} - x_\mathrm{r}) \\ \dot{y} &= x(r - z) - y & \dot{y}_\mathrm{r} &= x(r - z_\mathrm{r}) - y_\mathrm{r} \\ \dot{z} &= \underbrace{xy - bz}_{\text{transmitter}} & \dot{z}_\mathrm{r} &= \underbrace{xy_\mathrm{r} - bz_\mathrm{r}}_{\text{receiver}}. \end{aligned}$$

a. Simulate for $\sigma = 16$, $b = 4$, and $r = 45.6$. Demonstrate synchronization of the chaotic motion by plotting components of $\boldsymbol{x}_\mathrm{r}(t)$ vs. $\boldsymbol{x}(t)$. Also, plot the components of the error $\boldsymbol{e} \equiv \boldsymbol{x} - \boldsymbol{x}_\mathrm{r}$, and show that they decay exponentially.

b. Show that $V = \frac{1}{2}(\frac{1}{\sigma}e_x^2 + e_y^2 + e_z^2)$ is a Lyapunov function for this dynamical system. Why does this explain the synchronization?

c. Cuomo and Oppenheim built analog electronic circuits corresponding to these equations and demonstrated synchronization experimentally (Strogatz, 2014). They then communicated "secret messages" by both analog and digital techniques. To understand the latter, let the "signal" $m(t)$ be a stream of 0 and 1's – we will use a square wave but a random bit sequence will work equally well. Use m to modulate the b coefficient. That is, let $b \to b(t) = b + a\,m(t)$. Take $a = 0.4$. The altered $x(t)$ signal to the receiver, compute x_r, and then plot the quantity $e_x = (x_\mathrm{r} - x)^2$. Send e_x through a low-pass filter, and plot the result. Then apply a threshold: set the signal = 0 below a certain amplitude and 1 above it. Call this estimate $\hat{m}$ and compare to the original m (see below). Why does this scheme work?

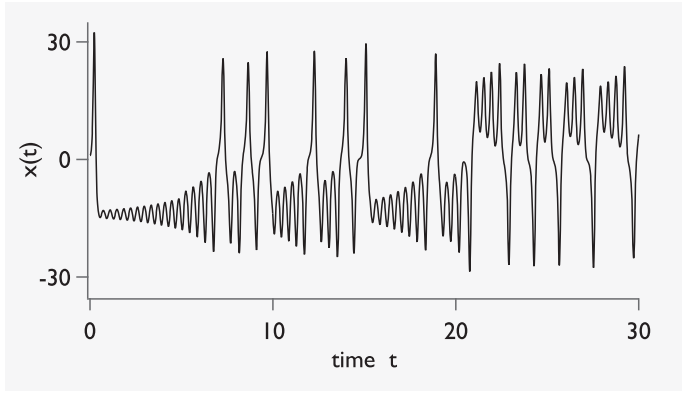


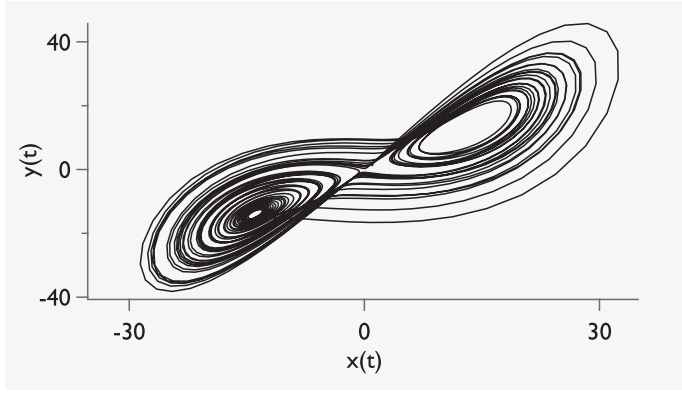
One note: although here the "carrier" of the message is chaotic and chaotic motion seems complicated, this problem shows that the message scheme is not

inherently secure, in that any eavesdropper could achieve the same synchronization as the intended receiver. Whatever robustness exists in the detection scheme will be available to all, and any desired security will arise by encrypting the message itself. For a discussion of why using a chaotic carrier might nonetheless be useful, see Abarbanel (2008). Quantum communication methods, by contrast, can provide inherently secure protection against eavesdroppers (Mermin, 2007).

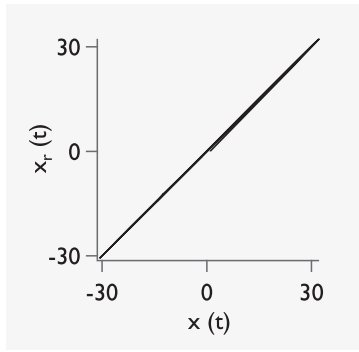
Solution.

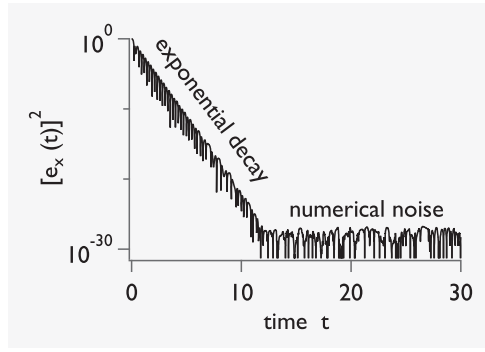
a. The parameters $\sigma = 16$, $b = 4$, and $r = 45.6$ are different from the set often used to demonstrate chaos in the Lorenz equations, but the motion and attractor are qualitatively the same. Below is a plot of $x(t)$ and $y(t)$ vs. $x(t)$. The latter shows a 2d projection of the 3d attractor.





Next, we look at synchronization. The plots below show, at left, $x_r(t)$ vs $x(t)$. As a diagonal line, it indicates synchronization. At right is $e_x = (x_r - x)^2$. It decays exponentially until numerical noise takes over. Thus, there is fast convergence to synchronization.





b. The candidate Lyapunov function is

$$V = \frac{1}{2}\left(\frac{1}{\sigma}e_x^2 + e_y^2 + e_z^2\right),$$

which is clearly positive definite, as it is the sum of squares. All three error components must vanish simultaneously, which is only possible at perfect synchronization. Next, we compute

$$\dot{V} = \frac{1}{\sigma}e_x\,\dot{e}_x + e_y\,\dot{e}_y + e_z\,\dot{e}_z\,.$$

The error dynamics are

$$\begin{aligned}\dot{e}_x &= \sigma\,(e_y - e_x)\\ \dot{e}_y &= x\,e_z - e_y\\ \dot{e}_z &= x\,e_y - b\,e_z\,.\end{aligned}$$

Substituting into the expression for $\dot{V}$ gives

$$\begin{aligned}\dot{V} &= \frac{1}{\sigma}e_x\,\dot{e}_x + e_y\,\dot{e}_y + e_z\,\dot{e}_z\\ &= e_x(e_y - e_x) + e_y(\cancel{x e_z} - e_y) + e_z(\cancel{x e_y} - b\,e_z)\\ &= -\left(e_x - \tfrac{1}{2}e_y\right)^2 - \tfrac{3}{4}e_y^2 - be_z^2\,,\end{aligned}$$

which is negative definite. Since the error converges to zero, the two trajectories are the same: this is what synchronization means.

c. The basic idea behind the digital communication scheme outlined in the text is that for $b = 0$, the two systems are identical and rapidly synchronize. When modulated to the "1" state, the value of b is different (4.4 in the example, rather than 4). The two systems no longer synchronize well. The mean-square error is then positive. Thus, in the scheme, we effectively look for period of small mean-square error and call them zero and one for larger amplitudes. The low-pass and threshold is just one algorithm to pick up this signature.

This example, due originally to K. Cuomo and A. Oppenheim, follows the description by Strogatz (2014). They also describe a way to transmit analog

signals, too, by adding the signal $m(t)$ to the output $x(t)$ and then setting $\hat{m} = x_r - x$. I found, numerically, that this can work approximately if the parameters are adjusted but that it is much less robust than the digital scheme.

This problem is nice because it combines several aspects of nonlinear control that are explored in this chapter: It uses Lyapunov functions; it synchronizes two dynamical systems; and, also, one system acts as a nonlinear observer.

11.19 Backstepping. Here is a recursive way to start from a Lyapunov function for a simple system and "extend" it to a new Lyapunov function for an enlarged dynamical system. Consider a system of the form $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})\,\nu$, with $\dot{\nu} = u$. Here, the input variable is u, as usual, and ν is an "extra" state variable, along with the n variables in x. Now pretend that ν is the control variable for $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})\,\nu$, and imagine that we know a Lyapunov function $V_0(\boldsymbol{x})$ that stabilizes $\boldsymbol{x} = \boldsymbol{0}$ for the system $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})\,\phi(\boldsymbol{x})$ for some appropriate "control" $\nu = \phi(\boldsymbol{x})$.

a. In the original problem, ν is not directly controlled. Define a new variable $z = \nu - \phi(\boldsymbol{x})$. Find a u to make a new Lyapunov function, $V = V_0 + \frac{1}{2}z^2$.
b. Use backstepping to find a control u and Lyapunov function V for $\dot{x}_1 = x_1^2 + x_2$, $\dot{x}_2 = u$ that stabilizes $x_1 = x_2 = 0$.

Khalil (2001) extends this idea to a chain of dynamical equations (not restricted to the simple integrator discussed here).

Solution.

a. Since $\nu = \phi(\boldsymbol{x})$ leads to a Lyapunov function V_0 for the first equation, let us try the change of variable $z = \nu - \phi$. The new equations of motion are then

$$\begin{aligned}\dot{\boldsymbol{x}} &= \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})\,(z + \phi)\\ \dot{z} &= \dot{\nu} - \dot{\phi} = u - \frac{\partial \phi}{\partial \boldsymbol{x}}\dot{\boldsymbol{x}} = u - \frac{\partial \phi}{\partial \boldsymbol{x}}[\boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})\,(z + \phi)]\,.\end{aligned}$$

Then $V = V_0 + \frac{1}{2}z^2$ is a Lyapunov function if

$$\begin{aligned}\dot{V} &= \dot{V}_0 + z\,\dot{z}\\ &= \frac{\partial V_0}{\partial \boldsymbol{x}}\dot{\boldsymbol{x}} + z\left\{u - \frac{\partial \phi}{\partial \boldsymbol{x}}[\boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})\,(z + \phi)]\right\}\,.\end{aligned}$$

Choosing

$$u = \frac{\partial \phi}{\partial \boldsymbol{x}}[\boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})\,(z + \phi)] - z$$

leads to $\dot{V} = \dot{V}_0 - z^2$. Since $\dot{V}_0$ is already negative (semi)-definite, so is V.

b. The dynamical system is

$$\begin{aligned}\dot{x}_1 &= x_1^2 + x_2\\ \dot{x}_2 &= u\,.\end{aligned}$$

If x_2 were a "control" for the x_1 equation, we could choose

$$x_2 = -x_1 - x_1^2 \equiv \phi(x_1)$$

and have $\dot{x}_1 = -x_1$, which has a Lyapunov function $V_0 = \frac{1}{2}x_1^2$.

Defining $z = \nu - \phi = x_2 + x_1 + x_1^2$ leads to new equations

$$\begin{aligned}\dot{x}_1 &= -x_1 + z \\ \dot{z} &= \dot{x}_2 + \dot{x}_1(1 + 2x_1) = u + (-x_1 + z)(1 + 2x_1)\,.\end{aligned}$$

We would like the full Lyapunov function to be $V = \frac{1}{2}(x_1^2 + z^2)$. To be a Lyapunov function, $\dot{V}$ must be negative (semi)-definite. Thus,

$$\begin{aligned}\dot{V} &= x_1\,\dot{x}_1 + z\,\dot{z} \\ &= x_1(-x_1 + z) + z(u + (-x_1 + z)(1 + 2x_1)) \\ &= -x_1^2 + z\,[x_1 + u + (-x_1 + z)(1 + 2x_1)]\,.\end{aligned}$$

Now choose u to cancel out the unwanted terms and leave $\dot{V}$ negative definite:

$$u = -x_1 + (x_1 - z)(1 + 2x_1) - z = 2\left[x_1^2 - z(1 + x_1)\right],$$

which implies $\dot{V} = -x_1^2 - z^2$.

As a check, we transform back to the original variables by substituting for z:

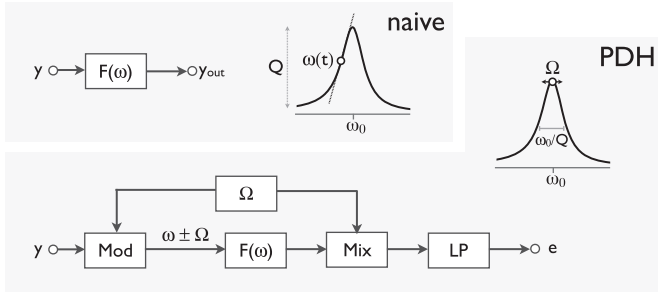
$$u = -2(x_1 + x_1^2 + x_1^3 + x_2 + x_1 x_2)\,.$$

11.20 Oscillator frequency stabilization. The *Pound–Drever–Hall* (PDH) technique is a way to stabilize the frequency of a tunable laser that is similar to *extremum-seeking control* (Section 10.1.2). It uses a standard PID loop based on an error signal proportional to the frequency shift between laser and reference cavity. The clever part is how the error signal is generated. Here, we discuss a simplified version of PDH that applies to a signal $y(t) = y(\omega)\ \mathrm{e}^{\mathrm{i}\omega t} + \text{c.c.}$, whose frequency $\omega(t)$ can drift in time and a reference passive filter, whose response has a fixed resonance frequency ω_0. Our focus will be on generating an error signal proportional to the shift in frequency, $e(t) \propto \omega_0 - \omega(t)$. In keeping with the optical applications, we will assume that ω is so large that we cannot measure the time-domain signal $y(t)$ directly but must rather measure its time-averaged power, $P(\omega) \equiv \langle |y(\omega)|^2 \rangle$, which can drift slowly.

a. We begin with a naive strategy that seems like it should work but has flaws. As shown below, we pass the signal, $y(t)$, through a second-order filter with highly underdamped, resonant response. Select ω_0 so that the $\omega(t)$ is located on the side of the resonance, at or near the point of maximum slope. Using a Taylor expansion of the power variation $\delta P = P - P_0$, show that we expect $\delta P = \left(\frac{\partial P}{\partial P_0}\right)\delta P_0 + \left(\frac{\partial P}{\partial \omega}\right)\delta\omega$, to first order in variations of signal power (δP_0)

and frequency ($\delta\omega$). We cannot distinguish between these variations without monitoring the signal power.

b. Now consider the PDH strategy, illustrated below. Set $\omega_0 \approx \omega(t)$ and add a deliberate phase modulation of amplitude a and frequency $\Omega \ll \omega_0/Q$, where Q is the enhancement factor of the resonance. Specifically, let the frequency be modulated so that $\omega(t) = \omega_0 + \delta\omega(t) + a\Omega\cos\Omega t$. Repeat the calculation from (a) and show, ignoring 2Ω and higher-frequency terms, that $\delta P = \left(\frac{\partial P}{\partial P_0}\right)\delta P_0 + \left(\frac{\partial^2 P}{\partial \omega^2}\delta\omega\right) a\Omega\cos\Omega t$, where $\left.\frac{\partial^2 P}{\partial \omega^2}\right|_{\omega_0} \delta\omega \equiv e(t)$ functions as an error signal. In the PDH strategy, we use a mixer to multiply the output of the filter $F(\omega)$ by $\cos\Omega t$ and a low-pass (LP) filter with cutoff frequency $\ll \Omega$. Explain why this isolates $e(t)$, the amplitude of the $\cos\Omega t$ term in δP and why the effects of amplitude variation in the original signal now give a higher-order contribution.



The real Pound–Drever–Hall technique not only adds the physics of laser cavities but also includes many subtleties, such as the advantages of working with a higher modulation frequency $\Omega > \omega_0/Q$, of working in reflection with an intensity minimum (as opposed to the maximum analyzed here), and the calculation of nonlinearities in the shape of the error function for larger frequency excursions (Black, 2001).

Solution.

a. The given expression is simply the Taylor expansion of the power at a frequency ω. We get both terms because the power has an overall factor $P_0(t)$ that can drift and well as a frequency $\omega(t)$ that can also drift. Choosing $\frac{\partial P}{\partial \omega}$ at the point of maximum slope helps in that it makes the coefficient of the variations that we care about ($\delta\omega$) as large as possible relative to the variations that we do not care about δP_0. Still, the fundamental problem is that if you use this signal to infer frequency shifts, you cannot be sure that you are not confusing an amplitude shift. One strategy, of course, is to do a separate monitoring of the power, eventually adding a feedback to stabilize it. That is a reasonable strategy, but the PDH strategy, as we will see below, does manage to isolate the frequency variation without the need for continual monitoring of the signal power.

b. The first term, $\frac{\partial P}{\partial P_0}$ is unchanged. To understand the second term, we calculate it in two stages. First, we Taylor expand in the frequency deviations. This gives a term of the form,

$$\left(\frac{\partial P}{\partial \omega}\right) a \cos \Omega t \,.$$

But now we have chosen a frequency near the top of the filter resonance where $\frac{\partial P}{\partial \omega} = 0$, so that we can do a further Taylor expansion for nearby frequencies:

$$\frac{\partial P}{\partial \omega} \approx \left.\frac{\partial^2 P}{\partial \omega^2}\right|_{\omega_0} \delta\omega \,.$$

Putting this result with the first gives the formula in the text. Choosing the filter frequency so that $\omega \approx \omega_0$ is important because changes to the intensity do not affect the equilibrium point (set point) in a feedback loop. (The error signal is zero whatever the intensity is.) Note that the second derivative is only an approximation that holds if Ω is small enough that $P(\omega)$ is approximately quadratic about the maximum (or minimum).

The mixer works by multiplying the filtered signal by $\cos \Omega t$. The reference has fixed frequency Ω, but the signal, because of the slow drifts, will actually have components at $\Omega' \approx \Omega$ (but not quite equal to it). The cosine-product trig identity

$$\cos \Omega t \, \cos \Omega' t = \frac{1}{2} \left(\cos(\Omega - \Omega')t + \cos(\Omega + \Omega')t\right) \,.$$

The term with $(\Omega - \Omega')$ is near DC, while the other is near 2Ω. By low-pass filtering, we can isolate the near-DC term, which is proportional to the amplitude of the $\cos \Omega t$ term in the drifting signal. All other terms, including the one near DC in the original signal, are filtered out after mixing with the reference signal.

11.21 Synchronization of the van der Pol oscillator to periodic forcing. Consider $\ddot{x} - \varepsilon(1 - x^2)\,\dot{x} + \omega_0^2 x = F \cos \omega t$, a forced version of the van der Pol equation introduced in Problem 7.17, with $0 < \varepsilon \ll 1$. The forcing amplitude is weak ($F \lesssim \varepsilon$), and the detuning small ($\omega/\omega_0 = 1 + O(\varepsilon)$).

a. Assume a solution of the form $x(t) = A\, \mathrm{e}^{\mathrm{i}\omega t} + \text{c.c.}$, with A an amplitude that can vary on time scales ε^{-1}. Show that $\dot{A} = O(\varepsilon)$ and $\ddot{A} = O(\varepsilon^2)$. Then show that

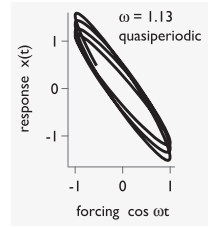
$$\dot{A} = -\mathrm{i}\left(\frac{\omega_0^2 - \omega^2}{2\omega}\right) A + \frac{1}{2}\varepsilon\left(1 - |A|^2\right) A - \mathrm{i}\,\frac{F}{4\omega} \,.$$

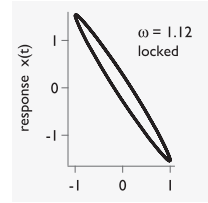
Hint: Write $\ddot{x} + \omega^2 x = \cdots$ and include all other $O(\varepsilon)$ terms on the RHS. Substitute for x, multiply by $\mathrm{e}^{-\mathrm{i}\omega t}$, and average over one period.

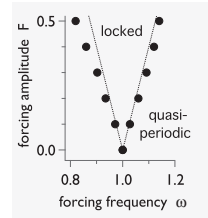
b. Rewrite (a) to give a generic amplitude equation, $a'(\tau) = -\mathrm{i}\nu a + (1 - |a|^2)a - \mathrm{i}f$, by rescaling and redefining quantities. Show that all terms are $O(1)$.

c. Write a in polar form, $a = r\,\mathrm{e}^{\mathrm{i}\phi}$, and show that $r' = (1-r^2)r - f\sin\phi$, with $\phi' = -\nu - \frac{f}{r}\cos\phi$. Does ϕ obey an Adler equation?

d. Simulate the forced van der Pol equation and find conditions for synchronization. Compare to the asymptotic analysis based on the Adler equation. Use the parametric plot (Lissajous figure) between $x(t)$ and $\cos\omega t$ to test for synchronization. Reproduce the plots above at left, where $\varepsilon = 0.1$, $\omega_0 = 1$, and (top two graphs) $f = 0.4$.







Solution.

a. We write the equations of motion

$$\ddot{x} + \omega^2 x = (\omega^2 - \omega_0^2)x + \varepsilon(1-x^2)\,\dot{x} + \omega_0^2 x + F\cos\omega t\,.$$

Let's substitute $x(t) = A\,\mathrm{e}^{\mathrm{i}\omega t} + \text{c.c.}$ into the LHS. Neglecting the $O(\varepsilon^2)$ term, we get

$$\begin{aligned}\ddot{x} + \omega^2 x &= -\omega^2 A\,\mathrm{e}^{\mathrm{i}\omega t} + 2\,\mathrm{i}\omega\dot{A}\,\mathrm{e}^{\mathrm{i}\omega t} + \cancel{\ddot{A}}\,\mathrm{e}^{\mathrm{i}\omega t} + \omega^2 A\,\mathrm{e}^{\mathrm{i}\omega t} \\ &\approx 2\,\mathrm{i}\omega\dot{A}\,\mathrm{e}^{\mathrm{i}\omega t}\,.\end{aligned}$$

We proceed similarly for the right-hand side. To simplify notation, we multiply both sides of the equation by $\mathrm{e}^{-\mathrm{i}\omega t}$. If we then average over a period $2\pi/\omega$, then all terms $\mathrm{e}^{\mathrm{i}m\omega t}$ with integer $m \neq 0$ will vanish. This leaves

$$2\,\mathrm{i}\omega\dot{A} = (\omega_0^2 - \omega^2)A + \mathrm{i}\omega\varepsilon A - \varepsilon|A|^2 A(\mathrm{i}\omega) + \frac{1}{2}F\,.$$

Here, we have changed $F\cos\omega t$ to $\frac{1}{2}F\,\mathrm{e}^{\mathrm{i}\omega t} + \text{c.c.}$ Also, in the nonlinear term, we have collected all contributions with two factors of $A\,\mathrm{e}^{\mathrm{i}\omega t}$ and one factor of $A^*\,\mathrm{e}^{-\mathrm{i}\omega t}$. These combine to give contributions proportional to $\mathrm{e}^{\mathrm{i}\omega t}$. Dividing by $2\,\mathrm{i}\omega$ gives the desired equation.

b. We define $\tau = \frac{1}{2}\varepsilon t$ to be the slow time variable and $A \to a$ (no change needed). Then

$$\frac{\varepsilon}{2}a'(\tau) = -\,\mathrm{i}\left(\frac{\omega_0^2 - \omega^2}{2\omega}\right)A + \frac{1}{2}\varepsilon\left(1 - |A|^2\right)A - \mathrm{i}\frac{F}{4\omega}\,.$$

Dividing by $\frac{\varepsilon}{2}$ gives

$$a'(\tau) = -\,\mathrm{i}\left(\frac{\omega_0^2 - \omega^2}{\omega\varepsilon}\right)A + \left(1 - |A|^2\right)A - \mathrm{i}\frac{F}{2\varepsilon\omega}\,.$$

Defining

$$\nu \equiv \left(\frac{\omega_0^2 - \omega^2}{\omega\varepsilon}\right), \qquad f \equiv \frac{F}{2\varepsilon\omega}$$

gives the desired equation. Notice that the assumptions on forcing and detuning imply that ν and f are both $O(1)$.

As a side remark, the forced van der Pol equation is special in lacking a nonlinear dependence of the limit cycle frequency on limit cycle amplitude (at $O(\varepsilon)$). More generally, the scaled amplitude equation takes the form

$$a'(\tau) = -\mathrm{i}\nu a + a - |a|^2 a - \mathrm{i}\alpha|a|^2 a - \mathrm{i}f\,,$$

where α gives the frequency shift due to amplitude a.

c. Let $a = r\,\mathrm{e}^{\mathrm{i}\phi}$. Then $a' = (r' + \mathrm{i}\phi' r)\,\mathrm{e}^{\mathrm{i}\phi}$. Substituting and then separating real and imaginary terms gives the desired equations. Note that we use $f\,\mathrm{e}^{-\mathrm{i}\phi} = f\,\cos\phi - \mathrm{i}f\,\sin\phi$.
We have already required F to be of $O(\varepsilon)$. If we further require $f \ll 1$, then

$$r' = (1 - r^2)f - f\sin\phi$$

has a solution $r \approx 1$. The f term is then a small perturbation. By contrast, using $r = 1$ in the ϕ equation gives

$$\phi' = -\nu - f\cos\phi\,,$$

which is a form of the Adler equation. (A trivial $\pi/2$ phase shift in ϕ converts the cosine into a sine.) Unlike the r equation, the phase ϕ is neutral to constant perturbations (the $\phi' = -\nu$ part). This is an explicit illustration of our assertion that phase perturbations are "soft" while amplitude perturbations are "hard."

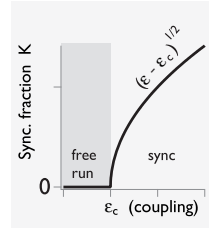
d. See code on book website. To reproduce the Lissajous figures, you have to eliminate transient motion. The figures shown simulate to $t = 500$ and show only the last 20 time units.

The plot shows the numerically obtained boundaries between locked and quasiperiodic (unlocked) motion, as judged crudely by Lissajous figures, as demonstrated by the top two graphs. We can compare to the predictions of the Adler equation. In scaled units, this is just $f = |\nu|$. Going back to the original quantities, F, ω, and $\omega_0 = 1$, we see that we expect lines $F = 1 \pm 4(\omega - 1)$. These are the two straight lines shown in the plot. They are consistent with the numerical results at small forcing F and start to deviate at higher values. (However, the method used to estimate the boundaries numerically is, as already noted, rather crude.)

Note that an alternative to using a Lissajous figure to detect phase relations is to use the *Hilbert transform*, which leads to a direct estimate of the phase of the response. In the locked regime, this phase will be a constant relative to that of the cos ωt term. In the quasiperiodic regime, it will increase (or decrease) linearly in time. This and many other details on this problem may be found in Pikovsky et al. (2001).

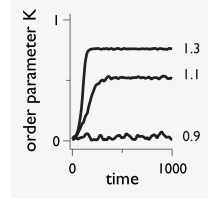
11.22 Synchronization of N globally coupled oscillators (Example 11.9). Consider N coupled oscillators obeying Eq. (11.53), with frequencies ω_k having symmetric distribution $g(\omega)$ and mean ω_0. Define the complex order parameter $\mathcal{K} = K\,\mathrm{e}^{\mathrm{i}\Theta}$, which can be viewed as a "mean field" that "externally" forces the oscillators.

a. Show that Eq. (11.53) can be rewritten as $\dot{\phi}_k = \omega_k + \varepsilon K \sin\big(\Theta - \phi_j\big)$.
b. Show that the synchronized oscillator frequency $\approx \omega_0$ and that $\psi_k \equiv \phi_k - \omega_0 t$ obeys the Adler equation $\dot{\psi}_k = \omega_k - \omega_0 - \varepsilon K \sin\psi_k$.
c. Argue that for $N \to \infty$, a self-consistent condition for the order parameter is $\mathcal{K} = K\, e^{i\Theta} = \int_{-\pi}^{\pi} d\psi\, e^{i\phi}\, n_s(\psi)$, where $n_s(\phi)$ is the distribution of synchronized oscillators having phase ϕ. Why do only the *synchronous* (and not the asynchronous) oscillators contribute? Why is the integral over ψ?
d. Rewrite the self-consistent equations for $\mathcal{K}$ as $1 = \varepsilon \int_{-\pi/2}^{\pi/2} d\psi \cos^2\psi\, g(\omega)$ and $0 = \varepsilon \int_{-\pi/2}^{\pi/2} d\psi \cos\psi\, \sin\psi\, g(\omega)$, where $\omega = \omega_0 + \varepsilon K \sin\psi$.
e. Argue that the second equation is satisfied if the mean frequency is ω_0, as we guessed. Taylor expand the first equation about ω_0 and show that there is a phase transition at $\varepsilon_c = \frac{2}{\pi g(\omega_0)}$ from $K = 0$ to $K^2 = \frac{8g(\omega_0)}{|g''(\omega_0)|\,\varepsilon_c^3}(\varepsilon - \varepsilon_c)$, as shown at right.
f. Integrate numerically the equations in (a), with $N = 10^5$ and $g(\omega) \sim \mathcal{N}(1, 0.1^2)$. Reproduce the order-parameter plot at right for $K(t)$. The values by the three curves indicate $\varepsilon/\varepsilon_c$. Show that the synchronization threshold $\varepsilon_c = \sqrt{8/\pi}\,\sigma$.



For details, see Kuramoto (1984) and also Pikovsky et al. (2001), Section 12.1.

Solution.



a. First, we note that we can separate the equation defining the order parameter,

$$\mathcal{K} = K\, e^{i\Theta} = \frac{1}{N}\sum_{j=1}^{N} e^{i\phi_k}\ .$$

into real and imaginary equations:

$$K\cos\Theta = \frac{1}{N}\sum_{j=1}^{N}\cos\phi_k\,, \qquad K\sin\Theta = \frac{1}{N}\sum_{j=1}^{N}\sin\phi_k\,.$$

Then,

$$\begin{aligned}\frac{d\phi_k}{dt} &= \omega_k + \frac{\varepsilon}{N}\sum_{j=1}^{N}\sin\big(\phi_j - \phi_k\big)\\ &= \omega_k + \frac{\varepsilon}{N}\sum_{j=1}^{N}\big(\sin\phi_j\cos\phi_k - \cos\phi_j\sin\phi_k\big)\\ &= \omega_k + \varepsilon K\,(\sin\Theta\cos\phi_k - \cos\Theta\sin\phi_k)\\ &= \omega_k + \varepsilon K\sin(\Theta - \phi_k)\,.\end{aligned}$$

Notice that if Θ and K were fixed external variables, we would have N separate equations describing how each oscillator is forces by the external periodic forcing. Here, the "external" forcing is provided by the synchronized oscillators themselves.

b. The symmetry of $g(\omega)$ implies that the synchronized oscillators must oscillate at the mean frequency ω_0, which is also the peak of the distribution g. Indeed, ω_0 is the only "special" frequency. With a more general, non-symmetric distribution, the synchronization frequency must be determined self-consistently, in a way that is similar to the equations we will write below for the synchronization condition. Then, using the definition $\psi_k = \phi_k - \omega_0 t$ and the result from part (a), we write

$$\begin{aligned}\frac{\mathrm{d}\psi_k}{\mathrm{d}t} &= \frac{\mathrm{d}\phi_k}{\mathrm{d}t} - \omega_0 \\ &= \omega_k - \omega_0 + \varepsilon K \sin(\Theta - \phi_k) \\ &= \omega_k - \omega_0 + \varepsilon K \sin(\omega_0 t - \phi_k) \\ &= \omega_k - \omega_0 - \varepsilon K \sin\psi_k \,.\end{aligned}$$

c. We can view the order parameter as the estimator of an average quantity. The self-consistent equation just writes down this average in terms of the probability distribution $n_s(\psi)$. Only the synchronized oscillators contribute: intuitively, the order parameter reflects the contribution of the synchronized oscillators, with the contributions of the unsynchronized fraction averaging to zero for $N \to \infty$. The integral should be over the population of synchronized oscillators. It is convenient to use ψ as the integration variable (and not ϕ) because that is the stationary variable.

d. To convert from the distribution of oscillators by frequency, $g(\omega)$, to the distribution of synchronous oscillators by phase, $n_s(\psi)$, we use the change-of-variables formula of probability theory:

$$n_s(\psi) = g(\omega)\left|\frac{\mathrm{d}\omega}{\mathrm{d}\psi}\right| = g(\omega_0 + \varepsilon K \sin\psi)\,\varepsilon K \cos\psi \,.$$

Because probability distributions must be positive functions (as reflected in the absolute value in the Jacobian factor), the range of ψ is restricted to $(-\frac{\pi}{2}, +\frac{\pi}{2})$, which makes $\cos\psi > 0$. Then, with $\Theta = \omega_0 t$ and $\phi = \psi + \omega_0 t$, the self-consistent order-parameter equation becomes

$$\begin{aligned}\mathcal{K} = K\,\mathrm{e}^{\mathrm{i}\Theta} &= \int_{-\pi}^{\pi} \mathrm{d}\psi\,\mathrm{e}^{\mathrm{i}\phi}\; n_s(\psi) \\ \cancel{K}\cancel{\mathrm{e}^{\mathrm{i}\omega_0 t}} &= \int_{-\pi}^{\pi} \mathrm{d}\psi\,\mathrm{e}^{\mathrm{i}\psi}\;\cancel{\mathrm{e}^{\mathrm{i}\omega_0 t}}\, g(\omega_0 + \varepsilon K \sin\psi)\,\varepsilon\cancel{K}\cos\psi \,, \\ 1 &= \varepsilon \int_{-\pi}^{\pi} \mathrm{d}\psi \cos\psi\;\mathrm{e}^{\mathrm{i}\psi}\; g(\omega_0 + \varepsilon K \sin\psi) \,.\end{aligned}$$

Writing the real and imaginary parts separately then gives,

$$\begin{aligned}1 &= \varepsilon \int_{-\pi/2}^{\pi/2} \mathrm{d}\psi \cos^2\psi\, g(\omega_0 + \varepsilon K \sin\psi) \,, \\ 0 &= \varepsilon \int_{-\pi/2}^{\pi/2} \mathrm{d}\psi \cos\psi\,\sin\psi\, g(\omega_0 + \varepsilon K \sin\psi) \,.\end{aligned}$$

The integration range is now $(-\frac{\pi}{2}, +\frac{\pi}{2})$ to reflect the support of $n_s(\psi)$.

e. The second equation determines the frequency, which we already assumed was equal to ω_0. Thus, all we must do is verify that it is satisfied. It is, since cos and g are even functions of ψ while $\sin\psi$ is odd. Had we not been smart enough to guess that symmetry implies that the synchronization frequency must be ω_0, this equation would have shown us that.
For the first equation, Taylor expanding g to second order about ω_0 gives

$$g(\omega_0 + \varepsilon K \sin\psi) \approx g(\omega_0) + \frac{1}{2}(\varepsilon K \sin\psi)^2 .$$

Then substitute:

$$\begin{aligned} 1 &= \varepsilon \int_{-\pi/2}^{\pi/2} \mathrm{d}\psi \cos^2\psi \left[g(\omega_0) + \frac{1}{2}(\varepsilon K \sin\psi)^2 \right] , \\ &\approx \varepsilon \frac{\pi}{2} g(\omega_0) + \varepsilon^3 K^2 g''(\omega_0) \frac{\pi}{16} . \end{aligned}$$

One solution to this equation is

$$K = 0 , \qquad \varepsilon = \varepsilon_\mathrm{c} = \tfrac{2}{\pi g(\omega_0)} .$$

For $\varepsilon > \varepsilon_\mathrm{c}$ and $|K|$ small, we can expand in $(\varepsilon - \varepsilon_\mathrm{c})$ and find

$$0 = (\varepsilon - \varepsilon_\mathrm{c}) \frac{\pi}{2} g(\omega_0) + \varepsilon_\mathrm{c}^3 K^2 g''(\omega_0) \frac{\pi}{16} ,$$

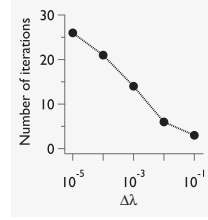
which implies

$$K^2 = \frac{8 g(\omega_0)}{|g''(\omega_0)|\, \varepsilon_\mathrm{c}^3} (\varepsilon - \varepsilon_\mathrm{c}) .$$

Again, we note that this equation is valid only near onset, i.e., for $\varepsilon \gtrsim \varepsilon_\mathrm{c}$ and small K. Note also that $K \geq 0$, so that we must take the positive square root.

f. See website for code.

11.23 Controlling chaos via the OGY method. Implement the OGY method and targeting, to reproduce the plots in Figure 11.6. For targeting, write a function that expands the range $\lambda \pm \Delta\lambda$ and compute the range of possible images $x_1 \pm \Delta x_1$. Then count the number of further iterations (using λ as control parameter) to expand Δx_1 to a larger range Δx_n that includes the target x^*. You now have a function between the range $\lambda \pm \Delta\lambda$ and $x_n \pm \Delta x_n$ that includes x^*. Use this function in a root-finder routine to predict the value of the perturbation λ' that brings the system to x^* in n iterations. Verify that the number of required iterations grows logarithmically as the perturbation tolerance $\Delta\lambda$ is reduced. For plots, use $\varepsilon = 0.02$ and $x_0 = 0.5$.



Solution.

See book website for code.

11.24 Time-delayed feedback for chaotic systems. Consider a control algorithm for a signal $y(t)$ that is based on $K[y(t) - y(t - \tau)]$, where K is a feedback gain and τ is the period of limit cycle that you wish to stabilize. The control signal will

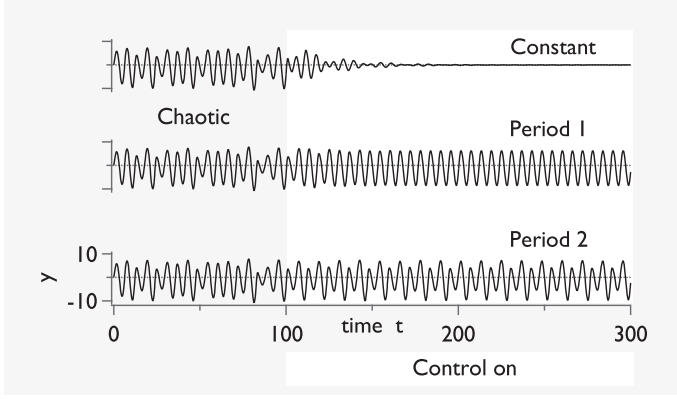
vanish when $y(t)$ is periodic with period τ. Use this idea to stabilize the Rössler equations, a canonical dynamical system exhibiting chaos. The equations for a three-dimensional state vector $(x\ y\ z)^{\mathsf{T}}$ are $\dot{x} = -y - z$ and, $\dot{y} = x + ay - K[y(t) - y(t-\tau)]$, and $\dot{z} = b + z(x - c)$.

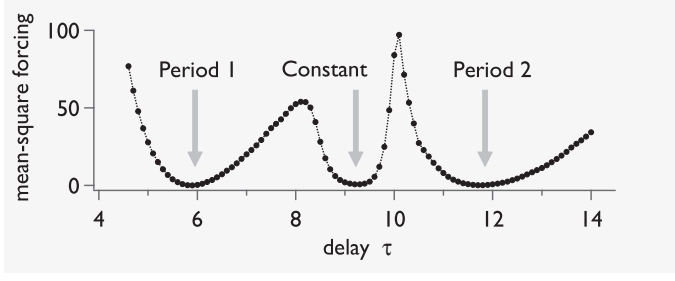
a. Simulate a time series $y(t)$ for $0 \le t \le 300$, using $a = b = K = 0.2$ and $c = 5.7$. Start the control at $t = 100$. For $K = 0$ (no control), the motion should be chaotic.
b. The method needs the period τ. To find τ from the motion itself, calculate the mean-square error of the control signal (after transients have died away), as a function of τ. From the minima locations, find the periods to use in part (a).

Cf. Pyragas and Pyragas (2011), who show how to tune τ adaptively. Note that using a delayed signal turns ordinary differential equations into *delay-differential equations*, whose infinite-dimensional state spaces are hard to analyze.

Solution.

See book website for code. You should find graphs resembling those below.





PART III

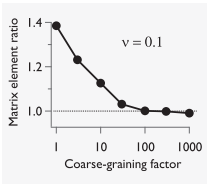
SPECIAL TOPICS

12 Discrete-State Systems

Problems

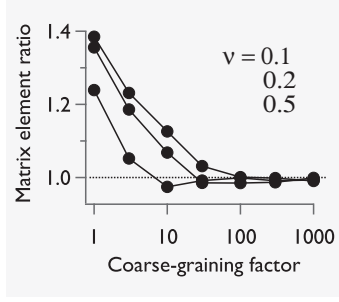
12.1 Coarse graining. There are two steps: time averaging and then a nonlinear "classification." Here, we investigate the choice of time-averaging scale.

a. Write a code to make time series plots similar to those in Figure 12.3 for $x_{k+1} = x_k + a\left(x_k - x_k^3\right) + \nu_k$, with $\nu_k \sim \mathcal{N}(0, \nu^2)$. The control parameters are a and ν.
b. Investigate the role of averaging time in coarse graining and reproduce Fig. 12.4.
c. Show that if the coarse-graining factor is too small, the resulting process is not Markov by using the update law, $p_{k+1} = \boldsymbol{A}\, p_k$ twice: $p_{k+2} = \boldsymbol{A}^2 p_k$. Compare with $p_{k+2} = (\boldsymbol{A}_2)\, p_k$, where $\boldsymbol{A}_2$ can be empirically estimated by looking at frequencies of the four different state combinations. Does $\boldsymbol{A}^2 = \boldsymbol{A}_2$? Average a simulated time series by a "coarse-graining factor" ; then compute the ratio of off-diagonal matrix elements $(\boldsymbol{A}^2)_{01} / (\boldsymbol{A}_2)_{01}$ as a function of ν and coarse graining (see left).



Solution.

a. Simulations were done using 10^7 time steps in the raw (fast) time series, before averaging by the coarse-graining factor.
b. Time series for $\nu = 0.15$. You should get something resembling Fig. 12.4.
c. The ratio converges faster for larger ν, as illustrated below.



Taking into account the constraints of accuracy (previous part) and Markov dynamics (present part), we see that a factor of about 100 is appropriate for

$\nu = 0.15$. More generally, for dynamics on a time scale τ, the averaging time should be somewhat shorter than τ. We want to average as much as possible to avoid high-frequency dynamics, but we do not want to lose any of the dynamics at the chosen time scale. Coarse-graining works best when there is a clear separation between the time scale of the desired dynamics and the slowest time scale of uninteresting dynamics.

Note: An alternative test is to see whether dwell-time distribution in each state is exponential.

12.2 Equilibrium and steady states of a Markov chain. A steady state is *reversible* if the stochastic process forward in (discrete) time is indistinguishable from the backward process. The steady state in a reversible Markov chain is also termed an *equilibrium state*, as it is closely connected to the notion of thermodynamic equilibrium. In the following, the matrix element A_{ij} is the $j \to i$ transition probability.

a. *Detailed balance* for a homogeneous Markov chain is defined by $A_{ij}\, p_j = A_{ji}\, p_i$, for all i and j. Show that reversibility implies detailed balance, and vice versa. Hints: Sum i over all n states. Also, consider $P(x_{k+1} = i, x_k = j)$ and its time reversal, along with sequences of N elements and their time reversal.
b. Show that the steady state of the two-state Markov model in Example 12.1 obeys detailed balance and is hence an equilibrium state.
c. For a diagonalizable, stochastic transition matrix $\boldsymbol{A}$ with no zero entries, show that $\lim_{N\to\infty} \boldsymbol{A}^N = \boldsymbol{P}$, where each column of $\boldsymbol{P}$ is the steady-state distribution $\boldsymbol{p}$. Please interpret. Use the fact (or prove) that $\lambda = 1$ is the largest eigenvalue. Hints: $\boldsymbol{p}$ is a right eigenvector, and there is also a left eigenvector of all ones.
d. In order to use the detailed-balance condition for equilibrium, we have to first solve for $\boldsymbol{p}$. Kolmogorov derived a condition for equilibrium that depends only on the transition probabilities A_{ij}. In particular, a stationary Markov chain is reversible and obeys detailed balance if and only if

$$A_{\ell_1\ell_2}\, A_{\ell_2\ell_3}\, \ldots\, A_{\ell_{N-1}\ell_N}\, A_{\ell_N\ell_1} = A_{\ell_N\ell_{N-1}}\, A_{\ell_{N-1}\ell_{N-2}}\, \ldots\, A_{\ell_2\ell_1}\, A_{\ell_1\ell_N}$$

for *every* finite sequence of states $\ell_1, \ell_2, \ldots, \ell_N$ for any length N. Show that detailed balance implies Kolmogorov's condition. The converse is trickier: consider a very long path, fix $\ell_1 = i$ and $\ell_N = j$, sum over all intermediate states $\ell_2, \ldots, \ell_{N-1}$, and use the identity from (c). Kolmogorov's condition implies that clockwise and counterclockwise probability currents around a loop are equal for an equilibrium (reversible) state. Here, the $j \to i$ current is $J_{ij} = A_{ij}\, p_j - A_{ji}\, p_i$. In equilibrium, detailed balance implies $J_{ij} = 0$: the *only* way to have a nonequilibrium steady state is to have a loop with differing clockwise and counterclockwise probability currents. Finally, as a corollary of Kolmogorov's criterion, show that steady states must be reversible for *trees* – graphs with no loops. Cf. Kelly (1979).

Solution.

a. First, we assume time-reversal symmetry and prove detailed balance. We denote the n states by $\{1, \ldots, n\}$, with probabilities p_1 to p_n. Then

$$P(x_{k+1} = i, x_k = j) = P(x_{k+1} = i | x_k = j)\, P(x_k = j) = A_{ij}\, p_j\,.$$

But reversibility implies that this is also equal to the time-reversed joint probability

$$P(x_{k+1} = j, x_k = i) = P(x_{k+1} = j | x_k = i)\, P(x_k = i) = A_{ji}\, p_i\,,$$

and hence we deduce the detailed-balance condition.

To prove that detailed balance implies reversibility, start with

$$A_{ij}\, p_j = A_{ji}\, p_i\,.$$

Since columns of $\boldsymbol{A}$ sum to unity (stochastic matrix), the sum over i is

$$\sum_{i=1}^{n} A_{ij}\, p_j = \left(\sum_{i=1}^{n} A_{ij} \right) p_j = (1)\, p_j = \sum_{i=1}^{n} A_{ji}\, p_i\,.$$

The last relation expresses $\boldsymbol{p} = \boldsymbol{A}\boldsymbol{p}$ in component form. Thus, detailed balance implies steady state. To prove that it also implies reversibility, write the joint probability of an N-element time series:

$$\begin{aligned} P(x_1 = \ell_1, \ldots, x_N = \ell_N) &= P(x_1 = \ell_1)\, P(x_2 = \ell_2 | x_1 = \ell_1) \ldots \\ &\qquad P(x_N = \ell_N | x_{N-1} = \ell_{N-1}) \\ &= p_{\ell_1}\, A_{\ell_2 \ell_1} \ldots A_{\ell_N \ell_{N-1}}\,, \end{aligned}$$

where ℓ_k denotes the state at time $k \in (0, N)$. The time-reversed sequence for the same set of states is

$$\begin{aligned} P(x_1 = \ell_N, \ldots, x_N = \ell_1) &= P(x_1 = \ell_N)\, P(x_2 = \ell_{N-1} | x_1 = \ell_N) \ldots \\ &\qquad P(x_N = \ell_1 | x_{N-1} = \ell_2) \\ &= p_{\ell_N}\, A_{\ell_{N-1} \ell_N} \ldots A_{\ell_1 \ell_2}\,. \end{aligned}$$

Then using detailed balance repeatedly shows that these forward and backwards sequences are identical.

b. From Example 12.1, we have

$$\boldsymbol{A} = \begin{pmatrix} 1 - a_0 & a_1 \\ a_0 & 1 - a_1 \end{pmatrix}, \qquad \boldsymbol{p} = \tfrac{1}{a_0 + a_1} \begin{pmatrix} a_1 \\ a_0 \end{pmatrix}.$$

Then, the detailed balance condition $A_{10}\, p_0 = A_{01}\, p_1$ is

$$a_0 \left(\frac{a_1}{a_0 + a_1} \right) = a_1 \left(\frac{a_0}{a_0 + a_1} \right),$$

which clearly holds for all valid a_0 and a_1.

c. The elements of the stochastic $\boldsymbol{A}$ are all in the range $(0, 1)$. Since the sum of each column is 1, the row vector $\boldsymbol{v} = (1\,1\,1, \ldots, 1)$ is a left eigenvector with eigenvalue 1. Thus, there is a (left) eigenvector with positive entries (all one) with eigenvalue one.

The *Perron Frobenius Theorem* asserts that, for a positive matrix $\boldsymbol{A}$, there is a *unique* eigenvector (up to constant scaling) with all positive entries whose simple eigenvalue $\lambda_{\max}$ is real and positive. Further, for all other eigenvalues, $|\lambda| < \lambda_{\max}$. For $\boldsymbol{A}$, we have proven that the (left) eigenvector is all positive and that the corresponding eigenvalue is also 1. Thus, $\lambda_{\max} = 1$.
Writing $\boldsymbol{A}$ in diagonal form, we have

$$\boldsymbol{A}^N = \boldsymbol{R}\boldsymbol{D}^N\boldsymbol{R}^{-1},$$

where the diagonal matrix is raised to the N'th power. Since the largest eigenvalue is 1, we have

$$\boldsymbol{D}^N = \begin{pmatrix} 1^N & & & \\ & \lambda_2^N & & \\ & & \ddots & \\ & & & \lambda_n^N \end{pmatrix} \to \begin{pmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix}.$$

Then we note that the matrix $\boldsymbol{R}$ has the structure $\boldsymbol{p}, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$. That is, each column is a right eigenvector of $\boldsymbol{A}$, with $\boldsymbol{p}$ corresponding to $\lambda = 1$ and $\boldsymbol{v}_2$ corresponding to λ_2 and so on. Similarly, $\boldsymbol{R}^{-1}$ is made up of row vectors that are the left eigenvectors of $\boldsymbol{A}$. The top row vector corresponds to $\lambda = 1$ and is given by all ones: $1\,1\,1\,\ldots\,1$. This arises from the normalization condition for all columns of $\boldsymbol{A}$. Now, we put this all together to write

$$\begin{aligned} \boldsymbol{P} \equiv \lim_{N\to\infty} \boldsymbol{A}^N &= \begin{pmatrix} p_1 & & & \\ p_2 & & & \\ \vdots & & & \\ p_n & & & \end{pmatrix} \begin{pmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ & & & \\ & & & \\ & & & \end{pmatrix} \\ &= \begin{pmatrix} p_1 & p_1 & \cdots & p_1 \\ p_2 & p_2 & & p_2 \\ \vdots & \vdots & & \vdots \\ p_n & p_n & & p_n \end{pmatrix}. \end{aligned}$$

To interpret this result, we notice that $\boldsymbol{P}\boldsymbol{v} = \boldsymbol{p}$, for *any* normalized initial distribution $\boldsymbol{v}$. We can see this by writing

$$\boldsymbol{P}\boldsymbol{v} = \begin{pmatrix} p_1 & p_1 & \cdots & p_1 \\ p_2 & p_2 & & p_2 \\ \vdots & \vdots & & \vdots \\ p_n & p_n & & p_n \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} p_1(\sum v_i) \\ p_2(\sum v_i) \\ \vdots \\ p_n(\sum v_i) \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix} = \boldsymbol{p},$$

since $\sum_{i=1}^{n} v_i = 1$ (normalization). Thus, whatever the initial distribution $\boldsymbol{v}$, the long-time dynamics of the Markov chain converges to $\boldsymbol{p}$, as expected.

d. We first assume reversibility. Then detailed balance holds, and we can write, for a cycle of N states,

$$\begin{aligned} A_{\ell_1\ell_2}\, p_{\ell_2} &= A_{\ell_2\ell_1}\, p_{\ell_1} \\ A_{\ell_2\ell_3}\, p_{\ell_3} &= A_{\ell_3\ell_2}\, p_{\ell_2} \\ &\vdots \\ A_{\ell_{N-1}\ell_N}\, p_{\ell_N} &= A_{\ell_N\ell_{N-1}}\, p_{\ell_{N-1}} \\ A_{\ell_N\ell_1}\, p_{\ell_1} &= A_{\ell_1\ell_N}\, p_{\ell_N}\,. \end{aligned}$$

Then multiplying all these equations together and canceling the common factors of p gives Kolmogorov's condition.

Going the other way—proving detailed balance starting from Kolmogorov's condition—is trickier. We outline the basic steps here, following Kelly (1979). We begin with Kolmogorov's condition, recopied for convenience:

$$A_{\ell_1\ell_2}\, A_{\ell_2\ell_3}\, \ldots\, A_{\ell_{N-1}\ell_N}\, A_{\ell_N\ell_1} = A_{\ell_N\ell_{N-1}}\, A_{\ell_{N-1}\ell_{N-2}}\, \ldots\, A_{\ell_2\ell_1}\, A_{\ell_1\ell_N}$$

Now, let us fix states $\ell_1 = i$ and $\ell_N = j$ and sum over all possible intermediate states $\ell_2, \ldots, \ell_{N-1}$. Recall from the definition of matrix multiplication that

$$\left(\boldsymbol{A}^2\right)_{ij} = \sum_{\ell} A_{i\ell}\, A_{\ell j}\,.$$

Using this identity repeatedly then gives

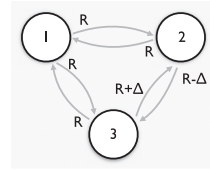
$$\begin{aligned} \sum_{\ell_2 \ldots \ell_{N-1}} A_{i\ell_2}\, A_{\ell_2\ell_3}\, \ldots\, A_{\ell_{N-1}j}\, A_{ji} &= \sum_{\ell_2 \ldots \ell_{N-1}} A_{j\ell_{N-1}}\, A_{\ell_{N-1}\ell_{N-2}}\, \ldots\, A_{\ell_2 i}\, A_{ij} \\ \left(\boldsymbol{A}^{N-1}\right)_{ij}\, A_{ji} &= \left(\boldsymbol{A}^{N-1}\right)_{ji}\, A_{ij} \\ p_i\, A_{ji} &= p_j\, A_{ij} \\ A_{ji}\, p_i &= A_{ij}\, p_j\,. \end{aligned}$$

To go from step two to three, take the limit $N \to \infty$ and use the result from (c). Thus, the matrix element only depends on the first (row) index and is independent of the second (column) index. Note that the $N \to \infty$ limit is justified because Kolmogorov's criterion holds for *every* path, including arbitrarily long sequences of states.

Kolmogorov's criterion immediately implies that for any network graph of transitions that can be represented as a *tree* (graph without cycles), the steady state must be an equilibrium, reversible state. The argument is simply that Kolmogorov's criterion is for a cycle of states. But for a tree, there are no cycles. Thus, at least one of the connecting transition probabilities must equal zero in both directions. That is, there must be at least one $A_{\ell_i\ell_j} = A_{\ell_j\ell_i} = 0$.

Then the criterion is *always* satisfied. Indeed, you need cycles to have the possibility of non-equilibrium steady states.

12.3 Kinetic proofreading. Many biological systems have error rates far below that predicted by the Boltzmann distribution of equilibrium thermodynamics. A simple *nonequilibrium* model can model this phenomenon. Consider the three-state, discrete-time Markov model depicted at right, with $\boldsymbol{A} = \begin{pmatrix} 1-2R & R & R \\ R & 1-2R-\Delta & R-\Delta \\ R & R+\Delta & 1-2R+\Delta \end{pmatrix}$.



a. Find the steady state of the Markov chain.
b. Using Kolmogorov's condition from Problem 12.2, show that the system is in equilibrium (reversible) if and only if $\Delta = 0$.
c. Show that choosing Δ controls the ratio $\frac{p_2}{p_1}$ to be in the range $(\frac{1}{3}, \frac{5}{3})$.
d. Find the nonequilibrium current $J(\Delta)$ circulating around the loop.
e. Compute the steady state when you convert the cycle into a linear chain by setting $A_{13} = A_{31} = 0$. For this chain, show that $\frac{p_2}{p_1}$ is independent of Δ.

The connection with kinetic proofreading is that adding a "useless" node 3 and creating a nonequilibrium cycle alters $\frac{p_2}{p_1}$ without changing A_{12} or A_{21}.

Solution.

a. The steady state is given by the eigenvector of $\boldsymbol{A}$ corresponding to the unit eigenvalue. Using *Mathematica*, we find

$$\boldsymbol{p} = \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} - \frac{2}{9}\,\frac{\Delta}{R} \\ \frac{1}{3} + \frac{2}{9}\,\frac{\Delta}{R} \end{pmatrix}.$$

b. For $\Delta = 0$, all states are equally likely ($p_i = \frac{1}{3}$), as is obvious from the transition diagram. For a three-state system, there is only one possible non-trivial Kolmogorov condition for equilibrium, which is

$$A_{32}\,A_{21}\,A_{13} = A_{23}\,A_{12}\,A_{31}\,,$$

which implies

$$(R+\Delta)\,R\,R = (R-\Delta)\,R\,R\,.$$

This relation obviously holds only when $\Delta = 0$. Note that two-state paths trivially satisfy the Kolmogorov condition, since they are simply

$$A_{\ell_1\ell_2}\,A_{\ell_2\ell_1} = A_{\ell_2\ell_1}\,A_{\ell_1\ell_2}\,.$$

The ratio $\frac{p_2}{p_1}$ is given by

$$\frac{p_2}{p_1} = 1 - \frac{2}{3}\,\frac{\Delta}{R}\,.$$

For $\Delta = 0$, we recover $\frac{p_2}{p_1} = 1$.

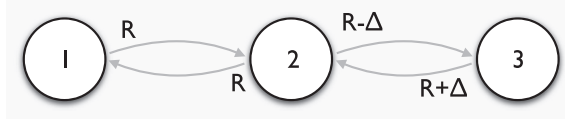
c. Since all elements of $\boldsymbol{A}$ must be positive (the elements are physically transition rates, which cannot be less than zero), $-R \le \Delta \le R$. At the upper bound, $\frac{p_2}{p_1} = \frac{1}{3}$; at the lower, $\frac{p_2}{p_1} = \frac{5}{3}$. (We could do better by altering the rates of $1 \to 3$ and $3 \to 1$ transitions or by not requiring $A_{23} + A_{32} = 2R$, as well.)

d. The current J_{21} in the clockwise direction ($1 \to 2 \to 3 \to 1$) is given by

$$J_{21} = A_{21}\, p_1 - A_{12}\, p_2 = +\tfrac{2}{9}\Delta\,.$$

Thus, the higher probability in p_1 relative to p_2 is accompanied by a current from 1 to 2. The direction of the current traces back to difference between the $2 \to 3$ transition probability, $R + \Delta$, and the $3 \to 2$ probability, $R - \Delta$.

e. Breaking the $1 \to 3$ and $3 \to 1$ transition creates a linear chain (see below),



and the transition matrix becomes

$$\boldsymbol{A}_{\text{linear}} = \begin{pmatrix} 1-R & R & 0 \\ R & 1-2R-\Delta & R-\Delta \\ 0 & R+\Delta & 1-R+\Delta \end{pmatrix}.$$

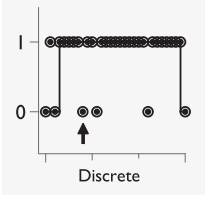
The single steady state is the eigenvector with unit eigenvalue:

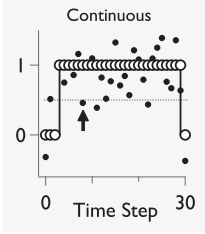
$$\boldsymbol{p}_{\text{linear}} = \begin{pmatrix} \frac{R-\Delta}{3R-\Delta} \\ \frac{R-\Delta}{3R-\Delta} \\ \frac{R+\Delta}{3R-\Delta} \end{pmatrix}.$$

Thus,

$$\frac{p_2}{p_1} = 1\,,$$

for all Δ in this linear-chain case. The result again illustrates the conclusion, proved using Kolmogorov's criterion above, that the steady state for a network with a tree graph must be in equilibrium and reversible.





12.4 HMM with continuous observations. Consider a two-state, discrete-time, symmetric Markov model with hopping probability a and states $x_k = \pm 1$. Let the observations $y_k = x_k + \xi_k$ be continuous, with $\xi_k \sim \mathcal{N}(0, \sigma^2)$.

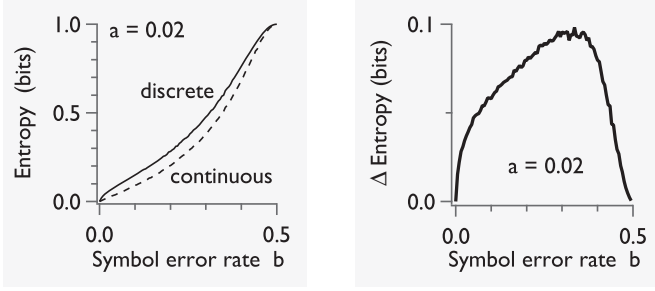
a. Write code to generate a hidden Markov sequence with both discrete and continuous observations. Match the noise variance ξ^2 to b, as in Eq. (12.7). Solve the filtering problem for continuous observations. Compare plots of a time series for $a = b = 0.1$. Plot the true state x_k, the observations (solid markers at left), and filter estimates (hollow markers). First, generate the continuous observations; then use them to form the discrete symbols via the sign operation.

b. Show that the filter performs slightly better using continuous observations. Hint: look at the vertical arrows.
c. For $0 < b < 0.5$, compare the entropy with that of the equivalent HMM having two observation symbols, defined as in Eq. (12.7). You should find something resembling the plots comparing filter and smoother in Section 12.2.1.

Solution.

a. For $b = 0.1$, we have $\sigma \approx 0.39$, to match the two in terms of error rates.
b. We solve $P(x_{k+1} = 0|y^k)\mathcal{N}(0, \sigma^2) = P(x_{k+1} = 1|y^k)\mathcal{N}(1, \sigma^2)$ for $y^k \equiv y^*$. The dividing value $y^* = 0.165627$.

   For discrete observations, the true state is 1 but the observation is 0. This fools the filter, which infers $\hat{x}_{92} = 0$. But in the continuous case, $y_{92} \approx 0.3$, which is below the midpoint, 0.5, between the two states. In this case, the prior "wins," and the filter estimates a 0 state, even though the observation is closer to 0 than to 1. The extra information in y_k helps. In more detail, show that if the prior probability to be in state 1 is 0.999937 (the value for the graph in question), then an observation $y^* \gtrsim 0.17$ will lead the filter to conclude that the most likely state is 1. That y^* is so much less than 0.5 reflects the strength of the prior.
c. You should find something like the plot below, at left for $a = 0.02$ and time series that are 10^5 long. The continuous observations lead to slightly lower entropy than do the discrete observations and thus to slightly greater certainty. As before, this difference goes to zero for $b \to 0$, 0.5 (very high and very low signal-to-noise ratios). As before, the maximum benefit is at intermediate values of the signal-to-noise ratio (right plot).



12.5 Confidence bound for state estimates.

a. For the filter, derive Eq. (12.11) and reproduce its associated plot.
b. For the smoother, show that the maximum confidence $q^* = \frac{1}{2}(1 + \frac{(1-a)(1-2b)}{\sqrt{a^2+(1-2a)(1-2b)^2}})$.

Solution.

a. We start from Eq. (12.10):

$$\underbrace{P(x_k = 1|y^k = 1)}_{p^*} = \frac{1}{Z_k} \underbrace{P(y_k = 1|x_k = 1)}_{1-b} \sum_{x_{k-1}} P(x_k = 1|x_{k-1}) P(x_{k-1}|y^{k-1} = 1) \,.$$

The terms in the sum are

$$P(x_k = 1|x_{k-1} = 1)P(x_{k-1} = 1|y^{k-1} = 1) = (1-a)p^*$$
$$P(x_k = 1|x_{k-1} = -1)P(x_{k-1} = -1|y^{k-1} = 1) = a(1-p^*) \,,$$

so that

$$p^* = \frac{1}{Z_k}(1-b)[(1-a)p^* + a(1-p^*)] \,.$$

Evaluating the other term in the partition function similarly gives

$$\begin{aligned} p^* &= \frac{(1-b)[(1-a)p^* + a(1-p^*)]}{(1-b)[(1-a)p^* + a(1-p^*)] + b[(1-a)(1-p^*) + ap^*]} \\ &= \frac{(1-b)(a + p^* - 2ap^*)}{(1-b)(a + p^* - 2ap^*) + b(1-a-p^*+2ap^*)} \\ &= \frac{(1-b)[1 + a + p^*(1-2a)]}{1 + p^*(1-2a)(1-2b)} \,. \end{aligned}$$

This gives a quadratic equation for p^*:

$$(1-2a)(1-2b)(p^*)^2 - [1-2b+a(4b-3)]p^* + a(b-1) = 0 \,,$$

whose relevant solution is

$$p^* = \frac{1-2b+a(4b-3) + \sqrt{a^2 + (1-2a)(1-2b)^2}}{2(1-2a)(1-2b)}$$

Setting $a = 0.2$ and $b = 0.3$ gives $p^* \approx 0.851629$. The confidence in a given state can go no higher than this value. The high chance ($b = 0.3$) that a symbol is wrong limits the confidence of an estimate, no matter what the observations.

For $b \to 0$, the confidence bound goes to 1. Thus, after observing a long series of +1 states, we will be sure that the true state is +1, as there is little chance the wrong symbol was received. The confidence limit then drops as b increases.

For $b = 0.5$, an observation gives no information at all about the underlying state and the confidence is accordingly $p^* = 0.5$: you may as well just flip a coin.

The case $b = 0.3$ is indicated by dashed lines.

b. For the smoother, we start from Eq. (12.12) and write

$$P(x_k = 1|y^N = 1) = P(x_k = 1|y^k = 1) \sum_{x_{k+1}} \frac{P(x_{k+1}|x_k = 1)\, P(x_{k+1}|y^N = 1)}{P(x_{k+1}|y^k = 1)} .$$

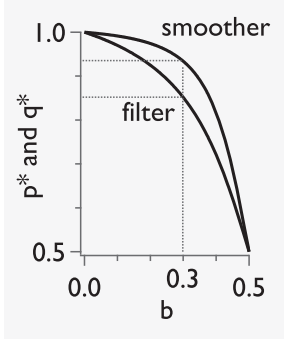
Substituting $q^* = P(x_k|y^N = 1)$ and $p^* = P(x_k = 1|y^k = 1)$ for any k gives

$$\begin{aligned} q^* &= p^* \sum_{x_{k+1}} \frac{P(x_{k+1}|x_k = 1)\, P(x_{k+1}|y^N = 1)}{P(x_{k+1}|y^k = 1)} \\ &= p^* \left(\frac{(1-a)q^*}{a + p^* - 2ap^*} + \frac{a(1-q^*)}{1 - a - p^* + 2ap^*} \right) , \end{aligned}$$

which we solve for q^*:

$$\begin{aligned} q^* &= \frac{p^*[p^* + a(1-2p^*)]}{1 - a(1-2p^*)^2 - 2p^*(1-p^*)} \\ &= \frac{1}{2} \left(1 + \frac{(1-a)(1-2b)}{\sqrt{a^2 + (1-2a)(1-2b)^2}} \right) \end{aligned}$$

where we have substituted for p^* in the last expression and simplified. Using a computer-algebra program is very helpful here! Plots of p^* for the filter and q^* for the smoother are given below, for $a = 0.2$, as a function of the symbol error probability b.



Again, we can look at limits. For $a = 0.5$, we have $q^* = 1 - b$, as before with p^*. Again, this corresponds to the idea that extra information beyond the current observation is useless. Thus, past and future information are equally useless, implying that we expect the same inferences whether based on filter or smoother.

For $a \to 0$, we can Taylor expand, to find

$$q^* = 1 - \frac{b(1-b)}{(1-2b)^2} a^2 + O(a^3) .$$

12.6 Phase transition in discord order parameter $\mathcal{D}$.

a. Using the argument given in Section 12.2.1, derive Eq. (12.15).

b. For the smoother case, let $y^{N\backslash k}$ refer to all observations except y_k. Explain why the condition to be imposed is now $P(x_k = 1|y_k = -1, y^{N\backslash k} = 1) = \frac{1}{2}$.
c. Show that $P(x_k = 1|y^{N\backslash k} = 1) = \frac{1}{Z} P(x_k = 1|y_{k+1}^N = 1)\, P(x_k = 1|y^{k-1} = 1)$.
d. Justify $P(x_k|y_{k+1}^N) = P(x_k|y^{k-1})$ and then derive Eq. (12.16).
e. Simulate the HMM and, for given a, scan b until $\mathcal{D} > 0.001$ in order to reproduce the numerical threshold data at left. Do for both filter and smoother.

Solution.

a. We start from Eq. (12.14) :

$$P(x_{k+1} = 1|y_{k+1} = -1, y^k = 1) = \frac{1}{2}\,.$$

From Bayes' theorem,

$$\begin{aligned} &P(x_{k+1} = 1|y_{k+1} = -1, y^k = 1) \\ &\quad = \frac{P(y_{k+1} = -1|x_{k+1} = 1, \cancel{y^k = 1})\, P(x_{k+1} = 1|y^k = 1)}{P(y_{k+1} = -1|y^k = 1)} \\ &\quad = \frac{P(y_{k+1} = -1|x_{k+1} = 1)\, P(x_{k+1} = 1|y^k = 1)}{\sum_{x_{k+1}} P(y_{k+1} = -1|x_{k+1})\, P(x_{k+1}|y^k = 1)} \\ &\quad = \frac{b[(1-a)p^* + a(1-p^*)]}{b[(1-a)p^* + a(1-p^*)] + (1-b)[ap^* + (1-a)(1-p^*)]}\,, \end{aligned}$$

where $p^* = P(x_k = 1|y^k = 1)$ is the maximum confidence for the filter estimate, a solution of Eq. (12.11). See Problem 12.5, too. We thus solve

$$\frac{b[(1-a)p^* + a(1-p^*)]}{b[(1-a)p^* + a(1-p^*)] + (1-b)[ap^* + (1-a)(1-p^*)]} = \frac{1}{2}\,.$$

where

$$p^* = \frac{1 - 2b + a(4b-3) + \sqrt{a^2 + (1-2a)(1-2b)^2}}{2(1-2a)(1-2b)}$$

From a symbolic-algebra program, we find that this equation reduces to

$$\frac{b(1 - a + \sqrt{a^2 + (1-2a)(1-2b)^2})}{(1-2b)(1 + a - \sqrt{a^2 + (1-2a)(1-2b)^2})} = \frac{1}{2}\,.$$

Squaring and simplifying gives

$$(2b-1)(b^2 - b + a) = 0\,,$$

which has solutions $b = \frac{1}{2}$ and $b = \frac{1}{2}(1 \pm \sqrt{1-4a})$. The relevant solution for the phase transition has $b < \frac{1}{2}$, which corresponds to the negative root, as described in Eq. (12.15). It is amazing that such a complicated expression simplifies so much!

b. For the smoother, the condition

$$P(x_k = 1|y_k = -1, y^{N\backslash k} = 1) = \frac{1}{2}\,,$$

is directly analogous to the condition for the filter,

$$P(x_k = 1|y_k = -1, y^{k-1} = 1) = \frac{1}{2}\,.$$

The difference is that we now assume that the future observations are also uniform. We then ask if the present observation contradicts both past and future, do we trust it or do we use the inference?

c. We use Bayes' theorem and the Markov condition:

$$\begin{aligned}
&P(x_k = 1|y^{N\backslash k} = 1) \\
&\quad = P(x_k = 1|y^{k-1} = 1, y_{k+1}^N = 1) \\
&\quad = \frac{1}{Z} P(y_{k+1}^N = 1|x_k = 1, \cancel{y^{k-1} = 1})\, P(x_k = 1|y^{k-1} = 1) \\
&\quad = \frac{1}{Z} P(x_k = 1|y_{k+1}^N = 1)\, \cancel{P(y_{k+1}^N = 1)}/\cancel{P(x_k = 1)}\, P(x_k = 1|y^{k-1} = 1) \\
&\quad = \frac{1}{Z} P(x_k = 1|y_{k+1}^N = 1)\, P(x_k = 1|y^{k-1} = 1)\,,
\end{aligned}$$

where we cancel $P(y_{k+1}^N = 1)$ and $P(x_k) = \frac{1}{2}$, as they do not depend on x_k and similar terms show up in the normalization coefficient Z. After cancellation, the normalization is, explicitly

$$Z = \sum_{x_k} P(x_k = 1|y_{k+1}^N = 1)\, P(x_k = 1|y^{k-1} = 1)$$

d. The condition, $P(x_k|y_{k+1}^N) = P(x_k|y^{k-1})$ is perhaps the trickiest to see. The idea is that the observations are just symbols. Thus, when evaluating the two conditional probabilities, we get exactly the same thing using future observations as with past. The sole difference is that $P(x_{k+1}|x_k) \to P(x_k|x_{k+1})$. But these are equal for a long time series that is in "equilibrium," as the Principle of Detailed Balance tells us:

$$P(x_{k+1}|x_k)\, P(x_k) = P(x_k|x_{k+1})\, P(x_{k+1})\,.$$

If the equilibrium unconditional probabilities $P(x_k) = P(x_{k+1}) = \frac{1}{2}$, then we conclude that $P(x_{k+1}|x_k) = P(x_k|x_{k+1})$.

Putting it all together, our equation is

$$\begin{aligned}
P(x_k = 1|y_k = -1, y^{N\backslash k} = 1) &= \frac{1}{Z} P(y_k = -1|x_k = 1)\, P(x_k = 1|y^{N\backslash k} = 1) \\
&= \frac{1}{Z} P(y_k = -1|x_k = 1)\, [P(x_k = 1|y^{k-1})]^2 = \frac{1}{2}\,.
\end{aligned}$$

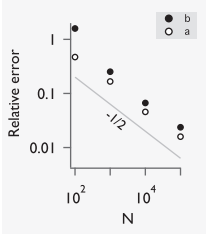
Using our earlier results for the filter, we have, for the explicit condition,

$$\frac{\frac{b(a+p^*-2ap^*)^2}{(a+p^*-2ap^*)^2+(1-a-p^*+2ap^*)^2}}{\frac{b(a+p^*-2ap^*)^2}{(a+p^*-2ap^*)^2+(1-a-p^*+2ap^*)^2}+\frac{(1-b)(1-a-p^*+2ap^*)^2}{(a+p^*-2ap^*)^2+(1-a-p^*+2ap^*)^2}} = \frac{1}{2},$$

where

$$p^* = \frac{1-2b+a(4b-3)+\sqrt{a^2+(1-2a)(1-2b)^2}}{2(1-2a)(1-2b)}$$

Again, an amazing simplification leads to the roots in Eq. (12.16).



12.7 Learning an HMM. Find the parameters a and b of a symmetric, two-state, two-symbol hidden Markov model. Write a program to generate an HMM time series of length N, given $a = 0.2$ and $b = 0.1$. Call a standard optimization program that can take the output series y^N and initial guesses for a and b and return estimates $\hat{a}$ and $\hat{b}$ based on minimizing the negative log likelihood function, Eq. (12.21). For given N, repeat enough times to estimate the mean and standard deviation of each parameter and then compute the relative error, vs. N. Use the true values of a and b as initial guesses for the optimization. Reproduce the plot at left.

Solution.

See book website for example code. With $N = 10^4$, the accuracy is $\approx 5\%$.

12.8 Gridworld. Code Example 12.2:

a. Create the 6×6 transition matrices $P(x'|x,u)$ for each of the four decisions $u = \{\text{N, E, S, and W}\}$. Each column should sum to one. If a move would leave gridworld, the system stays in its current state. Thus, for example, $P(x' = 1|x = 1, u = \text{N}) = 0.8 + 0.1 = 0.9$. You are in the NW corner trying to go N. You cannot go north, which adds 0.8 from the forward branch. You cannot go west, adding another 0.1 from the left branch. Include rules for the termination state, too.
b. Reproduce the tables of optimal utilities and policies with greater precision.
c. What happens if you always move forward and never go left, right, or back?

Solution.

a. The four transition matrices are

$$\underbrace{\begin{pmatrix} 0.9 & 0.8 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0.8 & 0 & 0.1 & 0 \\ 0 & 0 & 0.1 & 0 & 0 & 0.1 \\ 0.1 & 0 & 0 & 1 & 0.8 & 0 \\ 0 & 0.1 & 0 & 0 & 0.1 & 0.8 \\ 0 & 0 & 0.1 & 0 & 0 & 0.1 \end{pmatrix}}_{P(x'|x,\text{N})} \quad \underbrace{\begin{pmatrix} 0.1 & 0.1 & 0 & 0 & 0 & 0 \\ 0.1 & 0 & 0.1 & 0 & 0 & 0 \\ 0 & 0.1 & 0.1 & 0 & 0 & 0 \\ 0.8 & 0 & 0 & 1 & 0.1 & 0 \\ 0 & 0.8 & 0 & 0 & 0.8 & 0.1 \\ 0 & 0 & 0.8 & 0 & 0.1 & 0.9 \end{pmatrix}}_{P(x'|x,\text{E})}$$

$$\underbrace{\begin{pmatrix} 0.1 & 0 & 0 & 0 & 0 & 0 \\ 0.8 & 0.1 & 0 & 0 & 0.1 & 0 \\ 0 & 0.8 & 0.9 & 0 & 0 & 0.1 \\ 0.1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0.1 & 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0.1 & 0 & 0.8 & 0.9 \end{pmatrix}}_{P(x'|x,\mathrm{S})} \quad \underbrace{\begin{pmatrix} 0.9 & 0.1 & 0 & 0 & 0 & 0 \\ 0.1 & 0.8 & 0.1 & 0 & 0.8 & 0 \\ 0 & 0.1 & 0.9 & 0 & 0 & 0.8 \\ 0 & 0 & 0 & 1 & 0.1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.1 \\ 0 & 0 & 0 & 0 & 0.1 & 0.1 \end{pmatrix}}_{P(x'|x,\mathrm{W})}$$

b. The optimal utilities are

$$\underbrace{\begin{pmatrix} 0.590 & 10 \\ -1.099 & -8.594 \\ -1.225 & -1.249 \end{pmatrix}}_{\gamma=0.2} \quad \underbrace{\begin{pmatrix} 7.246 & 10 \\ 4.372 & -2.644 \\ 2.415 & 0.550 \end{pmatrix}}_{\gamma=0.9}$$

The algorithm converges much more quickly at small γ (0.2) than for larger values (0.9). Remember that we need to convert these 2×3 matrices to 6-dim. vectors for the algorithm.

c. If you always move forward, the problem becomes deterministic and you always move in the least-bad direction.

13 Quantum Control

Problems

13.1 Most general unitary operator on a qubit. Show that

a. the most general two-dimensional Hermitian operator can be written $H = c_0\mathbb{I} + c_1\sigma_x + c_2\sigma_y + c_3\sigma_z$, where the c_i are real and the $\sigma_{x,y,z}$ are the Pauli matrices;
b. the associated unitary transformation can be written $U = \exp -(\mathrm{i}Ht/\hbar) = \mathrm{e}^{\mathrm{i}\varphi\,\hat{\boldsymbol{m}}\cdot\boldsymbol{\sigma}}$;
c. and $\mathrm{e}^{\mathrm{i}\varphi\,\hat{\boldsymbol{m}}\cdot\boldsymbol{\sigma}} = \cos\varphi\,\mathbb{I} + \mathrm{i}\sin\varphi\,(\hat{\boldsymbol{m}}\cdot\boldsymbol{\sigma})$, where $\hat{\boldsymbol{m}}$ is a 3d unit vector and $\boldsymbol{\sigma}$ is the 3-vector of Pauli matrices. Hint: show that $(\hat{\boldsymbol{m}}\cdot\boldsymbol{\sigma})^2 = \mathbb{I}$, the 2×2 identity matrix.

Solution.

a. The most general 2×2 Hermitian matrix must be of the form

$$H = \begin{pmatrix} \alpha & \beta \\ \beta^* & \gamma \end{pmatrix},$$

where α, β, and γ are three arbitrary complex numbers. Because H is Hermitian, it satisfies $H = H^\dagger$, which implies that the off-diagonal elements must be β and β^* and that α and γ are real. Next, we expand

$$\begin{aligned} c_0\mathbb{I} + c_1\sigma_x + c_2\sigma_y + c_3\sigma_z &= \begin{pmatrix} c_0 & 0 \\ 0 & c_0 \end{pmatrix} + \begin{pmatrix} 0 & c_1 \\ c_1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & -\mathrm{i}c_2 \\ \mathrm{i}c_2 & 0 \end{pmatrix} + \begin{pmatrix} c_3 & 0 \\ 0 & -c_3 \end{pmatrix} \\ &= \begin{pmatrix} c_0 + c_3 & c_1 - \mathrm{i}c_2 \\ c_1 + \mathrm{i}c_2 & c_0 - c_3 \end{pmatrix}. \end{aligned}$$

Matching terms with the general expression for H shows that the two forms are equivalent.

b. From Eq. (13.6) and part (a),

$$\begin{aligned} U = \mathrm{e}^{-\mathrm{i}Ht/\hbar} &= \mathrm{e}^{-\mathrm{i}c_0\mathbb{I} + c_1\sigma_x + c_2\sigma_y + c_3\sigma_z} \\ &= \mathrm{e}^{-\mathrm{i}c_0 t/\hbar}\,\mathbb{I}\,\mathrm{e}^{-\mathrm{i}t/\hbar(c_1\sigma_x + c_2\sigma_y + c_3\sigma_z)} \\ &= \mathrm{e}^{-\mathrm{i}c_0 t/\hbar}\,\mathrm{e}^{-\mathrm{i}t/\hbar(c_1\sigma_x + c_2\sigma_y + c_3\sigma_z)}. \end{aligned}$$

The $e^{-ic_0t/\hbar}\,\mathbb{I}$ term separates from the rest of the expression because the identity matrix commutes with all other matrices. We can set the $e^{-ic_0t/\hbar}$ term to 1 because it is just a global phase factor. Thus,

$$U = e^{-it/\hbar(c_1\sigma_x+c_2\sigma_y+c_3\sigma_z)} \equiv e^{i\varphi\,\hat{\boldsymbol{m}}\cdot\boldsymbol{\sigma}} = \cos\varphi\,\mathbb{I} + i\sin\varphi\;(\hat{\boldsymbol{m}}\cdot\boldsymbol{\sigma})\,,$$

c. First, we show that $(\hat{\boldsymbol{m}}\cdot\boldsymbol{\sigma})^2 = \mathbb{I}$, the 2×2 identity matrix.

$$\begin{aligned}(\hat{\boldsymbol{m}}\cdot\boldsymbol{\sigma})^2 &= (m_x\sigma_x + m_y\sigma_y + m_z\sigma_z)^2\\ &= m_x^2\sigma_x^2 + m_y^2\sigma_y^2 + m_z^2\sigma_z^2\\ &\quad + m_xm_y\underbrace{(\sigma_x\sigma_y+\sigma_y\sigma_x)}_{0} + m_ym_z\underbrace{(\sigma_y\sigma_z+\sigma_z\sigma_y)}_{0} + m_zm_x\underbrace{(\sigma_z\sigma_x+\sigma_x\sigma_z)}_{0}\\ &= (m_x^2+m_y^2+m_z^2)\mathbb{I} + \mathbf{0} = \mathbb{I}\,,\end{aligned}$$

where $\sigma_x^2 = \sigma_y^2 = \sigma_z^2 = \mathbb{I}$ (as can be verified directly) and $\sigma_x\sigma_y = -\sigma_y\sigma_x$ (and similarly for the cyclic permutations). Finally, $\hat{\boldsymbol{m}}$ is a unit vector. Next, we expand the matrix exponential in a Taylor series:

$$\begin{aligned}e^{i\varphi\,\hat{\boldsymbol{m}}\cdot\boldsymbol{\sigma}} &= \mathbb{I} + i\varphi\,\hat{\boldsymbol{m}}\cdot\boldsymbol{\sigma} - \tfrac{1}{2!}\varphi^2\,(\hat{\boldsymbol{m}}\cdot\boldsymbol{\sigma})^2 - \tfrac{i}{3!}\varphi^3\,(\hat{\boldsymbol{m}}\cdot\boldsymbol{\sigma})^3 + \tfrac{1}{4!}\varphi^4\,(\hat{\boldsymbol{m}}\cdot\boldsymbol{\sigma})^4 + \cdots\\ &= \mathbb{I} + i\varphi\,\hat{\boldsymbol{m}}\cdot\boldsymbol{\sigma} - \tfrac{1}{2!}\varphi^2\mathbb{I} - \tfrac{i}{3!}\varphi^3\,\hat{\boldsymbol{m}}\cdot\boldsymbol{\sigma} + \tfrac{1}{4!}\varphi^4\mathbb{I} + \cdots\\ &= \Big(1 - \tfrac{1}{2!}\varphi^2 + \tfrac{1}{4!}\varphi^4 - \cdots\Big)\mathbb{I} + i\Big(\varphi - \tfrac{1}{3!}\varphi^3 + \cdots\Big)\,(\hat{\boldsymbol{m}}\cdot\boldsymbol{\sigma})\\ &= \cos\varphi\,\mathbb{I} + i\sin\varphi\;(\hat{\boldsymbol{m}}\cdot\boldsymbol{\sigma})\,.\end{aligned}$$

which is an explicit form for U that is a linear combination of the four matrices $\{\mathbb{I}, \sigma_x, \sigma_y, \sigma_z\}$.

13.2 Feedforward control of a qubit.

a. Show that if $|\alpha|^2 + |\beta|^2 = 1$, then $U = \left(\begin{smallmatrix}\beta & -\alpha\\ \alpha^* & \beta^*\end{smallmatrix}\right)$ is unitary and transforms the normalized state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ to the target state $|1\rangle$.
b. To interpret U as a rotation of $|\psi\rangle$ on the Bloch sphere, show that $R_z(\varphi) \equiv e^{-i\frac{\varphi}{2}\sigma_z}$ rotates $|\psi\rangle$ by φ about the z-axis on the Bloch sphere.
c. For a state $|\psi\rangle$ that lies in the x-z plane, show that $R_y(\varphi) \equiv e^{-i\frac{\varphi}{2}\sigma_y}$ rotates $|\psi\rangle$ by φ about the y-axis on the Bloch sphere.
d. Using (b) and (c), show that U can be decomposed into a rotation by $-\phi$ about the z-axis followed by a rotation $\pi - \theta$ about the y-axis. Write the two rotation matrices explicitly and show that the resulting U has the form supposed in (a).

Solution.

a. First, we show that U is unitary:

$$U = \begin{pmatrix}\beta & -\alpha\\ \alpha^* & \beta^*\end{pmatrix} \quad\Longrightarrow\quad U^\dagger = \begin{pmatrix}\beta^* & \alpha\\ -\alpha^* & \beta\end{pmatrix},$$

and

$$UU^\dagger = \begin{pmatrix} \beta & -\alpha \\ \alpha^* & \beta^* \end{pmatrix} \begin{pmatrix} \beta^* & \alpha \\ -\alpha^* & \beta \end{pmatrix} = \begin{pmatrix} |\beta|^2 + |\alpha|^2 & 0 \\ 0 & |\alpha|^2 + |\beta|^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

where $\alpha\alpha^* + \beta\beta^* = |\alpha|^2 + |\beta|^2 = 1$. Thus, U is unitary. Next, we apply it to $|\psi\rangle$:

$$U|\psi\rangle = \begin{pmatrix} \beta & -\alpha \\ \alpha^* & \beta^* \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 0 \\ \alpha\alpha^* + \beta\beta^* \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Thus, U controls the initial state as desired.

b. We start by constructing

$$\begin{aligned} R_z(\varphi) = \mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}\sigma_z} &= \cos\tfrac{\varphi}{2}\mathbb{I} - \mathrm{i}\sin\tfrac{\varphi}{2}\sigma_z \\ &= \begin{pmatrix} \cos\frac{\varphi}{2} - \mathrm{i}\sin\frac{\varphi}{2} & 0 \\ 0 & \cos\frac{\varphi}{2} + \mathrm{i}\sin\frac{\varphi}{2} \end{pmatrix} = \begin{pmatrix} \mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}} & 0 \\ 0 & \mathrm{e}^{\mathrm{i}\frac{\varphi}{2}} \end{pmatrix}. \end{aligned}$$

Next, we apply this operator to $|\psi\rangle$:

$$R_z(\varphi)\,|\psi\rangle = \begin{pmatrix} \mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}} & 0 \\ 0 & \mathrm{e}^{\mathrm{i}\frac{\varphi}{2}} \end{pmatrix} \begin{pmatrix} \cos\frac{\theta}{2} \\ \mathrm{e}^{\mathrm{i}\phi}\sin\frac{\theta}{2} \end{pmatrix} = \mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}} \begin{pmatrix} \cos\frac{\theta}{2} \\ \mathrm{e}^{\mathrm{i}(\phi+\varphi)}\sin\frac{\theta}{2} \end{pmatrix}.$$

In the last step, we pull out an overall phase factor, as we are free to do. Thus R_z makes $\phi \to (\phi + \varphi)$ in $|\psi\rangle$, which corresponds to a rotation by φ about the z-axis on the Bloch sphere, as claimed.

c. If $|\psi\rangle$ lies in the x-z plane, then $\phi = 0$ and

$$|\psi\rangle = \begin{pmatrix} \cos\frac{\theta}{2} \\ \sin\frac{\theta}{2} \end{pmatrix}.$$

A rotation $R_y(\varphi)$ about the y-axis has a matrix representation

$$R_y(\varphi) = \mathrm{e}^{-\mathrm{i}\frac{\varphi}{2}\sigma_y} = \cos\tfrac{\varphi}{2}\mathbb{I} - \mathrm{i}\sin\tfrac{\varphi}{2}\sigma_y = \begin{pmatrix} \cos\frac{\varphi}{2} & -\sin\frac{\varphi}{2} \\ \sin\frac{\varphi}{2} & \cos\frac{\varphi}{2} \end{pmatrix}.$$

Applying this to $|\psi\rangle$ in the x-z plane gives,

$$R_y(\varphi)\,|\psi\rangle = \begin{pmatrix} \cos\frac{\varphi}{2} & -\sin\frac{\varphi}{2} \\ \sin\frac{\varphi}{2} & \cos\frac{\varphi}{2} \end{pmatrix} \begin{pmatrix} \cos\frac{\theta}{2} \\ \sin\frac{\theta}{2} \end{pmatrix} = \begin{pmatrix} \cos\frac{(\theta+\varphi)}{2} \\ \sin\frac{(\theta+\varphi)}{2} \end{pmatrix},$$

which is indeed a rotation by φ about the y-axis.

d. We put these together to understand how U maps a general state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ to the target state $|\psi\rangle$. We first apply $R_z(-\phi)$, which rotates the state from ϕ to 0, which is in the x-z plane. It is not at an angle θ from the north pole, $|0\rangle$, and we want to rotate it to the south pole, $|1\rangle$, which is obviously done by rotating about y by an angle $(\pi - \theta)$.

The last step is to check that this gives a net U of the claimed form. Remembering that operators act first from the right and progress leftwards, we have

$$U = R_y(\pi-\theta)\,R_z(-\phi) = \begin{pmatrix} \cos\frac{(\pi-\theta)}{2} & -\sin\frac{(\pi-\theta)}{2} \\ \sin\frac{(\pi-\theta)}{2} & \cos\frac{(\pi-\theta)}{2} \end{pmatrix} \begin{pmatrix} e^{i\frac{\phi}{2}} & 0 \\ 0 & e^{-i\frac{\phi}{2}} \end{pmatrix}$$

$$= \begin{pmatrix} \sin\frac{\theta}{2} & -\cos\frac{\theta}{2} \\ \cos\frac{\theta}{2} & \sin\frac{\theta}{2} \end{pmatrix} \begin{pmatrix} e^{i\frac{\phi}{2}} & 0 \\ 0 & e^{-i\frac{\phi}{2}} \end{pmatrix}$$

$$= \begin{pmatrix} \sin\frac{\theta}{2}\,e^{i\frac{\phi}{2}} & -\cos\frac{\theta}{2}\,e^{-i\frac{\phi}{2}} \\ \cos\frac{\theta}{2}\,e^{i\frac{\phi}{2}} & \sin\frac{\theta}{2}\,e^{-i\frac{\phi}{2}} \end{pmatrix} \equiv \begin{pmatrix} \beta & -\alpha \\ \alpha^* & \beta^* \end{pmatrix}$$

where

$$\alpha \equiv \cos\tfrac{\theta}{2}\left(e^{-i\frac{\phi}{2}}\right), \qquad \beta \equiv \sin\tfrac{\theta}{2}\left(e^{i\frac{\phi}{2}}\right).$$

In other words, U has the form claimed in the problem. We could also express U as a *single* rotation on the Bloch sphere, but the decomposition into elementary rotations about the z- and y-axes given here seems more intuitive.

13.3 Spin-$\frac{1}{2}$ particle in a constant external field. Consider a spin-$\frac{1}{2}$ particle in a constant field $B_0\,\hat{z}$, whose normalized state at $t = 0$ is given by $|\psi(0)\rangle = \alpha|0\rangle + \beta|1\rangle$.

a. Solve the Schrödinger equation to find $|\psi(t)\rangle$.
b. Show that $\langle\mu_x\rangle$ and $\langle\mu_y\rangle$ precess about $\hat{z}$ at a frequency $\omega_0 = \gamma B_0$.
c. Show that the field does no work on the particle (expected energy is constant).

Solution.

a. In units where $\hbar = 1$, the solution to the Schrödinger Equation is

$$|\psi(t)\rangle = e^{-iHt}\,|\psi(0)\rangle\,.$$

For a constant magnetic field along the z-axis,

$$H = -\tfrac{1}{2}\gamma B_0\sigma_z = -\tfrac{1}{2}\omega_0\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Since H is diagonal and since $|\psi(0)\rangle = \alpha|0\rangle + \beta|1\rangle = \binom{\alpha}{\beta}$, we have

$$|\psi(t)\rangle = \underbrace{\begin{pmatrix} e^{i\omega_0 t/2} & 0 \\ 0 & e^{-i\omega_0 t/2} \end{pmatrix}}_{e^{-iHt}} \underbrace{\begin{pmatrix} \alpha \\ \beta \end{pmatrix}}_{|\psi(0)\rangle}$$

$$= \begin{pmatrix} \alpha\,e^{i\omega_0 t/2} \\ \beta\,e^{-i\omega_0 t/2} \end{pmatrix} \equiv \begin{pmatrix} \alpha(t) \\ \beta(t) \end{pmatrix} = \alpha(t)\,|0\rangle + \beta(t)\,|1\rangle\,.$$

Note that $|\psi(t)\rangle$ stays normalized. That is, $\langle\psi(t)\,|\,\psi(t)\rangle = 1$, since $|\alpha(t)|^2 + |\beta(t)|^2 = |\alpha^2| + |\beta^2| = 1$.

b. Since only the relative phase of α and β matters, we consider the initial condition variable α to be real and let $\beta \to \beta e^{i\theta}$, with β now taken as real and θ denoting the relative phase. We then evaluate the expectation values of the magnetic moment $\mu_x = \frac{1}{2}\gamma\boldsymbol{\sigma}_x$:

$$\begin{aligned}
\langle\mu_x\rangle &= \langle\psi(t)|\,\tfrac{1}{2}\gamma\boldsymbol{\sigma}_x\,|\psi(t)\rangle \\
&= \tfrac{1}{2}\gamma\begin{pmatrix}\alpha^*(t) & \beta^*(t)\end{pmatrix}\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}\begin{pmatrix}\alpha(t)\\ \beta(t)\end{pmatrix} \\
&= \tfrac{1}{2}\gamma\left[\alpha^*(t)\beta(t) + \beta^*(t)\,\alpha(t)\right] \\
&= \tfrac{1}{2}\gamma\alpha\beta\left[e^{-i(\omega_0 t-\theta)} + e^{+i(\omega_0 t-\theta)}\right] \\
&= \gamma\alpha\beta\,\cos(\omega_0 t - \theta)\,.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\langle\mu_y\rangle &= \langle\psi(t)|\,\tfrac{1}{2}\gamma\boldsymbol{\sigma}_y\,|\psi(t)\rangle \\
&= \tfrac{1}{2}\gamma\begin{pmatrix}\alpha^*(t) & \beta^*(t)\end{pmatrix}\begin{pmatrix}0 & -i\\ i & 0\end{pmatrix}\begin{pmatrix}\alpha(t)\\ \beta(t)\end{pmatrix} \\
&= \tfrac{i}{2}\gamma\left[-\alpha^*(t)\beta(t) + \beta^*(t)\,\alpha(t)\right] \\
&= \tfrac{i}{2}\gamma\alpha\beta\left[-e^{-i(\omega_0 t-\theta)} + e^{(i\omega_0 t-\theta)}\right] \\
&= -\gamma\alpha\beta\,\sin(\omega_0 t - \theta)\,.
\end{aligned}$$

The two components thus trace out a circle of radius $\gamma\alpha\beta$ that rotates at the Larmor frequency, $\omega_0 = \gamma B_0$. The classical picture is a unit vector at an angle with respect to the field axis ($\hat{z}$) and precessing at frequency ω_0 about the $\hat{z}$ axis. Note that if we define the x-axis to be along $\langle\mu_x\rangle$ at $t = 0$, then the phase $\theta = 0$.

c. We compute the energy:

$$\begin{aligned}
E = \langle H\rangle &= \langle\psi(t)|\,\tfrac{1}{2}\omega_0\boldsymbol{\sigma}_z\,|\psi(t)\rangle \\
&= \tfrac{1}{2}\omega_0\begin{pmatrix}\alpha^*(t) & \beta^*(t)\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\begin{pmatrix}\alpha(t)\\ \beta(t)\end{pmatrix} \\
&= \tfrac{1}{2}\omega_0\left[|\alpha(t)|^2 - |\beta(t)|^2\right] \\
&= \tfrac{1}{2}\omega_0\left(\alpha^2 - \beta^2\right) \\
&= \tfrac{1}{2}\omega_0\left(2\alpha^2 - 1\right)\,.
\end{aligned}$$

So, as $0 < \alpha < 1$, the average energy ranges from $-\frac{1}{2}\hbar\omega_0$ to $+\frac{1}{2}\hbar\omega_0$ and is constant.

13.4 Spin-$\frac{1}{2}$ particle in a rotating external field. Consider a spin-$\frac{1}{2}$ particle in a time-dependent field consisting of a stationary $\hat{z}$ component B_0 and a rotating horizontal component $\boldsymbol{B}_1(t) = B_1[\cos\omega t\,\hat{x} + \sin\omega t\,\hat{y}]$. Define $\omega_1 = \gamma B_1$.

a. Find the Hamiltonian in a frame that rotates about $\hat{z}$ at frequency ω by applying a rotation of angle ωt to the Hamiltonian.

b. Write the Schrödinger Equation in coordinates that rotate with phase $-\omega t$ and thereby show that the field becomes static, with $\gamma \boldsymbol{B}_{\rm eff} = -\frac{1}{2}(\gamma B_0 + \omega)\,\sigma_z + \frac{1}{2}\gamma B_1 \sigma_x$.

c. Using (a), solve the Schrödinger equation in the rotating frame and transform back to the original frame of reference to get $|\psi(t)\rangle$. With $|\psi(0)\rangle = |0\rangle$, show

$$\langle 1|\psi(t)\rangle = \mathrm{e}^{\mathrm{i}\omega t/2}\left[\frac{\mathrm{i}\omega_1}{a}\sin\left(\tfrac{1}{2}at\right)\right], \qquad a \equiv \sqrt{(\omega_0+\omega)^2+\omega_1^2}\,.$$

d. Show that at resonance, $\omega = -\omega_0$, the average energy $E(t) = -\frac{1}{2}\hbar\omega_0 \cos\omega_1 t$. The rotating field thus pumps energy in and out of the system periodically.

Solution.

a. In units where $\hbar = 1$, the Hamiltonian $H = -\boldsymbol{\mu}\cdot\boldsymbol{B} = -\frac{1}{2}\gamma\boldsymbol{B}\cdot\boldsymbol{\sigma}$. At time $t = 0$, the $\boldsymbol{B}$ vector lies in the z-x plane, and

$$H(0) = -\tfrac{1}{2}\gamma\,(B_0\,\sigma_z + B_1\,\sigma_x)\ .$$

In coordinates rotating about $\hat{\boldsymbol{z}}$ at frequency ω, we have $H(t) = U(t)\,H(0)\,U^\dagger(t)$, with $U(t) = \mathrm{e}^{-\mathrm{i}\omega t S_z} = \mathrm{e}^{-\mathrm{i}\omega t\sigma_z/2}$. The spin operator $S_z = \frac{1}{2}\hbar\sigma_z = \frac{1}{2}\sigma_z$, with $\hbar = 1$. Since $[U(t), \sigma_z] = 0$,

$$\begin{aligned} H(t) &= -\tfrac{1}{2}\gamma\left(B_0\,U\sigma_z U^\dagger + B_1\ \mathrm{e}^{-\mathrm{i}\omega t\sigma_z/2}\ \sigma_x\ \mathrm{e}^{+i\omega t\sigma_z/2}\right) \\ &= -\tfrac{1}{2}\left(\omega_0\,\sigma_z + \omega_1\ \mathrm{e}^{-\mathrm{i}\omega t\sigma_z/2}\ \sigma_x\ \mathrm{e}^{+\mathrm{i}\omega t\sigma_z/2}\right), \end{aligned}$$

where $\omega_0 = \gamma B_0$ defines the Larmor frequency and $\omega_1 = \gamma B_1$ the Rabi frequency.

b. We transform the Schrödinger Equation into coordinates rotating at $-\omega t$ by multiplying both sides of $\mathrm{i}\partial_t|\psi\rangle = H|\psi\rangle$ by $U^\dagger(t)$. Thus,

$$iU^\dagger(t)\,\partial_t|\psi(t)\rangle = U^\dagger(t)\,\underbrace{U(t)\,H(0)\,U^\dagger(t)}_{H(t)}\,|\psi(t)\rangle = H(0)\,|\Phi(t)\rangle\,,$$

where $U^\dagger|\psi\rangle \equiv |\Phi\rangle$ and $UU^\dagger = \mathbb{I}$. Taking a time derivative of the definition gives

$$\partial_t|\Phi\rangle = \partial_t\left(U^\dagger|\psi\rangle\right) = \tfrac{\mathrm{i}}{2}\omega\,\sigma_z\,U^\dagger|\psi\rangle + U^\dagger\partial_t|\psi\rangle\,,$$

so that $\mathrm{i}U^\dagger\,\partial_t|\psi\rangle = \mathrm{i}\partial_t|\Phi\rangle + \frac{1}{2}\omega\,\sigma_z\,|\Phi\rangle$ and

$$i\partial_t|\Phi\rangle = [H(0) - \tfrac{1}{2}\omega\sigma_z]\,|\Phi\rangle \equiv H_{\rm eff}\,|\Phi\rangle\,.$$

Then $H(0) = -\frac{1}{2}(\omega_0\sigma_z + \omega_1\sigma_x)$ is transformed to

$$H_{\rm eff}(t) = -\tfrac{1}{2}[(\omega_0+\omega)\sigma_z + \omega_1\sigma_x]\,,$$

which has no time dependence (in the rotating frame) and implies an effective static field via $H_{\rm eff} = -\frac{1}{2}\gamma\,(\boldsymbol{B}\cdot\boldsymbol{\sigma})$.

c. In order to be able to use the identity in Part (a) directly, we rewrite

$$H_{\text{eff}}(t) = -\tfrac{1}{2}[(\omega_0+\omega)\sigma_z + \omega_1\sigma_x] \equiv -\tfrac{1}{2}a\,(\hat{\boldsymbol{n}}\cdot\boldsymbol{\sigma})\,,$$

where

$$\hat{\boldsymbol{n}}\cdot\boldsymbol{\sigma} = \frac{1}{a}\begin{pmatrix}\omega_1\\0\\\omega_0+\omega\end{pmatrix}\cdot\begin{pmatrix}\sigma_x\\\sigma_y\\\sigma_z\end{pmatrix}, \qquad a \equiv \sqrt{(\omega_0+\omega)^2+\omega_1^2}$$

Since $\hat{\boldsymbol{n}}$ is a unit vector in the xz plane, $(\hat{\boldsymbol{n}}\cdot\boldsymbol{\sigma}) = \mathbb{I}$.

Now we can can solve the Schrödinger Equation in the counter-rotating frame:

$$|\Phi(t)\rangle = \mathrm{e}^{-\mathrm{i}H_{\text{eff}}(t)t}\,|\Phi(0)\rangle\,,$$

with $|\Phi(0)\rangle = U^\dagger(0)\,|\psi(0)\rangle = |0\rangle$.

From the identity proved in Part (a),

$$\begin{aligned}\mathrm{e}^{-\mathrm{i}tH_{\text{eff}}(t)} = \mathrm{e}^{\frac{\mathrm{i}t}{2}a\,(\hat{\boldsymbol{n}}\cdot\boldsymbol{\sigma})} &= \cos\left(\tfrac{1}{2}at\right)\mathbb{I} + \mathrm{i}\sin\left(\tfrac{1}{2}at\right)(\hat{\boldsymbol{n}}\cdot\boldsymbol{\sigma})\\ &= \begin{pmatrix}\cos\left(\frac{1}{2}at\right)+\frac{\mathrm{i}(\omega_0+\omega)}{a}\sin\left(\frac{1}{2}at\right) & \frac{\mathrm{i}\omega_1}{a}\sin\left(\frac{1}{2}at\right)\\ \frac{\mathrm{i}\omega_1}{a}\sin\left(\frac{1}{2}at\right) & \cos\left(\frac{1}{2}at\right)-\frac{\mathrm{i}(\omega_0+\omega)}{a}\sin\left(\frac{1}{2}at\right)\end{pmatrix}.\end{aligned}$$

With the given initial condition $|\Phi(0)\rangle = |0\rangle$,

$$\begin{aligned}|\Phi(t)\rangle &= \begin{pmatrix}\cos\left(\frac{1}{2}at\right)+\frac{\mathrm{i}(\omega_0+\omega)}{a}\sin\left(\frac{1}{2}at\right) & \frac{\mathrm{i}\omega_1}{a}\sin\left(\frac{1}{2}at\right)\\ \frac{\mathrm{i}\omega_1}{a}\sin\left(\frac{1}{2}at\right) & \cos\left(\frac{1}{2}at\right)-\frac{\mathrm{i}(\omega_0+\omega)}{a}\sin\left(\frac{1}{2}at\right)\end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix}\\ &= \begin{pmatrix}\cos\left(\frac{1}{2}at\right)+\frac{\mathrm{i}(\omega_0+\omega)}{a}\sin\left(\frac{1}{2}at\right)\\ \frac{\mathrm{i}\omega_1}{a}\sin\left(\frac{1}{2}at\right)\end{pmatrix}.\end{aligned}$$

Finally, we transform back to find

$$\begin{aligned}|\psi(t)\rangle = \mathrm{e}^{-\mathrm{i}\omega t\sigma_z/2}\,|\Phi(t)\rangle &= \begin{pmatrix}\mathrm{e}^{-\mathrm{i}\omega t/2} & 0\\ 0 & \mathrm{e}^{+i\omega t/2}\end{pmatrix}|\Phi(t)\rangle\\ &= \mathrm{e}^{-\mathrm{i}\omega t/2}\left[\cos\left(\tfrac{1}{2}at\right)+\frac{\mathrm{i}(\omega_0+\omega)}{a}\sin\left(\tfrac{1}{2}at\right)\right]|0\rangle + \mathrm{e}^{\mathrm{i}\omega t/2}\left[\frac{\mathrm{i}\omega_1}{a}\sin\left(\tfrac{1}{2}at\right)\right]|1\rangle\,.\end{aligned}$$

Notice that the initial condition is correct: $|\psi(0)\rangle = |0\rangle$.

d. To save space, we write $|\psi(t)\rangle = \alpha(t)|0\rangle + \beta(t)|1\rangle$. At resonance, $\omega = -\omega_0$ and $a = \omega_1$. Thus,

$$\alpha(t) = \mathrm{e}^{\mathrm{i}\omega_0 t/2}\cos\left(\tfrac{1}{2}\omega_1 t\right), \qquad \beta(t) = \mathrm{i}\,\mathrm{e}^{-\mathrm{i}\omega_0 t/2}\sin\left(\tfrac{1}{2}\omega_1 t\right),$$

and the Hamiltonian $H(t)$ becomes

$$H(t) = -\tfrac{1}{2}\left(\omega_0\,\sigma_z + \omega_1\,\mathrm{e}^{\mathrm{i}\omega_0 t\sigma_z/2}\,\sigma_x\,\mathrm{e}^{-\mathrm{i}\omega_0 t\sigma_z/2}\right).$$

The operators in the ω_1 term can be written

$$\begin{pmatrix} e^{i\omega t/2} & 0 \\ 0 & e^{-i\omega t/2} \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} e^{-i\omega_0 t/2} & 0 \\ 0 & e^{i\omega t/2} \end{pmatrix} = \begin{pmatrix} 0 & e^{i\omega t} \\ e^{-i\omega t} & 0 \end{pmatrix},$$

which implies that the Hamiltonian is

$$H(t) = -\tfrac{1}{2} \begin{pmatrix} \omega_0 & \omega_1 e^{i\omega t} \\ \omega_1 e^{-i\omega t} & -\omega_0 \end{pmatrix}.$$

Then

$$\begin{aligned} E(t) = \langle\psi(t)|\, H(t)\, |\psi(t)\rangle &= -\tfrac{1}{2} \begin{pmatrix} \alpha^* & \beta^* \end{pmatrix} \begin{pmatrix} \omega_0 & \omega_1 e^{i\omega t} \\ \omega_1 e^{-i\omega t} & -\omega_0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \\ &= -\tfrac{1}{2} \left[\omega_0 \left(|\alpha|^2 - |\beta|^2 \right) + 2\omega_1 \operatorname{Re} \left(\cancelto{0}{e^{i\omega_0 t} \alpha^* \beta} \right) \right] \\ &= -\tfrac{1}{2} \omega_0 \cos \omega_1 t \,, \end{aligned}$$

where $|\alpha|^2 - |\beta|^2 = \cos^2\left(\frac{1}{2}\omega_1 t\right) - \sin^2\left(\frac{1}{2}\omega_1 t\right) = \cos \omega_1 t$. The second term vanishes because the product is purely imaginary. Thus, the average energy $E(t)$ oscillates at the Rabi frequency ω_1 from $-\frac{1}{2}\hbar\omega_0$ to $+\frac{1}{2}\hbar\omega_0$. The classical picture is that the spin flips up and down along the z-axis.

13.5 Optimal control of a two-level system.

a. Derive Euler–Lagrange equations for $|\psi\rangle$, $|\lambda\rangle$, u_x, and u_y. Show that $|\psi\rangle$ and $|\lambda\rangle$ each obey Schrödinger equations and that $u_{\{x,y\}} = (1/\eta)\ \mathrm{Re}\ \langle\lambda|\sigma_{\{x,y\}}|\psi\rangle$.
b. Show that $\dot{u}_x = (\omega_0 - \frac{1}{\eta}\mathrm{Re}\,\langle\lambda|\sigma_z|\psi\rangle)u_y \equiv k\,u_y$, where k is constant in time. Similarly, show that $\dot{u}_y = -k\,u_x$ and hence, that the optimal control is of the form $u_x = A\cos\omega t$ and $u_y = A\sin\omega t$, where A and ω are free parameters.
c. Using the optimal controls, evaluate the cost function and confirm Eq. (13.24).
d. Show that $J(\omega, \omega_1)$ is minimized when $\omega = -\omega_0$ (resonance condition for the rotating field). Then minimize over the remaining variable, ω_1, and confirm the statements made in the text about the small- and large-η limits. Qualitatively, how would your conclusions change if J depended on $-(\langle\psi(\tau)|P_1|\psi(\tau)\rangle)^{1/2}$?

Solution.

a. From Eq. (13.20), the augmented cost function is

$$J' = 1 - |\langle\psi(\tau)|1\rangle|^2 + \tfrac{1}{2}\eta \int_0^\tau \mathrm{d}t \left(u_x^2 + u_y^2\right) + \mathrm{Re} \int_0^\tau \mathrm{d}t\, \langle\lambda|\,(-i\partial_t + H)\,|\psi\rangle\,,$$

with

$$H(t) = -\tfrac{1}{2}(\omega_0\sigma_z + u_x\sigma_x + u_y\sigma_y)\,.$$

The Euler-Lagrange equation for variations with respect to the wavefunction $|\psi\rangle$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial J'}{\partial(\partial_t|\psi\rangle)}\right) = \frac{\partial J'}{\partial|\psi\rangle} \quad \Longrightarrow \quad -i\partial_t\langle\lambda| = \langle\lambda|H\,, \qquad \langle\lambda(\tau)| = -\langle\psi(\tau)|P_1\,.$$

The Euler-Lagrange equation for variations with respect to the adjoint $\langle\lambda|$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial J'}{\partial(\partial_t\langle\lambda|)}\right) = \frac{\partial J'}{\partial\langle\lambda|} \quad \Longrightarrow \quad i\partial_t|\psi\rangle = H|\psi\rangle\,, \qquad \text{and } |\psi(0)\rangle = |0\rangle\,.$$

Notice that both $|\psi\rangle$ and $|\lambda\rangle$ obey the Schrödinger equation (remember $\mathrm{i} \to -\mathrm{i}$ when taking the adjoint). This identity is special and results from $H = H^\dagger$ and the fact that the running cost $L(|\psi\rangle, \boldsymbol{u})$ depends only on the controls $\boldsymbol{u}$. (One can make a similar statement, with analogous qualifications, for the classical case: if the running cost $L(\boldsymbol{x}, \boldsymbol{u})$ depends only on $\boldsymbol{u}$ and if the dynamics are linear and without dissipation, then the state obeys $\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x}$ and the adjoint $\dot{\boldsymbol{\lambda}} = -\boldsymbol{A}^{\mathsf{T}}\boldsymbol{\lambda}$.)

The Euler-Lagrange equation for variations with respect to u_x and u_y are

$$\frac{\partial J'}{\partial u_x} = 0\,, \quad \Longrightarrow \quad u_x = \frac{1}{\eta}\mathrm{Re}\,\langle\lambda|\sigma_x|\psi\rangle$$

$$\frac{\partial J'}{\partial u_y} = 0\,, \quad \Longrightarrow \quad u_y = \frac{1}{\eta}\mathrm{Re}\,\langle\lambda|\sigma_y|\psi\rangle\,.$$

Here, we have used $\partial_{u_x}H = -\frac{1}{2}\sigma_x$ and $\partial_{u_y}H = -\frac{1}{2}\sigma_y$.

b. Let us first take a time derivative:

$$\begin{aligned}\partial_t\langle\lambda|\sigma_x|\psi\rangle &= (\partial_t\langle\lambda|)\,\sigma_x|\psi\rangle + \langle\lambda|\sigma_x\,(\partial_t|\psi\rangle) \\ &= \mathrm{i}\langle\lambda|\,H\sigma_x|\psi\rangle - \mathrm{i}\langle\lambda\,|\sigma_x H|\psi\rangle \\ &= \mathrm{i}\langle\lambda|\,[H,\sigma_x]\,|\psi\rangle\end{aligned}$$

Then

$$\begin{aligned}[H,\sigma_x] &= -\tfrac{1}{2}(\omega_0[\sigma_z,\sigma_x] + u_x\cancelto{0}{[\sigma_x,\sigma_x]} + u_y[\sigma_y,\sigma_x]) \\ &= -\tfrac{1}{2}\{\omega_0(2\,\mathrm{i})\sigma_y + 0 + u_y(-2\,\mathrm{i})\sigma_z\} \\ &= -\,\mathrm{i}(\omega_0\sigma_y - u_y\sigma_z)\end{aligned}$$

Substituting then gives

$$\partial_t\langle\lambda|\sigma_x|\psi\rangle = \langle\lambda|(\omega_0\sigma_y - u_y\sigma_z)|\psi\rangle = \omega_0\langle\lambda|\sigma_y|\psi\rangle - u_y\langle\lambda|\sigma_z|\psi\rangle$$

and, hence, that

$$\dot{u}_x = \frac{1}{\eta}\mathrm{Re}\,\left(\omega_0\,\langle\lambda|\sigma_y|\psi\rangle - u_y\,\langle\lambda|\sigma_z|\psi\rangle\right) = \left(\omega_0 - \frac{1}{\eta}\mathrm{Re}\,\langle\lambda|\sigma_z|\psi\rangle\right)u_y$$

The calculation for the $\dot{u}_y$ equation is similar. We find

$$\partial_t \langle\lambda|\sigma_y|\psi\rangle = \mathrm{i}\langle\lambda|[H,\sigma_y]|\psi\rangle\,,$$

with

$$\begin{aligned}[H,\sigma_y] &= -\tfrac{1}{2}(\omega_0[\sigma_z,\sigma_y] + u_x[\sigma_x,\sigma_y] + u_y\cancelto{0}{[\sigma_y,\sigma_y]}) \\ &= -\tfrac{1}{2}\{\omega_0(-2\,\mathrm{i})\sigma_x + u_x(+2\,\mathrm{i})\sigma_z\} \\ &= -\,\mathrm{i}(-\omega_0\sigma_x + u_x\sigma_z)\,.\end{aligned}$$

Substituting then gives

$$\partial_t\langle\lambda|\sigma_y|\psi\rangle = \langle\lambda|(-\omega_0\sigma_x + u_x\sigma_z)|\psi\rangle = -\omega_0\langle\lambda|\sigma_x|\psi\rangle + u_x\langle\lambda|\sigma_z|\psi\rangle\,,$$

and, hence, that

$$\dot{u}_y = \frac{1}{\eta}\mathrm{Re}\ \left(-\omega_0\,\langle\lambda|\sigma_x|\psi\rangle + u_x\,\langle\lambda|\sigma_z|\psi\rangle\right) = \left(-\omega_0 + \frac{1}{\eta}\mathrm{Re}\,\langle\lambda|\sigma_z|\psi\rangle\right)u_x\,.$$

Now we claim that $\mathrm{Re}\,\langle\lambda|\sigma_z|\psi\rangle$ is constant. To see this,

$$\begin{aligned}\partial_t\langle\lambda|\sigma_z|\psi\rangle &= (\partial_t\langle\lambda|)\,\sigma_z|\psi\rangle + \langle\lambda|\,\sigma_z\,(\partial_t|\psi\rangle) \\ &= \mathrm{i}\langle\lambda|\,[H,\sigma_z]\,|\psi\rangle\,.\end{aligned}$$

But

$$\begin{aligned}[H,\sigma_z] &= -\tfrac{1}{2}\{u_x[\sigma_x,\sigma_z] + u_y[\sigma_y,\sigma_z]\} \\ &= -\tfrac{1}{2}\{u_x(-2i)\sigma_y + u_y(2i)\sigma_x\} \\ &= -\,\mathrm{i}(-u_x\sigma_y + u_y\sigma_x)\,.\end{aligned}$$

so that

$$\begin{aligned}\partial_t\,(\mathrm{Re}\,\langle\lambda|\sigma_z|\psi\rangle) &= (-u_x\mathrm{Re}\,\langle\lambda|\,\sigma_y\,|\psi\rangle + u_y\mathrm{Re}\,\langle\lambda|\,\sigma_x)\,|\psi\rangle \\ &= \eta(-u_xu_y + u_yu_x) = 0\end{aligned}$$

We can thus define a scalar constant $k = \omega_0 - \frac{1}{\eta}\mathrm{Re}\,\langle\lambda|\sigma_z|\psi\rangle$, so that

$$\dot{u}_x = k\,u_y\,, \qquad \dot{u}_y = -k\,u_x\,.$$

which has a solution

$$u_x = A\cos(\omega t - \theta_0)\,, \qquad u_y = A\sin(\omega t - \theta_0)\,,$$

where A, ω, and θ_0 are free parameters. As argued in the statement of the problem, rotational symmetry of the equations and initial state implies we can take $\theta_0 = 0$.

c. From Eq. (13.17), the cost function J is

$$J = 1 - |\langle\psi(\tau)|1\rangle|^2 + \tfrac{1}{2}\eta\int_0^\tau \mathrm{d}t\left[u_x^2(t) + u_y^2(t)\right]\,,$$

The terminal-cost term requires integrating the wavefunction from time $t = 0$ to τ. Using Eq. (13.23), we write,

$$\langle\psi(\tau)| = \langle 0| + \mathrm{e}^{-\mathrm{i}\omega t/2}\left[\frac{-\mathrm{i}\omega_1}{a}\sin\left(\tfrac{1}{2}at\right)\right]\langle 1|\,,$$

where we recall that $a = \sqrt{(\omega_0+\omega)^2+\omega_1^2}$. Thus,

$$-|\langle\psi(\tau)|1\rangle|^2 = -\left(\frac{\omega_1}{a}\right)^2\sin^2\left(\tfrac{1}{2}aT\right).$$

For the other term, we note that

$$\int_0^\tau \mathrm{d}t\,(u_x^2+u_y^2) = \int_0^\tau \mathrm{d}t\,\omega_1^2[\cos^2\omega t+\sin^2\omega t] = \omega_1^2\tau\,.$$

All together, the cost function is

$$J(\omega,\omega_1) = 1-\left(\frac{\omega_1}{a}\right)^2\sin^2\left(\tfrac{1}{2}a\tau\right)+\tfrac{1}{2}\eta\omega_1^2\tau\,.$$

d. Since there are no restrictions on ω and ω_1, we can set $\nabla J = 0$. Taking the derivative with respect to ω, we observe that the only ω dependence is through $a^2 = (\omega+\omega_0)^2+\omega_1^2$. Since $a'(\omega) = (\omega+\omega_0)/a$, we have

$$\frac{\partial J}{\partial\omega} = \frac{\partial J}{\partial a}\frac{da}{d\omega} \propto (\omega+\omega_0)\cdots$$

which vanishes when $\omega = -\omega_0$. Substituting this value then gives

$$\begin{aligned} J_1(\omega_1) &= 1-\sin^2\left(\tfrac{1}{2}\omega_1\tau\right)+\eta\omega_1^2\tau \\ &= \tfrac{1}{2}\left(1+\cos\omega_1\tau+2\eta\omega_1^2\tau\right) \\ &= \tfrac{1}{2}\left(1+\cos\theta+\tfrac{1}{2}\eta'\theta^2\right), \end{aligned}$$

where $\theta \equiv \omega_1\tau$ and $\eta' = \eta(4/\tau)$. We then differentiate to find

$$\partial_\theta J_1 = \tfrac{1}{2}\left(-\sin\theta+\eta'\theta\right) = 0\,.$$

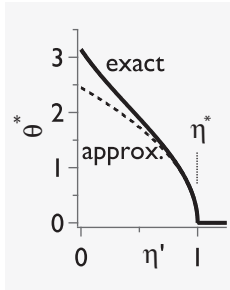
This leads to $\operatorname{sinc}\theta \equiv \frac{\sin\theta}{\theta} = \eta'$, or $\theta^* = \operatorname{sinc}^{-1}\eta'$ for $0 < \eta' < 1$ (and 0 for $\eta' > 1$). Alternatively, we can expand the sine to third order, which gives

$$\theta - \tfrac{1}{6}\theta^3 = \eta'\theta\,,$$

which has solutions $\theta^* = 0$ for $\eta' > 1$ and $\theta^* = \sqrt{6(1-\eta')}$ for $0 < \eta' < 1$. Thus,

$$\theta^* = (\gamma\tau)B^* = \begin{cases} \operatorname{sinc}^{-1}\eta' \approx \sqrt{6(1-\eta')} & 0 < \eta' \le 1 \\ 0 & \eta' > 1 \end{cases}.$$

The exact and approximate expressions for $\theta^*(\eta)$ are plotted below.



Going back to the original variables, $\eta' = 1$ translates to a critical value of η:

$$\eta^* = \tfrac{1}{4}\tau\,.$$

When $\eta = 0$, the exact expression for θ^* implies that $|\psi(\tau)\rangle = |1\rangle$ if $\omega_1\tau = \pi$, or

$$B_1^{\max} = \frac{\pi}{\gamma\tau}\,.$$

Finally, the problem asks what changes when we replace $-|\langle\psi(\tau)|1\rangle|^2$ by its square root in the cost function. If we look at the Taylor expansion in ω_1, we see that the bifurcation at η^* occurs because both the final cost and running cost have, for small ω_1, a leading-order term of ω_1^2. If we take the square root, then that no longer occurs (we balance ω_1 vs. $\eta\omega_1^2$). There is no longer a bifurcation but rather a crossover that is a continuous function of η. Thus, the details of the cost function determine whether a bifurcation or a crossover occurs.

14 Networks and Complex Systems

Problems

14.1 Deterministic graphs. Let us consider some properties of deterministic graphs.

a. Show that the average path length of a circle graph is $\langle d\rangle \sim \frac{1}{4}n$.
b. How does $\langle d\rangle$ of grid graphs scale with n if each node has $2k$ nearest neighbors?
c. Show that the average path length of a large star graph is $\langle d\rangle \sim 2$. (Hint: The paths from hub to periphery have negligible weight for large n).

Solution.

a. Consider a circle graph with n nodes. Pick a node at random. Starting from the two nearest neighbors, there are two paths of length 1. Clearly, there are then two paths of length 2, and so on, until we get half-way around the network. If n is large, it will be irrelevant whether n is odd or even, so assume it to be even. Then the longest path (half the circle) has length $n/2$, and there are two paths to this node. (There would only be one if n were odd.) Thus, we see that there is a uniform distribution of path lengths from 2 to $n/2$. Thus,

$$\langle d\rangle = \frac{1+2+\cdots+n/2}{n/2} = \frac{(n/2)(n/2+1)}{2(n/2)} = \frac{n}{4}+\frac{1}{2} \to \frac{n}{4}\,,$$

in the $n \gg 1$ limit. We use the identity $\sum_{i=1}^{N}(i) = \frac{1}{2}N(N+1)$.

A simpler approach is to recognize that for large n, we can replace the sum with an integral. Then,

$$\langle d\rangle = \frac{\int_0^{n/2} \mathrm{d}x\, x}{\int_0^{n/2}\mathrm{d}x} = \frac{\frac{1}{2}(n/2)^2}{n/2} = \frac{n}{4}\,.$$

b. If each node in a grid graph has $2k$ nearest neighbors, then the nodes form a k dimensional grid with periodic boundary conditions. Thus for $k = 1$, there are 2 nearest neighbors, and the graph is a circle, as already discussed. For $k = 2$, there are 4 nearest neighbors and the graph is a two-dimensional grid (see main text). Since there are n nodes total, the grid is $n^{1/2} \times n^{1/2}$. Arguing

heuristically, we expect average paths to be of this scale. Thus, in this case $\langle d\rangle \sim n^{1/2}$. More generally,

$$\langle d\rangle \sim n^{1/k}\,.$$

c. Pick two nodes. Either both will be *peripheral* nodes (path length =2) or one will be a hub and the other peripheral (path length = 1). The probability that both are peripheral is

$$\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n-1}\right) = \frac{n-2}{n}\,,$$

where the second probability is from the reduced set $n-1$ left over after having chosen the first node.

Then the average path length is

$$\langle d\rangle = \left(\frac{n-2}{n}\right)(2) + \left[1 - \left(\frac{n-2}{n}\right)\right](1) = 2 - \frac{4}{n} + \frac{2}{n} = 2 - \frac{2}{n}\,,$$

so that $\langle d\rangle \to 2$ as $n \to \infty$.

14.2 Random graphs. Consider the properties of random Erdős–Rényi graphs with n nodes and a probability p for a link between any two nodes chosen at random. Work in the "Poisson" limit where the only parameter is $\langle k\rangle$, the average node degree.

a. Give a reasonable argument that the average path length $\langle d\rangle \sim \ln n/\ln k$. Hint: consider the "fan out" of paths. Each node reaches roughly $\langle k\rangle$ nodes after paths of length one, $\langle k\rangle^2$ nodes after paths of length two, ...

b. Show that the graph becomes connected at $\langle k\rangle = \ln n$, for large n. Hints: Estimate the probability for a node to have no links and use $(1-p)^n \approx \mathrm{e}^{-np}$.

Solution.

a. Following the hint, the number of nodes reached after following paths of length d is approximately

$$n_d \approx \langle k\rangle + \langle k\rangle^2 + \langle k\rangle^3 + \cdots + \langle k\rangle^d \approx \frac{\langle k\rangle^{d+1} - 1}{\langle k\rangle - 1} \approx \langle k\rangle^d\,.$$

Since we expect the average path to cover a substantial portion of the network, the number of nodes explored n_d should be on the order of the total number of nodes n. Thus,

$$n \approx \langle k\rangle^d\,, \quad \text{or, inverting} \quad \langle k\rangle \approx (\ln n)/(\ln d)\,.$$

b. As the hint suggests, we estimate the probability p_0 that a node, chosen at random, does *not* have any links, given that the probability to have a link between any two nodes is p. We can take as a condition for threshold that the

expected number of such nodes is equal to one. Thus, we want $np_0 = 1$, for $p_0 = 1/n$. Since there are roughly n other nodes a given node can connect to

$$p_0 \approx (1-p)^n \approx \mathrm{e}^{-np} = \frac{1}{p}\,.$$

Inverting gives $p_c \approx \ln n/n$. Since $\langle k\rangle = p(n-1)$, we have

$$\langle k\rangle \approx \ln n\left(\frac{n-1}{n}\right) = \ln n\,.$$

Again, we work in the large-n limit.

14.3 Scale-free graphs. Consider node distributions $P(k) = Z^{-1}k^{-\gamma}$, with $\gamma \geq 2$ and $k > k_{\min}$. Treat, for simplicity, the node degree as a continuous density, $p(k)$.

a. Find the normalization constant Z for the continuous probability density.
b. Show that changing the units $k \to ak'$ does not change the form of the distribution.
c. Show that $\langle k\rangle = (\frac{\gamma-1}{\gamma-2})\,k_{\min}$.
d. For a finite network of n nodes, show that $k_{\max} = k_{\min}\,n^{\frac{1}{\gamma-1}}$.
e. Why then does $\gamma < 2$ imply multiple edges between node pairs?

Solution.

a. For continuous k, the normalization constant is

$$\int_{k_{\min}}^{\infty} \mathrm{d}k\,k^{-\gamma} = \frac{k_{\min}^{1-\gamma}}{\gamma-1}\,, \qquad \Longrightarrow \qquad p(k) = \left(\frac{\gamma-1}{k_{\min}}\right)\left(\frac{k}{k_{\min}}\right)^{-\gamma}\,.$$

b. We change scale by substituting $k = ak'$ and $k_{\min} = ak'_{\min}$ into the distribution:

$$\begin{aligned}\mathrm{d}k\,p(k) &= \mathrm{d}k\left(\frac{\gamma-1}{k_{\min}}\right)\left(\frac{k}{k_{\min}}\right)^{-\gamma} = a(\mathrm{d}k')\left(\frac{\gamma-1}{ak'_{\min}}\right)\left(\frac{ak'}{ak'_{\min}}\right)^{-\gamma} = \mathrm{d}k'\left(\frac{\gamma-1}{k'_{\min}}\right)\left(\frac{k'}{k'_{\min}}\right)^{-\gamma}\\ &= \mathrm{d}k'\,p(k')\,.\end{aligned}$$

Thus, changing the scale by a factor a does not change the probability density.

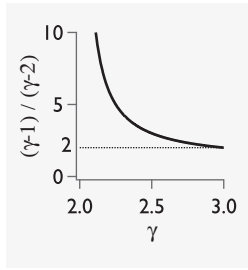
c. The average node degree is

$$\langle k\rangle = \int_{k_{\min}}^{\infty} \mathrm{d}k\,k\left(\frac{\gamma-1}{k_{\min}}\right)\left(\frac{k}{k_{\min}}\right)^{-\gamma}\,.$$

Let $k' = k/k_{\min}$. Then

$$\begin{aligned}\langle k\rangle &= (\gamma-1)k_{\min}\int_1^{\infty} \mathrm{d}k'\,(k')^{-(\gamma-1)}\\ &= (\gamma-1)k_{\min}\left(\frac{1}{\gamma-2}\right)k'\Big|_{\infty}^{1}\\ &= \left(\frac{\gamma-1}{\gamma-2}\right)k_{\min}\,.\end{aligned}$$

For $\gamma = 3$, this gives $\langle k \rangle = 2k_{\min}$. The prefactor diverges as $\gamma \to 2$ (see below).



d. $k_{\max}$ is set by asking that the expected number of nodes with $k_{\max}$ or greater links be less than one. Equivalently, the probability to have such a node is $\leq 1/n$. Then,

$$\begin{aligned}\frac{1}{n} &= \int_{k_{\max}}^{\infty} \mathrm{d}k\, p(k) = \int_{k_{\max}}^{\infty} \mathrm{d}k \left(\frac{\gamma-1}{k_{\min}}\right)\left(\frac{k}{k_{\min}}\right)^{-\gamma} \\ &= \int_{k'_{\max}}^{\infty} \mathrm{d}k'\, (\gamma-1)\, k'^{-\gamma} \\ &= \left(\frac{\gamma-1}{\gamma-1}\right) k'^{-(\gamma-1)}\Big|_{k'=k'_{\max}} = \left(\frac{k_{\min}}{k_{\max}}\right)^{\gamma-1} .\end{aligned}$$

Inverting gives

$$k_{\max} = k_{\min}\left(n^{\frac{1}{\gamma-1}}\right) .$$

For $2 \leq \gamma \leq 3$, the dependence on the number of network nodes ranges from linear (n) to square root ($n^{1/2}$).

e. For $\gamma < 2$, part (d) implies that $k_{\max}$ increases with n as a power law with exponent $1/(\gamma - 1) > 1$—faster than linear. Then, for large-enough n, the largest node will connect to more than n nodes, meaning that there must be multiple edges joining the same node pair. This can occur—think of chemical species that can transform into each other via different reactions—but is often not allowed.

14.4 Scaling in the Barabási-Albert model of preferential attachment. Consider a network that adds one node each time step. Let $n(k,t)$ be the number of nodes with degree k at time t and $p(k,t) = n(k,t)/n(t)$ the corresponding degree-node distribution at time t. Each new node adds m links. Each new link goes to an existing node, with the probability to connect to a node of degree k given by $\Pi(k) = k/\sum_j k_j = k/(2mt)$. For the denominator: at time t there are mt links, and each link connects two nodes.

a. Show that typically $\frac{k}{2}p(k,t)$ links are added to degree-k nodes at time t and that $(n+1)p(k,t+1) = np(k,t) + \left(\frac{k-1}{2}\right)p(k,t) - \left(\frac{k}{2}\right)p(k,t)$.

b. In the long-time limit, $p(k,t) \to p_k$. Show that the master equation in (a) becomes $p_k = \left(\frac{k-1}{2}\right) p_{k-1} - \left(\frac{k}{2}\right) p_k$.
c. Derive the continuum limit $p_k = -\frac{1}{2}\partial_k(kp_k)$ and verify that $p_k \sim k^{-3}$ is a solution.

Solution.

a. From the preferential-attachment law, the number of links that connect, on average, to the set of degree-k nodes at time t is given by the number of nodes with degree k at time t times the probability to attach to such nodes $\Pi(k)$ times the number of links added for each node m, or

$$[n(t)\, p(k,t)] \left(\frac{k}{2mt}\right) m = \frac{k}{2} p(k,t)\,,$$

using $n(t) = t$ as the number of nodes at time t. Then we note that adding a link to a node of degree k *increases* the population of $k+1$-degree nodes but *decreases* the population of degree-k nodes. The master equation captures this dynamic:

$$\underbrace{(n+1)p(k,t+1)}_{\text{degree-}k\text{ nodes at time }t+1} = \underbrace{np(k,t)}_{\text{degree-}k\text{ nodes at time }t} + \underbrace{\left(\frac{k-1}{2}\right)p(k-1,t)}_{(k-1)\to k} - \underbrace{\left(\frac{k}{2}\right)p(k,t)}_{k\to(k+1)}\,,$$

Note that this equation is modified for $k = m$, since each new node is automatically also a node of degree m. But we are interested in the large-k behavior and can therefore neglect the "boundary condition" at $k = m$.
b. In the long-time limit, we set $p(k,t+1) = p(k,t) = p_k$. Thus,

$$(n+1)p_k = np_k + \left(\frac{k-1}{2}\right)p_{k-1} - \left(\frac{k}{2}\right)p_k\,,$$

implying

$$p_k = \left(\frac{k-1}{2}\right)p_{k-1} - \left(\frac{k}{2}\right)p_k$$

c. The continuum limit is

$$p_k = -\tfrac{1}{2}\left[k\, p_k - (k-1)\, p_{k-1}\right] \approx -\tfrac{1}{2}\partial_k(k\, p_k)\,.$$

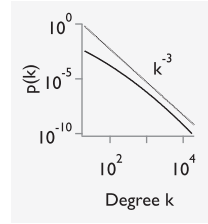
It is then straightforward to verify that $p_k \sim k^{-3}$ is a solution:

$$-\tfrac{1}{2}\partial_k(k\, p_k) \approx -\tfrac{1}{2}\partial_k k^{-2} \approx k^{-3}\,,$$

which is just p_k.

Note that parts of this problem are from Barabási (2016) but that there are typos in Eqs. 5.43 and 5.44.

14.5 Lognormal vs. power law. In practice, it is not easy to distinguish a power-law distribution from alternatives such as the lognormal distribution.



a. Show that a lognormal degree distribution $\ln k \sim \mathcal{N}(\mu, \sigma^2)$ is "scale free" in that a change of scale $k = ak'$ merely shifts the distribution on a log-log plot.
b. By creating a plot such as the one at right, show that you can find parameters in a lognormal distribution that are close to a given power law over some decades.

Solution.

a. The lognormal probability density function (pdf) is

$$p(k) = \left(\frac{1}{k\sigma\sqrt{2\pi}}\right) \mathrm{e}^{-\frac{(\ln k-\mu)^2}{2\sigma^2}} \ .$$

If we carry out the transformation $k = ak'$ in the expression $\mathrm{d}k\, p(k)$, we have

$$\begin{aligned} \mathrm{d}k\, p(k) &= a(\mathrm{d}k') \left(\frac{1}{ak'\sigma\sqrt{2\pi}}\right) \mathrm{e}^{-\frac{(\ln(ak')-\mu)^2}{2\sigma^2}} \\ &= \mathrm{d}k' \left(\frac{1}{k'\sigma\sqrt{2\pi}}\right) \mathrm{e}^{-\frac{(\ln k'-\mu')^2}{2\sigma^2}} \qquad \mu' = \mu - \ln a \,, \end{aligned}$$

which has the same form, except that the mean is shifted by $\ln a$. Notice that σ does not change. So, in this sense, one can think of the lognormal distribution as scale free. But, in another sense, it is not: The quantity $\ln k$ is normally distributed and has the obvious scale σ. A loose example is that a quantity could be $10^{2\pm1}$. There is a scale of ±1 in the uncertainty of the exponent. But because it is in the exponent, the range is 10–1000, which is pretty wide.

b. The plot in the book is generated from $\ln x \sim \mathcal{N}(\mu, \sigma^2)$, with $\mu = 0.1$ and $\sigma = 0.2$. The exercise is mainly to show that a mere fit of a model to data is not enough. There should be independent reasons for justifying a fit. Otherwise, some other model may fit nearly as well.

14.6 Controllability of an n-chain. Following an example from Sun and Motter (2013), we consider a chain of n one-dimensional systems with a single input at its head. The dynamics are that of an n-fold integrator: $\dot{x}_1 = u$, $\dot{x}_2 = x_1$, $\ldots$, $\dot{x}_n = x_{n-1}$.

a. Show that the controllability matrix $\boldsymbol{W}_\mathrm{c} = \mathbb{I}_n$, the n-dimensional identity matrix. The Kalman rank condition for controllability is thus satisfied.
b. A system has *strong* structural controllability if it is structurally controllable for all non-zero values of the weights. Show that the n-chain defined above has this property, allowing for arbitrary weights $b, a_{21}, a_{31}, \ldots a_{n1}$.

Solution.

a. From the dynamical equations, we can read off the system matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. For example, for $n = 4$, we have

$$\boldsymbol{A} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \qquad \boldsymbol{B} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Notice that powers, $\boldsymbol{A}^i$, make the row of 1s "move away" from the diagonal by one step per multiplication. That is,

$$\boldsymbol{A}^2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \qquad \boldsymbol{A}^3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Then, since $\boldsymbol{A}^m\boldsymbol{B}$ picks out the first column of $\boldsymbol{A}^m$, it is immediately clear that $\boldsymbol{W}_\mathrm{c} = \mathbb{I}_n$.

Alternatively, we can use index notation, where $\boldsymbol{A}_{ij} = \delta_{i,j+1}$. Then

$$\boldsymbol{A}^2_{ij} = \delta_{i,k+1}\delta_{k,j+1} = \delta_{i,j+2} \quad \Longrightarrow \quad \left(\boldsymbol{A}^2\boldsymbol{B}\right)_i = \delta_{i,j+2}\,\delta_{j,1} = \delta_{i,3}\,.$$

Continuing the same pattern, we have $\boldsymbol{A}^m_{ij} = \delta_{i,j+m}$. Since $\boldsymbol{B}_j = \delta_{j,1}$, we have $(\boldsymbol{A}^m\boldsymbol{B})_i = \delta_{i,m+1}$, which also implies that $\boldsymbol{W}_\mathrm{c}$ is the identity matrix.

b. For arbitrary weights, the state-space matrices become (for $n = 4$)

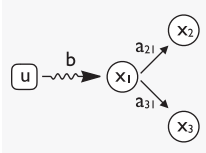
$$\boldsymbol{A} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ a_{21} & 0 & 0 & 0 \\ 0 & a_{32} & 0 & 0 \\ 0 & 0 & a_{43} & 0 \end{pmatrix}, \qquad \boldsymbol{B} = \begin{pmatrix} b \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Repeating the arguments from (a), we find a controllability matrix

$$\boldsymbol{W}_\mathrm{c} = \begin{pmatrix} b & 0 & 0 & 0 \\ 0 & a_{21}b & 0 & 0 \\ 0 & 0 & a_{32}b & 0 \\ 0 & 0 & 0 & a_{43}b \end{pmatrix},$$

which obviously has rank=4 (det $\neq 0$) for all non-zero values of b and the a_{i1} elements. The system is strongly structurally controllable.

14.7 Dilation. Consider the graph at left of an LTI dynamical system.



a. Show that the system is not controllable, for any values of the link weights.
b. Add a self-link of weight a_{22} to the node x_2. Show that the system is now controllable and has the strong structural controllability property.
c. Why does adding a self-interaction node remove the dilation?
d. Finally, add a second self-link of weight a_{33} to the node x_3. Show that the system is controllable but does not have the strong structural controllability property.

Solution.

a. The graph corresponds to the dynamical system

$$\dot{x}_1 = b\,u(t)\,, \qquad \dot{x}_2 = a_{21}x_1\,, \qquad \dot{x}_3 = a_{31}x_1\,.$$

If we multiply the x_2 equation by a_{31} and the x_3 equation by a_{21} and then subtract, we have

$$a_{31}\dot{x}_2 - a_{21}\dot{x}_3 = 0\,, \quad \Longrightarrow \quad a_{31}x_2 - a_{21}x_3 = \text{ constant}\,,$$

meaning that motion in the x_2–x_3 plane is limited to the line $a_{31}x_2 - a_{21}x_3 =$ constant. More formally, the state-space matrices are

$$\boldsymbol{A} = \begin{pmatrix} 0 & 0 & 0 \\ a_{21} & 0 & 0 \\ a_{31} & 0 & 0 \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} b \\ 0 \\ 0 \end{pmatrix},$$

which implies that the controllability matrix is

$$\boldsymbol{W}_{\mathrm{c}} = \begin{pmatrix} \boldsymbol{B} & \boldsymbol{A}\boldsymbol{B} & \boldsymbol{A}^2\boldsymbol{B} \end{pmatrix} = \begin{pmatrix} b & 0 & 0 \\ 0 & a_{21}b & 0 \\ 0 & a_{31}b & 0 \end{pmatrix}.$$

The matrix $\boldsymbol{W}_{\mathrm{c}}$ has det=0 (rank=2). The system is therefore not controllable.

b. Now we add a self-interaction node, meaning that the dynamics are

$$\dot{x}_1 = b\,u(t)\,, \qquad \dot{x}_2 = a_{21}x_1 + a_{22}x_2\,, \qquad \dot{x}_3 = a_{31}x_1\,.$$

The state-space and controllability matrices are

$$\boldsymbol{A} = \begin{pmatrix} 0 & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & 0 & 0 \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} b \\ 0 \\ 0 \end{pmatrix}, \quad \Longrightarrow \quad \boldsymbol{W}_{\mathrm{c}} = \begin{pmatrix} b & 0 & 0 \\ 0 & a_{21}b & a_{21}a_{22}b \\ 0 & a_{31}b & 0 \end{pmatrix}.$$

The determinant of $\boldsymbol{W}_{\mathrm{c}} = -a_{21}a_{31}a_{22}b^3$, which vanishes only if one of the system parameters a_{21}, a_{22}, a_{31}, of b equals zero. Thus, the system has strong structural controllability.

c. Without the self-interaction, the network illustrated below at left has a dilation and is not controllable. As stated in the main text, the subset $S = \{x_2, x_3\}$ has two elements, while its neighborhood $T(S) = \{x_1\}$ has but one. Since the neighborhood has fewer elements than the set, there is a dilation and the system is not controllable.

With self-interaction (network below at right), the neighborhood $T(S) = \{x_1, x_2\}$ has the same cardinality as S. Thus, because a self-interaction means that a node is a neighbor of itself, we no longer have a situation where a subset of nodes has a smaller neighborhood than itself. Consequently, we do not violate Lin's conditions for structural controllability, and the system is structurally controllable.

d. Finally, if we add another self-link, this time to node x_3, the dynamics are

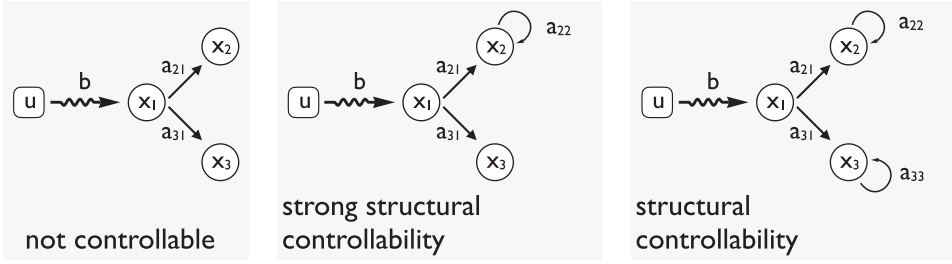
$$\dot{x}_1 = b\,u(t)\,, \qquad \dot{x}_2 = a_{21}x_1 + a_{22}x_2\,, \qquad \dot{x}_3 = a_{31}x_1 + a_{33}x_3\,.$$

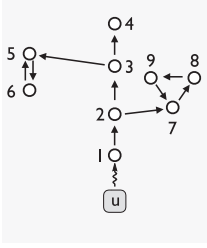
The state-space and controllability matrices are

$$\boldsymbol{A} = \begin{pmatrix} 0 & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & 0 & a_{33} \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} b \\ 0 \\ 0 \end{pmatrix}, \quad \Longrightarrow \quad \boldsymbol{W}_{\mathrm{c}} = \begin{pmatrix} b & 0 & 0 \\ 0 & a_{21}b & a_{21}a_{22}b \\ 0 & a_{31}b & a_{31}a_{33}b \end{pmatrix}.$$

The determinant of $\boldsymbol{W}_{\mathrm{c}} = -a_{21}a_{31}(a_{22} - a_{33})b^3$, which vanishes only if $a_{22} = a_{33}$. Thus, the system has structural controllability but not strong structural controllability.

The digraphs for the three situations are given below.





14.8 Cactus is a minimum controllable structure. Consider the cactus. At left is a cactus with two buds, reproduced from Figure 14.6d and then annotated. Show that removing *any* edge renders the resulting network uncontrollable.

Solution.

There are three types of edges:

- Edges along the stem (here, $u \to 1$, $1 \to 2$, $2 \to 3$, $3 \to 4$). Removing any of these edges will make the target vertex (and downstream ones) inaccessible. For example, if we remove the edge between vertex 3 and vertex 4, then vertex 4 is inaccessible.
- Edges within a cycle of the bud (here, $5 \to 6$, $6 \to 5$, along with $7 \to 8$, $8 \to 9$, $9 \to 7$). Removing any of these edges will create a dilation. For example, if we remove the edge $9 \to 7$, then vertex 2 becomes a dilation. The set of vertices $\{3, 7\}$ would then have a neighborhood $T(S)$ that consists only of vertex 2.
- Edges to a cycle that turn the cycle into a bud ($3 \to 5$, $2 \to 7$). Removing either of these edges will create cycles that are isolated. All the vertices in the cycle are inaccessible.

Thus, removing any edge makes the network uncontrollable, in the formal sense of the notion of controllability.

14.9 Self-interactions. By computing the determinant of the controllability matrix $\boldsymbol{W}_{\mathrm{c}}$, give a non-graphical proof that adding self-interactions generically implies

that a system is controllable from one input. Hint: Set all non-self interactions equal to zero and all input couplings to one. (Justify these assumptions.)

Solution.

Following the hint, we consider dynamics of the form, for each component $1 \le i \le n$,

$$\dot{x}_i = -\lambda_i x_i + u\,.$$

The controllability matrix is then

$$\boldsymbol{W}_{\mathrm{c}} = \begin{pmatrix} 1 & (-\lambda_1) & (-\lambda_1)^2 & \cdots & (-\lambda_1)^{n-1} \\ 1 & (-\lambda_2) & (-\lambda_2)^2 & \cdots & (-\lambda_2)^{n-1} \\ & & \vdots & & \\ 1 & (-\lambda_n) & (-\lambda_n)^2 & \cdots & (-\lambda_n)^{n-1} \end{pmatrix}.$$

Since generically the λ_i are all different, then it is clear that $\det \boldsymbol{W}_{\mathrm{c}} \neq 0$. Having proven structural controllability for this special case, it is clear that introducing generic non-zero edges (A_{ij} for $i \neq j$) or letting the coupling constants (the B_i) be different cannot change this situation (the values where accidental vanishing occurs will vary and depend on the constants chosen).

14.10 Control effort in one dimension. For $\dot{x} = \lambda x + u$, with $x(0) = x_0$ and $x(\tau) = x_\tau$:

a. Show that the minimum control effort is $\mathcal{E} = 2\lambda(x_\tau - \mathrm{e}^{\lambda\tau} x_0)^2/(\mathrm{e}^{2\lambda\tau} - 1)$.
b. Deduce the short- and fast-protocol limits given in the text.
c. For $\tau^* = \lambda^{-1} \ln(x_\tau/x_0)$, the minimum effort $\mathcal{E} = 0$. What is going on?
d. Show that the minimum-effort trajectory $x(t)$ is identical for $\lambda \to -\lambda$.

Solution.

a. The Gramian for this problem is

$$P(\tau) = \int_0^\tau \mathrm{d}t\, \mathrm{e}^{\lambda t}\,.1.1.\,\mathrm{e}^{\lambda t} = \frac{\mathrm{e}^{2\lambda\tau} - 1}{2\lambda}\,.$$

The minimum-effort input is then

$$u(t) = \mathrm{e}^{\lambda(\tau - t)}\, P^{-1}(x_\tau - \mathrm{e}^{\lambda\tau} x_0) = \frac{2\lambda\, \mathrm{e}^{\lambda(\tau-t)}(x_\tau - \mathrm{e}^{\lambda\tau} x_0)}{\mathrm{e}^{2\lambda\tau} - 1}\,,$$

which leads to a minimum control effort

$$\mathcal{E} = \frac{2\lambda(x_\tau - \mathrm{e}^{\lambda\tau} x_0)^2}{\mathrm{e}^{2\lambda\tau} - 1}\,.$$

b. The short-time limit is $\lambda\tau \ll 1$, which leads to

$$\mathcal{E} \approx \frac{2\lambda(x_\tau - (1)x_0)^2}{1 + 2\lambda\tau - 1} = \frac{(x_\tau - x_0)^2}{\tau}\,.$$

The long-time is $\lambda\tau \gg 1$, which leads to

$$\mathcal{E} \approx \frac{2\lambda x_0^2}{-1} = -2\lambda x_0^2 = 2|\lambda| x_0^2\,.$$

In the same long-time limit, for $\lambda > 0$, we have

$$\mathcal{E} \approx \frac{2\lambda\, \mathrm{e}^{2\lambda\tau}\, x_0^2}{\mathrm{e}^{2\lambda\tau}} = 2\lambda x_0^2 .$$

The fact that the limits depend only on one or the other corresponds, as briefly mentioned in the text, to the idea that only movement against the flow is costly. The other part of the trajectory is nearly free. But the direction of flow does depend on whether the local equilibrium $x = 0$ is stable or unstable. More generally, we can partition the dynamics into stable and unstable subspaces and draw corresponding conclusions about the difficulty of particular control movements accordingly.

c. It is easy to verify that $\tau^* = \lambda^{-1} \ln(x_\tau / x_0)$ makes $\mathcal{E} = 0$. The case corresponds to "natural," uncontrolled motion, with $u = 0$. The effort is obviously zero, and the solution $x(t) = x_0\, \mathrm{e}^{\lambda t}$. If the desired x_0, x_τ, and τ are all compatible with these values, then you can go from x_0 to x_τ for free!

d. We substitute the optimal control $u(t)$ into the equations of motion and find

$$x(t) = \frac{x_\tau \sinh \lambda\tau + x_0 \sinh \lambda(\tau - t)}{\sinh \lambda\tau} ,$$

which is invariant under $\lambda \to -\lambda$ because $\sinh(\cdot)$ is odd: $\sinh(-x) = -\sinh x$.

14.11 Control effort diverges in a nearly uncontrollable system. Consider two first-order equations driven by a common input: $\dot{x}_1 = -x_1 + u\,, \quad \dot{x}_2 = -(1+\delta)x_2 + u.$

a. Calculate the Gramian $\boldsymbol{P}(\tau)$ and show that its determinant $\sim \delta^2$, for $\delta \ll 1$. Argue that this implies that the control effort $\mathcal{E} \sim \delta^{-2}$.

b. Calculate numerically and then plot the minimum-effort trajectory connecting $\binom{0}{0} \to \binom{0}{1}$ for $0 \le t \le 1$. Plot, too, the control effort $\mathcal{E}$ as a function of δ.

Solution.

a. The state-space matrices are

$$\boldsymbol{A} = \begin{pmatrix} -1 & 0 \\ 0 & -(1+\delta) \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

The finite-time control Gramian for a protocol going from 0 to τ is then

$$\begin{aligned} \boldsymbol{P}(\tau) &= \int_0^\tau \mathrm{d}t\, \mathrm{e}^{\boldsymbol{A}^\mathsf{T} t}\, \boldsymbol{B}^\mathsf{T} \boldsymbol{B}\, \mathrm{e}^{\boldsymbol{A} t} \\ &= \int_0^\tau \mathrm{d}t \begin{pmatrix} \mathrm{e}^{-t} & 0 \\ 0 & \mathrm{e}^{-(1+\delta)t} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \mathrm{e}^{-t} & 0 \\ 0 & \mathrm{e}^{-(1+\delta)t} \end{pmatrix} \\ &= \int_0^\tau \mathrm{d}t \begin{pmatrix} \mathrm{e}^{-2t} & \mathrm{e}^{-2(1+\delta/2)t} \\ \mathrm{e}^{-2(1+\delta/2)t} & \mathrm{e}^{-2(1+\delta)t} \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} 1 - \mathrm{e}^{-2\tau} & \frac{1}{1+\delta/2} \left(1 - \mathrm{e}^{-(1+\delta/2)\tau}\right) \\ \frac{1}{1+\delta/2} \left(1 - \mathrm{e}^{-(1+\delta/2)\tau}\right) & \frac{1}{1+\delta} \left(1 - \mathrm{e}^{-(1+\delta)\tau}\right) \end{pmatrix} . \end{aligned}$$

The determinant can be calculated using a symbolic-manipulation program:

$$\det \boldsymbol{P}(\tau) = \frac{\mathrm{e}^{-2(\delta+2)\tau}\left(\delta^2\,\mathrm{e}^{2(\delta+2)\tau} + \delta^2 - (\delta+2)^2\,\mathrm{e}^{2\tau} - (\delta+2)^2\,\mathrm{e}^{2(\delta+1)\tau} + 8(\delta+1)\,\mathrm{e}^{(\delta+2)\tau}\right)}{4(\delta+1)(\delta+2)^2}$$

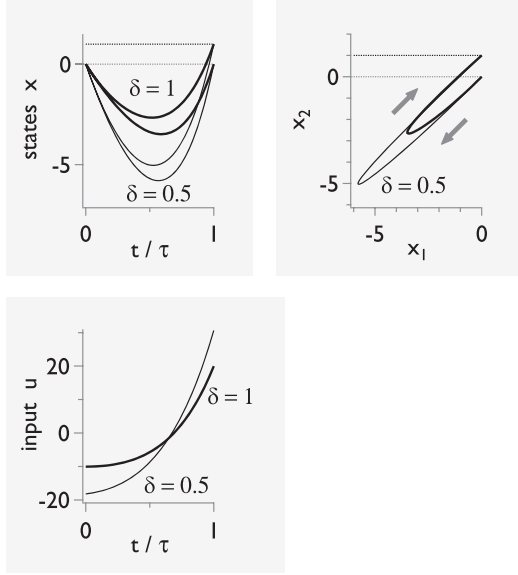
$$= \frac{\delta^2}{16}\left[1 - 2\left(2\tau^2+1\right)\mathrm{e}^{-2\tau} + \mathrm{e}^{-4\tau}\right] + O(\delta^3)\,.$$

The term in brackets interpolates between $\frac{4}{3}\tau^4$ and 1, for $\tau \ll 1$ and $\tau \gg 1$, respectively.

The control effort $\mathcal{E} \sim \boldsymbol{P}^{-1}(\tau)$. Above, we have shown that $\det \boldsymbol{P} \sim \delta^2$. The determinant of the inverse is thus $\sim \delta^{-2}$ and is the product of the inverse of the eigenvalues of $\boldsymbol{P}$. Now it is easy to see that one eigenvalue of $\boldsymbol{P}$ is of order unity and the other of order δ^2. Physically, the controllable subspace requires $O(1)$ effort. The other eigenvalue is thus $\sim \delta^2$. The size of the control effort involves matrix products of $\boldsymbol{P}^{-1}$ and is therefore on the scale of $\lambda_{\min}^{-1}$, where $\lambda_{\min}$ is the smallest eigenvalue of $\boldsymbol{P}$.

All of these properties are much easier to establish in the long-time limit, $\tau \gg 1$. However, that limit leads to large control values that are not caused by degeneracy. In the long-time limit, an input $u(t)$ that tended to a constant value would force states to approach λ^{-1}, where λ is the rate constant (1 or $1+\delta$) of the relaxation. One needs a violently changing $u(t)$ to get two simultaneously different values.

b. The plots are shown below. Thicker lines are calculated with greater mismatch ($\delta = 1$) than thinner lines ($\delta = 0.5$). The parametric plot at top right of $x_2(t)$ versus $x_1(t)$ illustrates the nonlocality of these minimum-effort trajectories, explored in Problem 4.3. The bottom plot shows the common input used to drive both subsystems. Its magnitude is larger for $\delta = 0.5$.



14.12 Control effort can be sensitive to direction and dimension. Consider the linear system $\boldsymbol{A} = \left(\begin{smallmatrix} -3.2 & 1.3 \\ 1.3 & -2.7 \end{smallmatrix}\right)$, $\boldsymbol{B} = \left(\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right)$, with protocol duration $\tau = 3$ (Yan et al., 2015).

a. Find the eigenvalues of the system dynamics $\boldsymbol{A}$ and also the controllability ($\boldsymbol{W}_c$) and Gramian matrices $\boldsymbol{P}$; confirm the plots and numbers given in the text. Plot $u(t)$ for the minimum- and maximum-effort inputs.
b. Show that applying powers of the Gramian $\boldsymbol{P}$ to an arbitrary unit vector (normalizing at each step) gives the target direction requiring the least control effort, and powers of $\boldsymbol{P}^{-1}$ give the target direction requiring the most control effort.
c. Enlarge the system to three dimensions. For $\boldsymbol{A} = \left(\begin{smallmatrix} -3.2 & 1.3 & 1 \\ 1.3 & -2.7 & 0.7 \\ 1 & 0.7 & -2.2 \end{smallmatrix}\right)$, $\boldsymbol{B} = \left(\begin{smallmatrix} 0 \\ 1 \\ 0 \end{smallmatrix}\right)$, show that the ratio of largest to smallest control efforts ≈ 979.
d. Show that the efforts for the easiest direction are roughly the same in the $n = 2$ and $n = 3$ cases, whereas the efforts for the hardest direction differ significantly.

Solution.

a. For the dynamical system defined in the text,

$$\boldsymbol{A} = \begin{pmatrix} -3.2 & 1.3 \\ 1.3 & -2.7 \end{pmatrix}, \qquad \boldsymbol{B} = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

the eigenvalues and eigenvectors of the dynamical matrix $\boldsymbol{A}$ are given by

$$\lambda \approx (-4.2738, -1.6262) \quad \leftrightarrow \quad \approx \begin{pmatrix} -0.7710 \\ 0.6368 \end{pmatrix}, \quad \begin{pmatrix} -0.6368 \\ -0.7710 \end{pmatrix}.$$

Because $\boldsymbol{A}$ is symmetric, the eigenvectors are orthogonal. (If there are degenerate eigenvalues, the eigenvectors can still be chosen orthogonal.)
The controllability matrix is

$$\boldsymbol{W}_c = \begin{pmatrix} 0 & 1.3 \\ 1 & -2.7 \end{pmatrix},$$

which has rank = 2 and determinant -1.3. Notice that all the numbers so far are of order one. The Gramian for $\tau = 3$, by contrast, is

$$\boldsymbol{P} \approx \begin{pmatrix} 0.0206 & 0.0507 \\ 0.0507 & 0.210 \end{pmatrix},$$

with eigenvalues $\approx 0.222, 0.00785$. The ratio is now more than 28. Note that the finite-time Gramian given here for $\tau = 3$ is identical to the infinite-time limit, at the three-digit precision used here.

The control effort in direction $\hat{\boldsymbol{n}}$ is

$$\mathcal{E}(\hat{\boldsymbol{n}}) = \hat{\boldsymbol{n}}^{\mathsf{T}} \boldsymbol{P}^{-1}(\tau = 3)\hat{\boldsymbol{n}},$$

which takes on maximum and minimum values at the eigenvectors corresponding to the minimum and maximum eigenvalues of $\boldsymbol{P}$ (since they are the inverse of the eigenvalues of $\boldsymbol{P}^{-1}$). We can then confirm that

$$\mathcal{E}_{\min} \approx 4.50\,, \quad \text{for } \hat{\boldsymbol{n}}_{\min} \approx \begin{pmatrix} 0.24 \\ 0.97 \end{pmatrix}$$

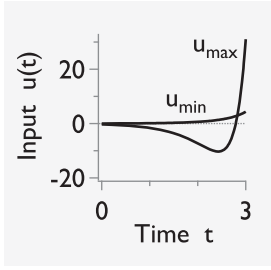
$$\mathcal{E}_{\max} \approx 127.4\,, \quad \text{for } \hat{\boldsymbol{n}}_{\max} \approx \begin{pmatrix} -0.97 \\ 0.24 \end{pmatrix}.$$

The ratio of largest to smallest control efforts $\mathcal{E}_{\max}/\mathcal{E}_{\min} = 28.32$.

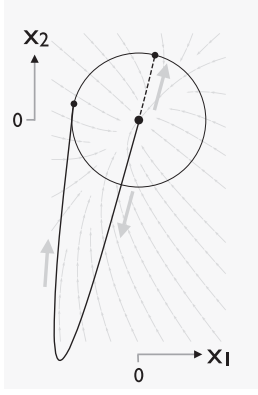
The optimal input to go to a target state $\hat{\boldsymbol{n}}$ on the unit circle is given by

$$u_{\text{opt}}(t, \hat{\boldsymbol{n}}) = \boldsymbol{B}^{\mathsf{T}}\, \mathrm{e}^{\boldsymbol{A}^{\mathsf{T}}(\tau - t)}\, \boldsymbol{P}^{-1}(\tau)\, \hat{\boldsymbol{n}}\,.$$

We plot $u(t)$ for $\hat{\boldsymbol{n}}_{\min}$ and $\hat{\boldsymbol{n}}_{\max}$ below. The input $u_{\max}(t) \equiv u_{\text{opt}}(t, \hat{\boldsymbol{n}}_{\max})$ requires much higher amplitudes than $u_{\min}(t)$.



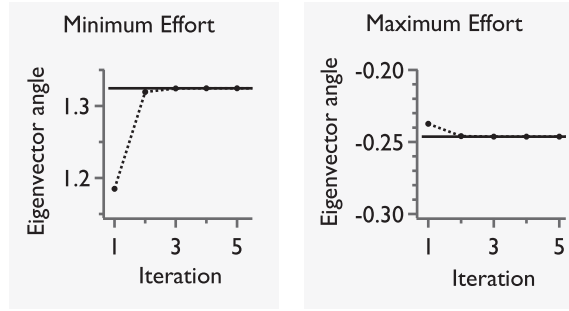
Below, we replot for convenience the phase portrait of the two solutions that result from applying $u_{\min}(t)$ and $u_{\max}(t)$. Notice the nonlocal nature of the $u_{\max}(t)$-produced trajectory. More direct trajectories are possible but would require even more effort.



b. We compute successive powers of $\boldsymbol{P}\,\boldsymbol{v}_0$, normalizing the result to a unit vector after each application of the Gramian. We start from the (arbitrarily chosen) vector $\boldsymbol{v}_0 = \binom{1}{0}$, which corresponds to $\theta = 0$. At left, we plot the angle of the unit vector, which rapidly converges (in about $N = 3$ iterations) to that of the

associated eigenvector of $\boldsymbol{P}$. The rapid convergence arises because the ratio of eigenvalues is almost 30, so that the deviation decreases as $\sim 30^{-N}$.

Similarly, we find the direction of maximum effort by looking at the eigenvector for the largest eigenvalue of $\boldsymbol{P}^{-1}$. Again, convergence is very rapid. Notice that the directions differ by $1.32\ldots - (-0.25\ldots) = \pi/2 = 90°$, as they must for the eigenvectors of a symmetric matrix.



c. The calculations follow those from the two-dimensional case. The Gramian for $\tau = 3$ is

$$\boldsymbol{P} \approx \begin{pmatrix} 0.045825 & 0.0798624 & 0.042819 \\ 0.0798624 & 0.241266 & 0.0679964 \\ 0.042819 & 0.0679964 & 0.0410984 \end{pmatrix},$$

with eigenvalues $\approx$ {0.29, 0.034, 0.00031} and directions of minimum and maximum effort (over the unit sphere) given by

$$\hat{\boldsymbol{n}} \approx \begin{pmatrix} -0.34 \\ -0.90 \\ -0.29 \end{pmatrix}, \quad \begin{pmatrix} -0.73 \\ -0.048 \\ -0.69 \end{pmatrix},$$

The ratio $\mathcal{E}_{\max}/\mathcal{E}_{\min} \approx 979$. Note that computing the infinite-time Gramian is easier, as you need only solve the Lyapunov equation. This gives a condition number ≈ 948, which is not so different, since $\tau > 1$. The condition number becomes larger as τ is reduced.

d. The efforts for $n = 2$ and $n = 3$ in the *easiest* directions are $\approx$ 4.5 and 3.4, respectively. The efforts in the *hardest* directions are $\approx$ 127 and 3361 respectively.

14.13 Effective controllability of an n-chain. (Continuation of Problem 14.6.)

a. Show that the control Gramian $\boldsymbol{P}(\tau)$ has elements $\boldsymbol{P}(\tau)_{ij} = \frac{(\tau)^{i+j-1}}{(i+j-1)\,(i-1)!\,(j-1)!}$.
b. Evaluate the eigenvalues of $\boldsymbol{P}$ numerically, for $n = 1, 2, \ldots, 20$, and show that the condition number increases exponentially as $\sim \mathrm{e}^{8.4n}$.

Solution.

a. From the pattern in (a), we deduce that $\boldsymbol{A}^n = \boldsymbol{0}$, and, thus,

$$\mathrm{e}^{\boldsymbol{A}t} = \mathbb{I} + t\boldsymbol{A} + \frac{t^2}{2!}\boldsymbol{A}^2 + \cdots + \frac{t^{n-1}}{(n-1)!}\boldsymbol{A}^{n-1} .$$

For $n = 4$, this is, explicitly,

$$\mathrm{e}^{\boldsymbol{A}t} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ t & 1 & 0 & 0 \\ \frac{t^2}{2!} & t & 1 & 0 \\ \frac{t^3}{3!} & \frac{t^2}{2!} & t & 1 \end{pmatrix}$$

We then see that

$$\mathrm{e}^{\boldsymbol{A}t}\,\boldsymbol{B} = \begin{pmatrix} 1 \\ t \\ \frac{t^2}{2!} \\ \vdots \\ \frac{t^{n-1}}{(n-1)!} \end{pmatrix} .$$

The ij element of the integrand is then

$$\frac{t^{i-1}}{(i-1)!}\,\frac{t^{j-1}}{(j-1)!} = \frac{t^{i+j-2}}{(i-1)!\,(j-1)!} ,$$

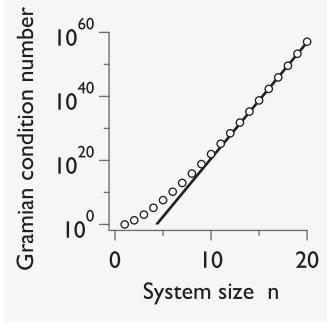
implying, after integration over $t \in [0, \tau]$, that

$$\boldsymbol{P}(\tau)_{ij} = \frac{\tau^{i+j-1}}{(i+j-1)\,(i-1)!\,(j-1)!} .$$

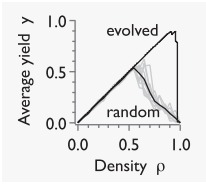
b. We calculate the eigenvalues of the Gramian $\boldsymbol{P}(\tau)$ numerically and plot the condition number vs. system size. Asymptotically, it scales as

$$\sim 1.7 \times 10^{-16}\,\mathrm{e}^{8.4n} ,$$

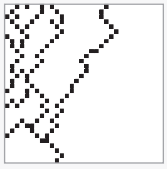
with n the system size. See plot below.



14.14 Some like it HOT. Consider a simple model of a forest management in the face of forest fires. Plant trees on a square lattice of side N, with site probability



random, ρ=0.6



evolved, ρ=0.9

ρ. The forest is subject to "sparks" that have a source above and left of the forest and fall on the ij site in the forest with probability $p_{ij} \equiv p_i p_j$, where $p_x \propto 2^{-(m_x+x/N)/\sigma_x^2}$. Choose $m_i = 1$, $\sigma_i = 0.4$, $m_j = 0.5$, and $\sigma_j = 0.2$. If a spark falls on an unoccupied site, there is no damage. If it falls on an occupied site, all contiguous occupied sites are burned. The goal is to maximize the yield $y(\rho)$, the average number of trees left after a fire.

a. For randomly planted trees and $N \to \infty$, show that $y(\rho) = \rho - P_\infty^2$, where P_∞ is the probability that a lattice site is in the infinite *percolation* cluster.
b. Write code to find the yield plotted at left (top), for $N = 32$. Ten individual trials are shown in light gray, and the average in thick black. Hints: Find contiguous lattice sites of ones (the trees) by finding morphological components of a binary image and then counting their size. Find this number for every lattice site and average over the p_{ij} to determine the expected loss from a spark.
c. To increase the yield, *evolve* a design for planting trees as follows: When going from a density $\rho = n/N^2$ to $(n + 1)/N^2$, explore D possibilities for where to plant the next tree. For each candidate position, calculate the average loss and then choose the position that minimizes the loss. $D = 1$ is equivalent to the random-forest case. The figure at left is for $D = N = 32$ (cf. thin black line for yield).

This simple design procedure naturally leads to *firebreaks* of unplanted sites that stop fires from spreading too far. See the lines of unplanted sites in black at left (bottom). The firebreaks become even more clearly organized if one optimizes over all free sites instead of only up to D sites. We emphasize that this organization arises from the repeated evolutionary cycle of trial planting and evaluation and not from any kind of self-organization. The distribution of fire sizes turns out to be approximately a power law, although that fact does not play an important role in the organization (*function*) of the tree-planting algorithm. For more, see Carlson and Doyle (2000).

Solution.

a. Below the percolation threshold of $\rho_c \approx 0.59$, there are only finite-size clusters of trees on an infinite lattice. At ρ_c, an infinite cluster first forms and its size increases with ρ. In the limit of large lattices, only the sparks hitting the infinite cluster will decrease the yield by a number of $O(N^2)$. Others have negligible impact. Let P_∞ be the probability to hit the infinite cluster. Then the yield

$$y(\rho) = (1 - P_\infty)\,\rho + P_\infty(\rho - P_\infty) = \rho - P_\infty^2 .$$

The first term is the probability to miss the infinite cluster and retain full yield. The second term is the probability to hit the infinite cluster and reduces the yield by the size of the infinite cluster. These effects are "softened" in a finite lattice. The sharp, non-analytic transition for the yield at ρ_c between isolated

clusters (no infinite cluster) and above is rounded because of edge effects and finite sampling.

b. See book website for code. As the hint suggests, you can ease the trickiest part of the programming by leveraging standard routines for image processing of a binary image. A common task is to find all connected objects, and most image-processing packages have a routine to find all such sites. Then count the number in all such objects and return that number to each lattice site. Then the average loss is just the sum of the element-by-element product of this matrix with the matrix of probabilities for a spark to fall on each site. (In *Mathematica*, the `MorphologicalComponents` and `ComponentMeasurements` commands can carry out the necessary operations.)

c. The evolution algorithm is straightforward to program, particularly if efficiency is not a major goal. On my laptop, the code for the $N = D = 32$ case took a bit less than 4 minutes for a single iteration (all I did), running in *Mathematica*. No doubt this can be improved! See book website for code.

The fall in yield for $\rho \gtrsim 0.95$ results mainly from "filling in" the firebreaks. The firebreaks are essentially one-dimensional curves in the two-dimensional forest and thus have zero measure in the limit of an infinite forest. In that case, we expect the yield to go to one. Of course, in a finite forest, the yield is maximized at a lower value (still quite close to 1, as our 0.95 result for $N = D = 32$ shows). We did not have space to include it, but another interesting feature to explore is that the yield decreases sharply if the distribution p_{ij} changes. For example, the source of sparks might move to a different corner. The firebreaks were optimized for the given distribution p_{ij}. Such an optimization implicitly assumes that the statistics are stationary over a relatively long time (long enough to accumulate the statistics to estimate the form of the distribution, for example). Finally, we would expect a similar fall in yield if the firebreaks ever have a defect. The increase in yield comes from limiting the size of the largest cluster that is likely to form. That is why the area in the upper left, which is where most sparks fall, has a high fraction of unplanted sites.

15 Limits to Control

Problems

15.1 Causality and Kramers-Kronig for a first-order system. Consider a first-order, low-pass-filter system with transfer function $G(s) = \frac{1}{1+s}$. Using contour integration,

a. invert the Fourier transform of $G(\mathrm{i}\omega)$ and verify that the impulse response function $G(t)$ (the Green function) is causal;
b. verify the Kramers-Kronig relations for G.

Solution.

a. The response function $G(t)$ is given by the inverse Fourier transform:

$$G(t) = \int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} \left(\frac{1}{1 + \mathrm{i}\omega} \right) \mathrm{e}^{\mathrm{i}\omega t} \, .$$

We can do this integral by contour integration about the closed contour γ. We note that there is a single pole at $\omega = \mathrm{i}$, whose residue is

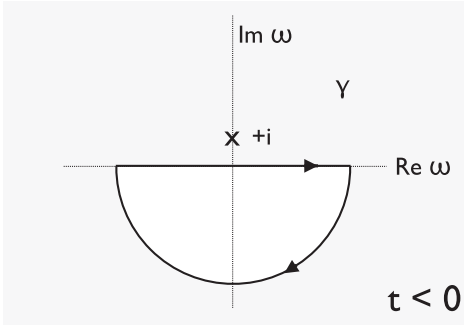
$$\mathrm{Res}(i) = \frac{1}{2\pi} \left(\frac{\mathrm{e}^{\mathrm{i}(\mathrm{i})t}}{\mathrm{i}} \right) = \frac{\mathrm{e}^{-t}}{2\pi \mathrm{i}} \, .$$

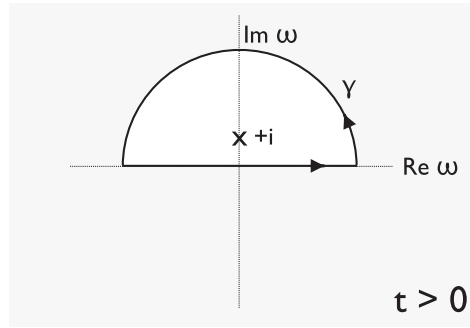
- For $t < 0$, we close the contour in the lower half of the complex ω plane. Since there are no poles inside γ, the integral = 0.
- For $t > 0$, we close the contour in the upper half of the complex ω plane. The residue theorem then implies that the integral is e^{-t}.

Putting the two results together, we have

$$G(t) = \mathrm{e}^{-t} \, \theta(t) \, ,$$

where the theta function explicitly shows that the response is causal.





b. The real and imaginary parts of the response functions are

$$G'(\omega') = \mathrm{Re}\left(\frac{1}{1+\mathrm{i}\omega'}\right) = \frac{1}{1+\omega'^2} \qquad G''(\omega') = \mathrm{Im}\, G(\mathrm{i}\omega') = \frac{-\omega'}{1+\omega'^2}\,.$$

Then, the second Kramers-Kronig relation in Eq. (15.8) asks us to verify that

$$\begin{aligned} G''(\omega) &= \frac{2\omega}{\pi}\,\mathrm{P}\int_0^\infty d\omega'\,\frac{G'(\omega')}{\omega'^2-\omega^2} \\ &= \frac{2\omega}{\pi}\,\mathrm{P}\int_0^\infty \mathrm{d}\omega'\,\frac{1}{(1+\omega'^2)(\omega'^2-\omega^2)} \\ &= \frac{\omega}{\pi}\oint_\gamma \mathrm{d}\omega'\,\frac{1}{(1+\omega'^2)(\omega'^2-\omega^2)} \\ &= \frac{\omega}{\pi}\,(2\pi\,\mathrm{i})\,\frac{1}{\mathrm{i}[1-(-1)](-1-\omega^2)} \\ &= -\frac{\omega}{1+\omega^2}\,. \end{aligned}$$

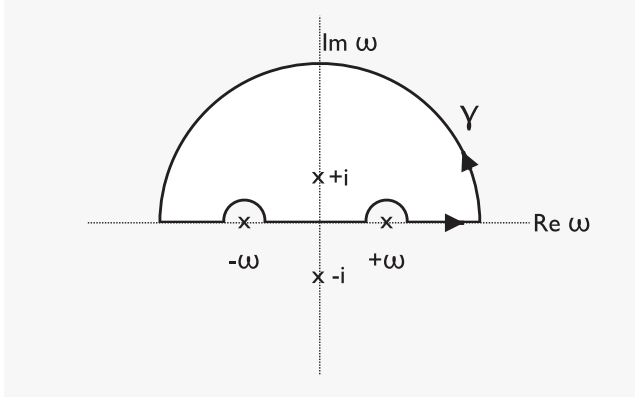
Similarly,

$$\begin{aligned} G'(\omega) &= -\frac{2}{\pi}\,\mathrm{P}\int_0^\infty \mathrm{d}\omega'\,\frac{\omega' G''(\omega')}{\omega'^2-\omega^2} \\ &= +\frac{2}{\pi}\,\mathrm{P}\int_0^\infty \mathrm{d}\omega'\,\frac{\omega'^2}{(1+\omega'^2)(\omega'^2-\omega^2)} \\ &= +\frac{1}{\pi}\oint_\gamma \mathrm{d}\omega'\,\frac{\omega'^2}{(1+\omega'^2)(\omega'^2-\omega^2)} \\ &= \frac{1}{\pi}\,(2\pi\,\mathrm{i})\,\frac{-1}{2\,\mathrm{i}(-1-\omega^2)} \\ &= \frac{1}{1+\omega^2}\,. \end{aligned}$$

Here, the contour γ is chosen as depicted below. The poles are at $\pm\,\mathrm{i}$, and we can close the contour either in the top or bottom planes. Note the factor of $\frac{1}{2}$ that arises when we extend the range from $(0,\infty)$ to $(-\infty,\infty)$.

The last thing is to show that the contributions of the big and little semicircles go to zero as the radii go to ∞ and 0, respectively. For the big semicircle, the integrand $\sim \omega'^{-4}$, which clearly converges. The integral around the little

semicircle is more subtle. It does not go to zero as $r \to 0$ when $\omega' = \omega + r\,\mathrm{e}^{\mathrm{i}\theta}$. However, the contribution turns out to be *odd* in ω, meaning that the sum of the contributions from the two semicircles (at $\pm\omega$) vanishes.



15.2 Sensitivity function for oscillator. Verify numerically that the waterbed integral of a second-order system with proportional feedback gain is zero. That is, show for $G(s) = \frac{1}{1+2\zeta\omega+\omega^2}$ and $K(s) = K_\mathrm{p}$ that $\int_0^\infty \mathrm{d}\omega \ln|S(\mathrm{i}\omega)| = 0$. Plot for $K_\mathrm{p} = 1$ and $\zeta = 0.5$, and show its numerical integral = 0.

Solution.

We have

$$L(s) = \frac{K_\mathrm{p}}{1 + 2\zeta s + s^2}\,,$$

which implies

$$S(s) = \frac{1 + 2s\zeta + s^2}{1 + K_\mathrm{p} + 2s\zeta + s^2}$$

$$\implies \quad S(\mathrm{i}\omega) = \frac{1 + 2\,\mathrm{i}\omega\zeta - \omega^2}{1 + K_\mathrm{p} + 2\,\mathrm{i}\omega\zeta - \omega^2}$$

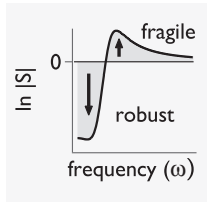
$$\implies \quad |S(\mathrm{i}\omega)| = \sqrt{\frac{(1-\omega^2)^2 + 4\zeta^2\omega^2}{(1 + K_\mathrm{p}^2 - \omega^2)^2 + 4\zeta^2\omega^2}}\,.$$

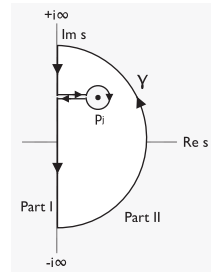
We then evaluate numerically the integral and confirm that

$$\int_0^\infty \mathrm{d}\omega \ln|S(\mathrm{i}\omega)| = 0\,, \qquad \forall \zeta > 0\,, K_\mathrm{p} > 0\,.$$

The plot at left shows the waterbed plot for $K_\mathrm{p} = 1$ and $\zeta = 0.5$.

Note that $\ln|S(\mathrm{i}\omega)| \to K_\mathrm{p}/\omega^2$ as $\omega \to \infty$, implying that the integral converges relatively slowly, as K_p/ω. We can see this in the plot below.





15.3 Bode's waterbed theorem. Derive Eqs. (15.22) and (15.25). For $S(s)$ analytic:

a. Show that Re $[\ln S(\mathrm{i}\omega)]$ is an even function of ω and Im $[\ln S(\mathrm{i}\omega)]$ is odd.
b. Fill in the missing steps leading to Eq. (15.24).
c. Show that Part II, evaluated along a circle of radius $R \to \infty$, vanishes if $L(s)$ is of relative order 2 or greater.
d. Deal with RHP poles using the contour at right, with a similar detour for each pole p_j. Evaluate the added contributions to prove Eq. (15.25).

Solution.

a. We have

$$\ln S = \ln|S| + \mathrm{i}\arg S\,.$$

Thus,

$$\mathrm{Re}\,[\ln S(\mathrm{i}\omega)] = \ln|S(\mathrm{i}\omega)|\,.$$

We can write $S(\mathrm{i}\omega) = u(\omega) + \mathrm{i}v(\omega)$, where u and v are real functions of the frequency ω. Thus, $\ln|S| = \frac{1}{2}\ln\left(u^2+v^2\right) + \mathrm{i}\tan^{-1}\frac{v}{u}$ is even in ω, and Im $[\ln S(\mathrm{i}\omega)]$ is odd (since arc tan is odd).

Another approach to this problem is to note that for any real function $f(t)$, the Fourier transform $F(\omega) = F^*(-\omega)$. Writing $F(\omega) = F'(\omega) + \mathrm{i}F''(\omega)$ immediately shows that the real part, $F'(\omega)$ is even and the imaginary part, $F''(\omega)$ is odd. Applying this observation to $S(\mathrm{i}\omega)$ and $\ln S(\mathrm{i}\omega)$ gives the result.

b. Using the result from (a), we have that

$$\int_R^{-R} \mathrm{d}\omega \ln S(\mathrm{i}\omega) = -\int_{-R}^{R} \mathrm{d}\omega \ln S(\mathrm{i}\omega) = -2\int_0^R \mathrm{d}\omega \ln|S(\mathrm{i}\omega)|\,,$$

because the even contribution from Re $[\ln S]$ contributes a factor of 2 and the odd contribution cancels out. (We are integrating over a domain symmetric about $\omega = 0$.)

c. Part II is

$$\int_\mathrm{II} \mathrm{d}s \ln S = \int_\mathrm{II} \mathrm{d}s \ln\frac{1}{1+L} \approx \int_\mathrm{II} \mathrm{d}s\,[-L(s)]\,,$$

since L is small for $s = R\,\mathrm{e}^{\mathrm{i}\theta}$ and $R \to \infty$. Assuming that $L \lesssim s^{-2}$, we have

$$\int_{\mathrm{II}} \mathrm{d}s\, L(s) \sim \frac{1}{R^2} R \int_0^{\pi} \mathrm{d}\theta\, \mathrm{d}\left(\mathrm{e}^{\mathrm{i}\theta}\right) \to 0\,,$$

as $R \to \infty$. Thus, Part II goes to 0 and Part I becomes an integral $\omega \in (0, \infty)$, which proves the theorem.

d. Since the new contours do not include the poles p_j, the integral around the modified γ continues to be zero. Because of the pole at p_j, we have to add a branch cut starting from each pole and going horizontally to the left. Across the branch cut, the argument jumps by 2π.

We now break the new part of the contour for pole p_j into three parts: "$\circ$", the little circle of radius $r \to 0$ that goes clockwise around the pole, "$\to$", the straight line from the imaginary axis to p_j (after the radius of the little circle, $r \to 0$), and "$\leftarrow$", which goes from p_j to the imaginary axis.

We first note that

$$\int_{\circ} \mathrm{d}s \ln S \approx \int_{\pi}^{-\pi} \mathrm{d}\theta\, (\ln r)(r)\, \mathrm{e}^{\mathrm{i}\theta} \sim r \ln r \to 0\,,$$

as $r \to 0$. Thus, the contribution around the little circle vanishes in the limit of small r.

The contributions $\int_{\to}$ and $\int_{\leftarrow}$ are evaluated by noting that the magnitude of $\ln S$ is equal to that of its corresponding partner. Only the argument is different, having shifted by 2π. Thus,

$$\int_{\to} + \int_{\leftarrow} = (2\pi\,\mathrm{i})(\mathrm{Re}\ p_j)\,.$$

The factor $\mathrm{Re}\ p_j$ is just the length of either contour.

Adding up all the contours then gives

$$-2\,\mathrm{i} \int_0^{\infty} \mathrm{d}\omega \ln |S(\mathrm{i}\omega)| + 2\pi\,\mathrm{i}\ \mathrm{Re}\ p_j = 0\,.$$

We can extend the argument to all the poles p_j, with a keyhole contour and branch cut for each, to find

$$\int_0^{\infty} \mathrm{d}\omega \ln |S(\mathrm{i}\omega)| = \pi \sum_j \mathrm{Re}\ p_j = \pi \sum_j p_j\,.$$

The last identity is true because the poles come in complex-conjugate pairs. This proof is adapted from Åström and Murray (2008).

15.4 Waterbed theorem for relative degree 1 systems. For a stable, first-order loop transfer function $L(s)$ with $L(s) \to \alpha/s$ as $s \to \infty$, show that the Bode sensitivity integral is $\int_0^{\infty} \mathrm{d}\omega \ln |S(\mathrm{i}\omega)| = -\frac{\pi}{2}\alpha$ and that, for degree ≥ 1, $\int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} \ln |S(\mathrm{i}\omega)| = \sum_j p_j - \frac{1}{2} \lim_{s\to\infty} s\, L(s)$. Hint: in deriving Eq. (15.22), the contribution of the

big semicircle no longer vanishes. Uncertainties in the dynamics and delays mean that most practical systems have unmodeled high-frequency dynamics that are effectively higher order. The first-order case has relatively few practical consequences.

Solution.

As in the original derivation, $\oint_\gamma \mathrm{d}s \ln S(s) = 0$ for the half disk of radius R, as $R \to \infty$. Part I is unchanged. For Part II, we have

$$\int_{\mathrm{II}} \mathrm{d}s \ln S(s) = \int_{\mathrm{II}} \mathrm{d}s \ln\left(\frac{1}{1+L}\right) \approx -\int_{\mathrm{II}} \mathrm{d}s\, L(s) = -\alpha \int_{\mathrm{II}} \frac{ds}{s}\,.$$

For the large semicircle of radius R, we have $s = R\,\mathrm{e}^{\mathrm{i}\theta}$ and $ds = \mathrm{i}\, s\, d\theta$, giving

$$\lim_{R\to\infty} \int_{\mathrm{II}} \mathrm{d}s \ln S(s) = -\alpha \int_{-\pi/2}^{\pi/2} \mathrm{d}\theta\, \mathrm{i} = -\,\mathrm{i}\alpha\pi\,.$$

The relation Part I + Part II = 0 then gives

$$-2\,\mathrm{i}\int_0^\infty \mathrm{d}\omega \ln|S(\mathrm{i}\omega)| - \mathrm{i}\alpha\pi = 0 \qquad \Longrightarrow \qquad \int_0^\infty \mathrm{d}\omega \ln|S(\mathrm{i}\omega)| = -\frac{\pi}{2}\alpha\,,$$

For the general expression, let $\alpha = \lim_{s\to\infty} s\,L(s)$. Then

$$\int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} \ln|S(\mathrm{i}\omega)| = \sum_j p_j - \frac{1}{2}\lim_{s\to\infty} s\,L(s)$$

applies for $\nu \geq 1$ for an L with unstable poles p_j that are stabilized in closed loop.

15.5 Waterbed theorem for T. Prove Eq. (15.26). Hint: Define $\bar{s} = 1/s$ and $\bar{L}(s) \equiv 1/L(1/s) = 1/L(\bar{s})$. From Åström and Murray (2008).

Solution.

Following the hint, we write

$$T(\bar{s}) = \frac{L(\bar{s})}{1+L(\bar{s})} = \frac{1}{1+L^{-1}(\bar{s})} = \frac{1}{1+\bar{L}(s)} \equiv \bar{S}(s)\,.$$

We thus apply the Bode waterbed theorem to $\bar{S}(\mathrm{i}\omega)$, which gives

$$\int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} \ln|\bar{S}(\mathrm{i}\omega)| = \sum_j \bar{p}_j\,.$$

We rewrite this in terms of $T(\bar{s})$, remembering that we need to change variables in the integral to express it in terms of s. Thus, we redefine frequency by $\bar{\omega} = 1/\omega$. This gives rise to

$$\mathrm{d}\omega = -\frac{\mathrm{d}\bar{\omega}}{\bar{\omega}^2}\,.$$

In rewriting the integral, the minus sign cancels out because the limits are also flipped. We also note that the poles of S are the zeros of T, meaning that $\bar{p}_j \to 1/z_j$. Taking into account the change of variable, we have then

$$\int_{-\infty}^{\infty} \frac{\mathrm{d}\bar{\omega}}{2\pi} \ln|T(\mathrm{i}\bar{\omega})| = \sum_j \frac{1}{z_j} .$$

Then rename the dummy variable $\bar{\omega}$ as ω.

15.6 Waterbed theorem for discrete dynamics. Derive Eq. (15.27): for an open-loop transfer function $L(z)$ with relative degree $\nu \geq 1$ and unstable poles $|p_j| > 1$, we have $\int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \ln|S(\mathrm{e}^{\mathrm{i}\omega})| = \sum_j \ln|p_j|$. Assume that the gain K of L stabilizes the closed-loop system. Hint: Use Jensen's relation, Eq. (A.51), to show $I(p) \equiv \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \ln\left|\mathrm{e}^{\mathrm{i}\omega} - p\right| = \ln|p|$ for $|p| > 1$ and 0 otherwise. Then write $L(z)$ in pole-zero form.

Solution.

We write the loop gain in terms of its poles and zeros:

$$L(z) = K \frac{\prod_{j'=1}^{m}(z - z_{j'})}{\prod_{j=1}^{n}(z - p_j)} ,$$

where the loop gain K is chosen so that $S = \frac{1}{1+L}$ is stable and where the relative degree $\nu = n - m \geq 1$. Since S is stable, we can write it in pole-zero form as

$$S(z) = \frac{1}{1 + K\frac{\prod_{j'=1}^{m}(z-z_{j'})}{\prod_{j=1}^{n}(z-p_j)}} = \frac{\prod_{j=1}^{n}(z - p_j)}{\prod_{j'=1}^{n}(z - r_{j'})} ,$$

where $r_{j'}$ denote the closed-loop poles. For our purposes, we do not care where they are, except that, since S is stable, we know that $|r_{j'}| < 1$, for all j'. Note that S has $\nu = 0$.

Substituting the expression for S and using the suggested integral $I(p)$ then gives

$$\begin{aligned}
&= \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \ln\left|\frac{\prod_{j=1}^{n}(z - p_j)}{\prod_{j'=1}^{n}(z - r_{j'})}\right|_{z=\mathrm{e}^{\mathrm{i}\omega}} \\
&= \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \left[\sum_{j=1}^{n} \ln|\,\mathrm{e}^{\mathrm{i}\omega} - p_j| - \sum_{j'=1}^{n} \ln|\,\mathrm{e}^{\mathrm{i}\omega} - r_{j'}|\right] \\
&= \sum_{j=1}^{n_u} \ln|p_j| ,
\end{aligned}$$

where n_u is the number of unstable poles in $L(z)$. We note that the $I(p)$ identity shows that all the poles $r_{j'}$ do not contribute because they are, by hypothesis, stable. Likewise, the stable parts of $L(z)$ do not contribute. If $L(z)$ is stable to begin with, then we have shown that

$$\int_0^{\pi} \mathrm{d}\omega \ln|S(\mathrm{e}^{\mathrm{i}\omega})| = 0 .$$

The $I(p)$ identity is a special case of Jensen's relation, Eq. (A.51):

$$\int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \ln|f(\mathrm{e}^{\mathrm{i}\omega})| = \ln|f(0)| + \sum_{j=1}^{m} \ln|p_j| - \sum_{j'=1}^{n} \ln|z_{j'}|\,,$$

where p_j are the poles and $z_{j'}$ the zeros *inside* the unit circle. Choose $f(z) = z - p$, which has a *zero* at p. Then,

$$\begin{aligned} I(p) &= \frac{1}{2} \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \ln\left|\mathrm{e}^{\mathrm{i}\omega} - p\right|^2 \\ &= \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \ln\left|\mathrm{e}^{\mathrm{i}\omega} - p\right| . \end{aligned}$$

- $|p| < 1$. Jensen implies that $I(p) = \ln|p| + 0 - \ln|p| = 0$.
- $|p| > 1$. Jensen implies that $I(p) = \ln|p| + 0 - 0 = \ln|p|$.

Finally, we give an alternate, more direct proof for the case of stable L, where all poles are inside the unit circle, implying that $\ln S$ is analytic outside the unit circle. We write

$$\begin{aligned} \int_{-\pi}^{\pi} \mathrm{d}\omega \ln|S(z)|^2 &= \oint \frac{\mathrm{d}z}{\mathrm{i}z} \ln|S(z)|^2 \\ &= \oint \frac{\mathrm{d}z}{\mathrm{i}z} \left[\ln|S(z)| + \ln\left|S(z^{-1})\right|\right] \\ &= 2 \oint \frac{\mathrm{d}z}{\mathrm{i}z} \ln|S(z)| \end{aligned}$$

where, in the last step, we substitute $z' = z^{-1}$ and remember that the limits reverse, absorbing the minus sign from $\mathrm{d}z' = -\,\mathrm{d}z\,/z^2$.

We next deform the unit circle contour of the integral to a larger circle of radius $R \to \infty$ via the Cauchy Integral Theorem. For z on this enlarged contour, we have

$$\ln S(z) = \ln \frac{1}{1+L} \approx \ln(1-L) \approx -L\,.$$

Then, since $L(R\,\mathrm{e}^{\mathrm{i}\omega}) \sim R^{-\nu}$ as $R \to \infty$, we have

$$\oint \frac{1}{\mathrm{i}z}\,\mathrm{d}z\,\ln[S(z)] \approx -\lim_{R\to\infty} \int_0^{2\pi} \mathrm{d}\omega\, L(R\,\mathrm{e}^{\mathrm{i}\omega}) = 0\,.$$

This proves the waterbed theorem for the discrete, minimum-phase case and makes clear the role that $\nu \geq 1$ plays.

Although not asked for in the problem, there is a straightforward generalization to loop transfer functions L of relative degree $\nu = 0$. Let $\alpha = \lim_{z\to\infty} L(z)$. Then

$$L(z) = \alpha + \beta z^{-1} + O(z^{-2})\,,$$

and

$$S(z) = \frac{1}{1+L} = \frac{1}{1+\alpha+\beta z^{-1}+\cdots}$$
$$= \frac{1}{1+\alpha}\left[1-\left(\frac{\beta}{1+\alpha}\right)z^{-1}+\cdots\right].$$

Then

$$\ln S(z) \approx \ln\frac{1}{1+\alpha} - \left(\frac{\beta}{1+\alpha}\right)z^{-1}+\cdots,$$

and the Bode relation generalizes to

$$\int_{-\pi}^{\pi}\frac{\mathrm{d}\omega}{2\pi}\ln|S(\mathrm{e}^{\mathrm{i}\omega})| = \sum_{j=1}^{n_u}\ln|p_j| - \ln|1+\alpha|.$$

15.7 One-dimensional, discrete dynamics: Bode's waterbed theorem. For the dynamics of Example 15.1, reproduce the graphs in Figure 15.3 and show that the variance of observations is given by $\langle y^2\rangle = \nu^2/[1-(a-K)^2]$.

Solution.

The graphs are simply plots of

$$\log\left|S(\mathrm{e}^{\mathrm{i}\omega})\right|^2 = \log\left(1+a^2-2a\cos\omega\right).$$

We can derive the expression for the variance by manipulating correlation functions and taking advantage of the time invariance of the equations. From the equations of motion,

$$y_{k+1} = (a-K)y_k + \nu_k,$$

which implies the correlations

$$\langle y\,\nu_{-1}\rangle = \nu^2$$
$$\left\langle y^2\right\rangle = (a-K)\langle y\,y_{-1}\rangle + \langle y\,\nu_{-1}\rangle$$
$$\langle y\,y_{-1}\rangle = (a-K)\left\langle y^2\right\rangle.$$

Substituting and then solving for $\left\langle y^2\right\rangle$ gives

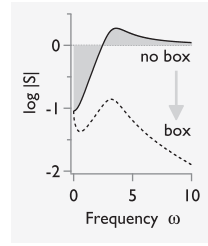
$$\left\langle y^2\right\rangle = \frac{\nu^2}{1-(a-K)^2}.$$

A faster solution is to square each side of the equation for y_{k+1} and average:

$$\left\langle y_{k+1}^2\right\rangle = (a-K)^2\left\langle y_k^2\right\rangle + \nu^2,$$

where we use the fact that ν_k and y_k are independent, so that $\langle y_k\,\nu_k\rangle = 0$. After initial transients have decayed, the statistics should be stationary, so that $\left\langle y_{k+1}^2\right\rangle = \left\langle y_k^2\right\rangle = \left\langle y^2\right\rangle$. Solving for $\left\langle y^2\right\rangle$ also gives our result.

15.8 Temperature control and the waterbed theorem. As a simplified temperature response let $G(s) = \frac{1}{(1+s)^2}$. Add a PID controller, $K(s) = K_\mathrm{p} + \frac{K_\mathrm{i}}{s} + K_\mathrm{d}\left(\frac{s}{1+s/\omega_f}\right)$. Place the system in a box of thickness ℓ and thermal diffusivity D, whose temperature transfer function for high frequencies is approximately $G_\mathrm{box}(s) = \mathrm{e}^{-(\ell/\sqrt{D})\sqrt{s}}$ (see Problem 2.5). Use $\{D, \ell, K_\mathrm{p}, K_\mathrm{i}, K_\mathrm{d}, \omega_f\} = \{1, 2, 10, 3, 7, 10\}$.



a. With no box, compute the sensitivity function $S = \frac{1}{1+L}$, with $L = K(s)\,G(s)$. Confirm the "no box" Bode integral plot at right.
b. Add the insulating box response and confirm the "box" plot at right.
c. Investigate the response to an impulse disturbance at the output, using a Padé approximant to the box transfer function or the inverse Laplace transform of $G_\mathrm{box}(s)$: $G(x = \ell, t) = \frac{\ell}{\sqrt{4\pi D t^3}} \exp\left\{\left[-\frac{\ell^2}{4Dt}\right]\right\}$. Plot the disturbance as filtered by the box, $G(x = \ell, t)$, along with the closed-loop response that it provokes. Plot, too, the responses to a step input, showing the case of no control (just the system), PID control, and PID control augmented with a first-order feedforward filter between the reference signal and the controller input that eliminates the overshoot of the simple PID controller.

Solution.

a. The sensitivity function is

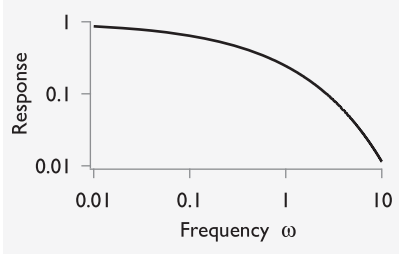
$$S(s) = \frac{1}{1+L} = \frac{(1+s)^2}{(1+s)^2 + K(s)},$$

with

$$K(s) = K_\mathrm{p} + \frac{K_\mathrm{i}}{s} + K_\mathrm{d}\left(\frac{s}{1+\frac{s}{\omega_f}}\right).$$

Then we compute, numerically, $\ln|S(\mathrm{i}\omega)|$.

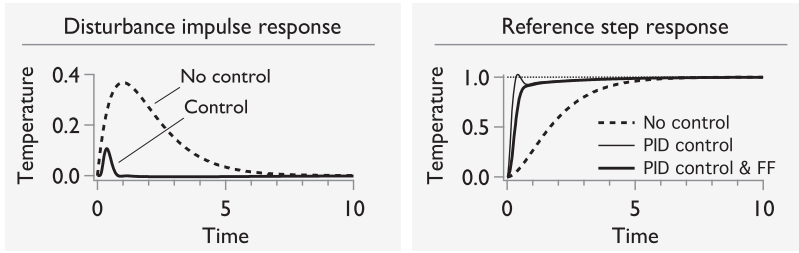
b. The Bode response, $|G_\mathrm{box}(\mathrm{i}\omega)| = |\,\mathrm{e}^{-2\sqrt{\mathrm{i}\omega}}\,| = \mathrm{e}^{-\sqrt{2\omega}}$ is shown below. Recall that $\sqrt{\mathrm{i}} = \pm\frac{1}{\sqrt{2}}(1+\mathrm{i})$. Select the positive root to make $|G| \to 0$ as $\omega \to \infty$.



See book for plot of $|S(\mathrm{i}\omega)|\,|G_\mathrm{box}(\mathrm{i}\omega)|$.

c. The time responses are generated by using a linear response routine for $G(s)$ in Mathematica with the appropriate driving function. We note that the PID parameters chosen give reasonable reference control (the step response) and reasonable disturbance rejection (using the passive insulation from the box). It is instructive to play around with other combinations, too.

The solution shown uses a feedforward filter with transfer function $G_{\text{ff}}(s) = \frac{1}{1+s/\omega_{\text{ff}}}$, with $\omega_{\text{ff}} = 10$. No attempt was made to systematically tune the PID and other controller parameters, and you can probably find better ones. Here are the plots:



15.9 Acausal control of a first-order system. In reference to Example 15.2:

a. Show that the acausal feedforward input $u(t)$ given leads to the desired control. Hint: Solve the time-domain equations for $u(t < 0)$, then $u(t) \approx 0$, then $u(t) > 0$.

b. Explain physically (in words) how this solution works. How can a non-zero input $u(t)$ for $t < 0$ nonetheless produce a zero output? If that output is zero, where does the step response come from?

Solution.

a. First the math. For a transfer function

$$G(s) = \frac{1 - \tau s}{1 + \tau s},$$

the corresponding time domain equation is

$$\dot{y} + \tau^{-1} y = -\dot{u} + \tau^{-1} u\,.$$

We substitute the desired solution $y(t) = \theta(t)$, to find a first-order differential equation for $u(t)$:

$$-\dot{u} + \tau^{-1} u = \tau^{-1}\theta(t) + \delta\,(t)\,,$$

where the delta function arises from differentiating $\theta(t)$.

We begin by noting that an "obvious" solution for $t > 0$ is $u(t) = 1$. This clearly both satisfies the differential equation for positive time and meets the final condition $y = 1$ at $t = +\infty$.

To solve near $t = 0$, we integrate from $-\varepsilon$ to $+\varepsilon$. All terms that are continuous or have a finite jump continuity, such as $\theta(t)$, give a finite output times the interval, 2ε. Taking the limit $\varepsilon \to 0$ eliminates those terms. Integrating the other terms then gives,

$$-u(0^+) + u(0^-) = 1\,, \qquad \Longrightarrow \qquad u(0^-) = 2\,.$$

In the last step, we use the previous result that $u(0^+) = 1$, since we evaluate it for $t > 0$.

In the last regime, $t < 0$, we solve

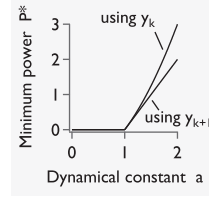
$$-\dot{u} + \tau^{-1}u = 0\,, \qquad u(0^-) = 2\,,$$

for negative times. This has a solution in this regime of $u(t) = 2\,\mathrm{e}^{t/\tau}$. If integrating backwards in time seems strange to you, change variables $t \to -t$ and solve forward in the new time variable.

b. Now let us try to understand more physically what is going on. The first question is how a non-zero input $u(t)$ leads to a zero output $y(t)$ for negative times. Well, this is the meaning of a zero! That is, the zero dynamics implies a signal $u(t)$ that leads to a zero output. Here, the zero is at $s = +\tau^{-1}$, corresponding to the unstable (in forward time) signal shown. In the text, we have discussed how zeros can arise due to cancellations that often depend on the precise placement of input and output.

How then, do we get a step if "nothing" has happened so far? Here we get to the subtlety of unstable internal states, discussed in Chapters 3 and 4. The input-output relation of a transfer function misses some internal quantities. Here, the internal variable blows up. Altering $u(t)$ suddenly from the value that "maintains" the zero output allows for a sudden output swing.

15.10 Anticipating the future improves control. Consider the system of Problem 15.7, for $|a| > 1$, with unstable uncontrolled dynamics. Scale $\nu^2 = 1$. What is the minimum power $P^* = \langle u^2(K^*)\rangle$ required to stabilize the system?



a. Use only current information. Assume $u_k = -Ky_k$, and show that choosing $K^* = a - 1/a$ leads to a minimum power $P^* = a^2 - 1$. (See Section 15.2.4.)

b. Assume that somehow you know y_{k+1} at time k and choose $u_k = -K_0 y_k - K_1 y_{k+1}$. Show that choosing $K_0^* = K_1^* = a - 1$ minimizes the power, with $P_+^* = 2(|a| - 1)$. For unstable systems, note that $P_+^* < P^*$ (see right).

c. Show that using y_{k-1} does not help. That is, if $u_k = -k_0 y_k - k_1 y_{k-1}$, the best feedback gains are $k_0 = a - 1/a$ and $k_1 = 0$, leading again to $P_-^* = a^2 - 1$.

Solution.

To recap, the equations of motion are

$$y_{k+1} = ay_k + u_k + v_k\,, \qquad \left\langle v_k^2 \right\rangle = \nu^2\,.$$

Hereafter, we set $\nu^2 = 1$. (In all cases, the power $\propto \nu^2$.)

a. This problem is mostly done in the text. We choose $u_k = -Ky_k$ and use the result from Problem 15.7 for the variance:

$$\left\langle y^2 \right\rangle = \frac{1}{1 - (a - K)^2}\,.$$

The power is

$$P = \left\langle u^2 \right\rangle = K^2 \left\langle y^2 \right\rangle = \frac{K^2}{1-(a-K)^2} .$$

The minimum is found by solving (preferably using a symbolic-algebra program)

$$\frac{\partial P}{\partial K} = \frac{2K\left(-a^2 + aK + 1\right)}{\left[1-(a-K)^2\right]^2} = 0$$

for K. The solutions are

$$K^* = \begin{cases} 0 & |a| < 1 \\ a - 1/a & |a| \geq 1 \end{cases} \qquad \Longrightarrow \qquad P^* = \begin{cases} 0 & |a| < 1 \\ a^2 - 1 & |a| \geq 1 \end{cases}$$

b. Now assume that we know y_{k+1} at time k, so that we can have $u_k = -K_0 y_k - K_1 y_{k+1}$. Then a similar calculation shows

$$\left\langle y^2 \right\rangle = \frac{1+K_1}{(1+K_1)^2 - (a-K_0)^2} , \qquad \langle y\, y_{-1} \rangle = \frac{a-K_0}{(1+K_1)^2 - (a-K_0)^2} .$$

The power in this case is denoted by P_+ and is

$$\begin{aligned} P_+ = \left\langle u^2 \right\rangle &= (K_0^2 + K_1^2)\left\langle y^2 \right\rangle + 2K_0K_1 \langle y\, y_{-1} \rangle \\ &= \frac{\left(K_0^2 + K_1^2\right)(1+K_1) + 2K_0K_1(a-K_0)}{(1+K_1)^2 - (a-K_0)^2} . \end{aligned}$$

We find the minimum by solving the simultaneous equations

$$\frac{\partial P}{\partial K_0} = \frac{\partial P}{\partial K_1} = 0$$

for K_0 and K_1. The solutions are $K_0^* = K_1^* = a-1$, with a corresponding power $P_+^* = 2(a-1)$.

c. What if we had no future information but tried to reduce the stabilization power by using past information? Let $u_k = k_0 y_k + k_1 y_{k-1}$. Then

$$\begin{aligned} \left\langle y^2 \right\rangle &= \frac{1+k_1}{(1-k_1)\left[(1+k_1)^2 - (a-k_0)^2\right]} , \\ \langle y\, y_{-1} \rangle &= \frac{a-k_0}{(1-k_1)\left[(1+k_1)^2 - (a-k_0)^2\right]} , \\ \langle y\, y_{-2} \rangle &= \frac{(a-k_0)^2 - k_1(1+k_1)}{(1-k_1)\left[(1+k_1)^2 - (a-k_0)^2\right]} . \end{aligned}$$

and the power is

$$P_- = \left\langle u^2 \right\rangle = (k_0^2 + k_1^2)\left\langle y^2 \right\rangle + 2k_0k_1 \langle y\, y_{-1} \rangle$$
$$= \frac{(k_0^2 + k_1^2)(1 + k_1) + 2k_0k_1(a - k_0)}{(1 - k_1)\left[(1 + k_1)^2 - (a - k_0)^2\right]} .$$

Minimizing this expression for $P(k_0, k_1)$ leads to $k_0^* = a - 1/a$ and $k_1^* = 0$, which is exactly the same expression that we found for a feedback $u_k = -Ky_k$. In other words, adding past information does not allow us to reduce the minimum power, and $P_-^* = P^*$.

15.11 Entropy-rate paradox. Equation (15.37) claims that, for a stable open-loop linear dynamics, the entropy rate of the output, $\mathcal{H}(Y)$, equals the entropy rate of an output disturbance, $\mathcal{H}(v)$. Yet Eq. (A.263) claims that if $y_k = a\, v_k$, then $\mathcal{H}(Y) = \mathcal{H}(v) + \ln|a|$. Reconcile these two statements mathematically and physically.

Solution.

The theorem applies to the sensitivity function $S = \frac{1}{1+L}$, where L is the open-loop transfer function, assumed stable. Physically, $L \to 0$ at high frequencies. Mathematically, this implies $\lim_{z\to\infty} S(z) = \frac{1}{1+0} = 1$. The initial-value theorem then implies $s_0 = \lim_{z\to\infty} S(z) = 1$, and Example A.26 shows that $\mathcal{H}(Y) = \mathcal{H}(v) + \ln|s_0| = \mathcal{H}(v)$.

Physically, an output disturbance immediately affects the output with unit gain (by definition). Then, because the closed-loop dynamics is stable, the information contained in the disturbance fades away. But because "all of it" entered in the initial step, it is present in the time series y_k.

By contrast, the situation with $y_k = av_k$ is different. It implies an immediate, constant gain a that applies for all frequencies. The "stretching" by the factor $|a|$ alters the amount of information gained in a measurement with fixed resolution (assuming a continuous alphabet for the values of Y and v). Notice that in Problem 15.12 below, we consider similar dynamics of the form $y_{k+1} = ay_k + v_k$ and show that for $|a| < 1$, the entropy rate truly converges to $\mathcal{H}(Y) = \mathcal{H}(v)$. (The a has a different interpretation in that problem.) In Problem 15.12, the coefficient of v is again 1. If not, the entropy rate would also have been altered.

Physically, then, the equality of entropy rates comes from the fact that output disturbances affect the output instantaneously and with unit gain.

15.12 Entropy rate of the output of a stable 1d system. Let $x_{k+1} = ax_k + v_k$, with $v_k \sim \mathcal{N}(0, v^2)$ and $|a| < 1$. Let the output $y_k = x_k$ (no measurement noise).

a. Show that the variance of the output is $\langle y_k^2 \rangle = \frac{v^2}{1-a^2}$.
b. By direct calculation in the time domain, show that the entropy rate $\mathcal{H}(Y) = \mathcal{H}(v)$, where the time series Y has realizations y_k and the series v has realizations v_k.

Solution.

a. Assuming stationarity, which requires stable motion ($|a| < 1$), and substituting $y_k = x_k$, we have

$$\langle y\, y_{-1}\rangle = a\left\langle y^2\right\rangle + 0$$
$$\left\langle y^2\right\rangle = a\,\langle y\, y_{-1}\rangle + \nu^2\,,$$

which implies that $\langle y^2\rangle = \frac{1}{1-a^2}\nu^2$ and $\langle y\, y_{-1}\rangle = \frac{a}{1-a^2}\nu^2$.

b. The entropy rate $\mathcal{H}(\nu)$ of the independent random variables ν_k is that of a single variable, $H(\nu) = \ln\sqrt{2\pi e} + \frac{1}{2}\ln\nu^2$. For $\mathcal{H}(Y)$, we calculate the entropy of $H(Y^N)$, which, from Eq. (A.255) and Problem A.10.2, is given by

$$H(Y^N) = \ln\left(\sqrt{2\pi e}\right)^N + \frac{1}{2}\ln|\det\mathbf{\Sigma}|\,,$$

where $\mathbf{\Sigma}$ is the covariance matrix with elements $\langle y_i\, y_j\rangle$. Continuing the argument from Part (a) gives, for the delayed correlations,

$$\langle y_i\, y_j\rangle = \left(\frac{a^{|i-j|}}{1-a^2}\right)\nu^2\,,$$

which corresponds, explicitly, to a covariance matrix $\mathbf{\Sigma}$ that is a symmetric, banded-diagonal (Toeplitz) $N \times N$ matrix:

$$\mathbf{\Sigma} = \frac{\nu^2}{1-a^2}\begin{pmatrix} 1 & a & a^2 & \dots & a^N \\ a & 1 & a & & a^{N-1} \\ a^2 & a & 1 & \ddots & \\ \vdots & & \ddots & \ddots & a \\ a^N & & a^2 & a & 1 \end{pmatrix}$$

By evaluating explicitly low-order cases or using a symbolic algebra program, it is easy to see that

$$\det\mathbf{\Sigma} = \left(\frac{\nu^2}{1-a^2}\right)^N (1-a^2)^{N-1} = \frac{\nu^{2N}}{1-a^2}\,.$$

Then

$$\begin{aligned}\mathcal{H}(Y) &= \lim_{N\to\infty}\frac{H(Y^N)}{N} \\ &= \lim_{N\to\infty}\frac{1}{N}\left[\ln(2\pi e)^{N/2} + \frac{1}{2}\ln\nu^{2N} - \cancelto{0}{\frac{1}{2}\log\left(1-a^2\right)}\right] \\ &= \ln\sqrt{2\pi e} + \frac{1}{2}\ln\nu^2 \\ &= \mathcal{H}(\nu)\,.\end{aligned}$$

15.13 Causal conditioning. Prove that Eqs. (15.38) and (15.39) for ordinary and causal conditioning are equivalent. Hints: Use Bayes repeatedly and do $N = 2$ explicitly.

Solution.

Let us first do the $N = 2$ case explicitly, to get some hints as to how to proceed more generally. For the ordinary decomposition of conditional probability,

$$\begin{aligned} P(X^N, Y^N) &= P(Y^N|X^N)\, P(X^N) \\ &= \prod_{k=1}^{N} P(Y_k|Y^{k-1}, X^N)\, P(X_k|X^{k-1}) \\ \text{for } n = 2 \rightarrow\ & P(Y_2|Y_1, X^2)\, P(Y_1|X^2)\, P(X_2|X_1)\, P(X_1) \end{aligned}$$

For the causal decomposition of conditional probability,

$$\begin{aligned} P(X^N, Y^N) &= P(Y^N\|X^N)\, P(X^N\|Y^{N-1}) \\ &= \prod_{k=1}^{N} P(Y_k|Y^{k-1}, X^k)\, P(X_k|X^{k-1}, Y^{k-1}) \\ \text{for } n = 2 \rightarrow\ & P(Y_2|Y_1, X^2)\, P(Y_1|X_1)\, P(X_2|X_1, Y_1)\, P(X_1)\,. \end{aligned}$$

Canceling common terms, equating the remaining one, and using Bayes' theorem gives

$$\begin{aligned} P(Y_1|X^2)\, P(X_2|X_1) &\stackrel{?}{=} P(Y_1|X_1)\, P(X_2|X_1, Y_1) \\ &= \frac{\cancel{P(Y_1|X_1)}\, P(Y_1|X_1, X_2)\, P(X_2|X_1)}{\cancel{P(Y_1|X_1)}}\,. \end{aligned}$$

Thus, the two decompositions of conditional probability are equivalent for this simple case. More generally, for

$$P(X^N, Y^N) = P(Y^N\|X^N)\, P(X^N\|Y^{N-1}) = \prod_{k=1}^{N} P(Y_k|Y^{k-1}, X^k)\, P(X_k|X^{k-1}, Y^{k-1})\,,$$

we apply Bayes' theorem to $P(X_k|X^{k-1}, Y^{k-1}) = P(X_k|Y^{k-1}, X^{k-1})$:

$$P(X_k|Y^{k-1}, X^{k-1}) = \frac{P(Y_{k-1}|Y^{k-2}, X^k)\, P(X_k|Y^{k-2}, X^{k-1})}{P(Y_{k-1}|Y^{k-2}, X^{k-1})}\,.$$

The denominator cancels all the terms in the first product except $k = N$, leaving

$$P(X^N, Y^N) = P(Y_N|Y^{N-1}, X^N) \prod_{k=1}^{N} P(Y_{k-1}|Y^{k-2}, X^k)\, P(X_k|Y^{k-2}, X^{k-1})\,.$$

We use Bayes' theorem again on the second product:

$$P(X_k|Y^{k-2}, X^{k-1}) = \frac{P(Y_{k-2}|Y^{k-3}, X^k)\, P(X_k|Y^{k-3}, X^{k-1})}{P(Y_{k-2}|Y^{k-3}, X^{k-1})}\,.$$

Again, the denominator cancels all but the $k = N$ term of the first product, giving

$$P(X^N, Y^N) = P(Y_N|Y^{N-1}, X^N)\, P(Y_{N-1}|Y^{N-2}, X^N) \prod_{k=1}^{N} P(Y_{k-2}|Y^{k-3}, X^k)\, P(X_k|Y^{k-3}, X^{k-1})\,.$$

Repeating, we see that we will eventually generate from the first term

$$P(Y_N|Y^{N-1}, X^N)\, P(Y_{N-1}|Y^{N-2}, X^N)\, \ldots = \prod_{k=1}^{N} P(Y_k|Y^{k-1}, X^N) = P(Y^N|X^N)\,.$$

The remaining terms will be of the form

$$\prod_{k=1}^{N} P(X_k|Y^{k-N}, X^{k-1}) = \prod_{k=1}^{N} P(X_k|X^{k-1}) = P(X^N)\,,$$

where we note that $Y_0, Y_{-1}, \ldots$ are all equal to the empty set (there is only $Y_1, \ldots, Y^N$) and can thus be eliminated from the conditioning. Thus we have proven that

$$P(X^N, Y^N) = P(Y^N||X^N)\, P(X^N||Y^{N-1}) = P(Y^N|X^N)\, P(X^N)\,.$$

Note that I found it helpful to work out the $N = 3$ case explicitly, as well.

15.14 Directed information decomposition. Prove Eq. (15.43). Hint: Use Eq. (A.287).

Solution.

The version of the mutual information definition in Eq. (A.287) makes the symmetric of its arguments explicit:

$$I(X^N; Y^N) = H(X^N) + H(Y^N) - H(X^N, Y^N)\,.$$

The joint entropy is

$$H(X^N, Y^N) = -\sum P(X^N, Y^N) \log P(X^N, Y^N)\,.$$

But, from Eq. (15.39), we have

$$P(X^N, Y^N) = P(Y^N||X^N)\, P(X^N||Y^{N-1})\,,$$

Then

$$\begin{aligned} H(X^N, Y^N) &= -\sum P(X^N, Y^N) \log P(Y^N||X^N)\, P(X^N||Y^{N-1}) \\ &= -\sum P(X^N, Y^N) \log P(Y^N||X^N) - \sum P(X^N, Y^N) \log P(X^N||Y^{N-1}) \\ &= H(Y^N||X^N) + H(X^N||Y^{N-1})\,. \end{aligned}$$

Going back to the definition of mutual information, we have

$$\begin{aligned} I(X^N; Y^N) &= H(X^N) + H(Y^N) - H(X^N, Y^N) \\ &= H(X^N) + H(Y^N) - H(Y^N||X^N) - H(X^N||Y^{N-1}) \\ &= \left[H(Y^N) - H(Y^N||X^N)\right] + \left[H(X^N) - H(X^N||Y^{N-1})\right] \\ &= I(X^N \to Y^N) + I(Y^{N-1} \to X^N)\,, \end{aligned}$$

which is the identity we set out to demonstrate.

15.15 Mutual vs. directed information. Using the chain rule, show the results for mutual and directed information claimed in Example 15.5.

Solution.

For this problem, let $X^N = \{X_0, X_1, \ldots, X_N\}$ and $Y^N = \{Y_1, \ldots, Y^N\}$. That is, X starts at $k = 0$ and Y starts at $k = 1$. Then the dynamics $Y_k = X_{k-1}$ is valid for $k = 1$ to N.

- Mutual-information rate, $\mathcal{I}(X \to Y)$.

$$\begin{aligned}
I(X^N; Y^N) &= \sum_{k=1}^{N} I(Y_k; X^N | Y^{k-1}) && \text{chain rule} \\
&= \sum_{k=1}^{N} H(Y_k|Y^{k-1}) - H(Y_k|Y^{k-1}, X^N) \\
&= \sum_{k=1}^{N} H(X_{k-1}|X^{k-2}) - H(X_{k-1}|X^{k-2}, X^N) && (Y_k = X_{k-1}) \\
&= \sum_{k=1}^{N} H(X_{k-1}) - H(X_{k-1}|X_{k-1}) && (X_k \text{ are i.i.d.}) \\
&= \sum_{k=1}^{N} H(X) - 0 \\
&= N\, H(X)\,.
\end{aligned}$$

Thus, $\mathcal{I}(X; Y) = H(X)$. Intuitively, since X and Y are the same time series shifted by 1 unit, each measurement of Y reduces the uncertainty about X by $H(X)$. Note, however, that our intuition depends on the i.i.d. assumption about the stochastic process for X. When there are correlations, information is "spread out" and not "localized" to a particular variable.

- Directed-information rate, $\mathcal{I}(X \to Y)$.

$$\begin{aligned}
I(X^N \to Y^N) &= \sum_{k=1}^{N} I(Y_k; X^k | Y^{k-1}) && \text{definition of dir. info.} \\
&= \sum_{k=1}^{N} H(Y_k|Y^{k-1}) - H(Y_k|Y^{k-1}, X^k) \\
&= \sum_{k=1}^{N} H(X_{k-1}|X^{k-2}) - H(X_{k-1}|X^{k-2}, X^k) \\
&= \sum_{k=1}^{N} H(X_{k-1}) - H(X_{k-1}|X_{k-1}) \\
&= \sum_{k=1}^{N} H(X) - 0 \\
&= N\, H(X)\,.
\end{aligned}$$

Thus, $\mathcal{I}(X \to Y) = H(X)$.

- Directed-information rate, $\mathcal{I}(Y \to X)$.

$$\begin{aligned} I(Y^N \to X^N) &= \sum_{k=1}^{N} I(X_k; Y^k | X^{k-1}) \\ &= \sum_{k=1}^{N} H(X_k|X^{k-1}) - H(X_k|X^{k-1}, Y^k) \\ &= \sum_{k=1}^{N} H(X_k|X^{k-1}) - H(X_k|X^{k-1}, X^{k-1}) \\ &= \sum_{k=1}^{N} H(X_k) - H(X_k) \\ &= 0\,. \end{aligned}$$

Thus, $\mathcal{I}(Y \to X) = 0$.

The last two results show that X causes Y, and not the reverse. Note that our results are consistent with Problem 15.14, which proves that

$$I(X^N; Y^N) = I(X^N \to Y^N) + I(Y^{N-1} \to X^N)\,.$$

(Remember that we extended the X variables by adding X_0 here.)

15.16 Information rates for a finite-bandwidth, continuous system. Consider an amplifier that acts also as a low-pass filter, with transfer function $G(s) = G_0/(1+s)$, that is used as a transducer between a continuous input signal $u(t)$ and a continuous output signal $y(t)$. Let $y(s) = G(s)\,u(s) + \xi(s)$, with $\langle \xi(t)\,\xi(t')\rangle = \xi^2 \delta\,(t - t')$.

a. Using integration by parts, show that $\int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} \ln\left[\prod_i \left(\frac{\omega^2+a_i^2}{\omega^2+b_i^2}\right)\right] = \sum_i (a_i - b_i)$.

b. Show that $\mathcal{I}(U; Y)$ is given by Eq. (15.48).

Solution.

a. We have

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} \ln\left[\prod_i \left(\frac{\omega^2 + a_i^2}{\omega^2 + b_i^2}\right)\right] &= \int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} \ln\left[\prod_i \left(\frac{1 + \frac{a_i^2}{\omega^2}}{1 + \frac{b_i^2}{\omega^2}}\right)\right] \\ &= \sum_i \int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} \left[\ln\left(1 + \frac{a_i^2}{\omega^2}\right) - \left(1 + \frac{b_i^2}{\omega^2}\right)\right]. \end{aligned}$$

We thus focus on

$$\int \frac{\mathrm{d}\omega}{2\pi} \ln\left(1 + \frac{a^2}{\omega^2}\right) = \frac{a}{2\pi} \int \mathrm{d}\omega \ln\left(1 + \frac{1}{\omega^2}\right).$$

Isolating the dimensionless part, we integrate by parts:

$$\begin{aligned}\int \mathrm{d}\omega \ln\left(1+\frac{1}{\omega^2}\right) &= \omega \ln\left(1+\frac{1}{\omega^2}\right) - \int \mathrm{d}\omega\, \omega \frac{(-2/\omega^3)}{1+1/\omega^2} \\ &= \omega \ln\left(1+\frac{1}{\omega^2}\right) + 2\int \mathrm{d}\omega \frac{1}{1+\omega^2} \\ &= \omega \ln\left(1+\frac{1}{\omega^2}\right) + 2\tan^{-1}\omega\,.\end{aligned}$$

With limits $\omega = \pm\infty$, we have $\omega \ln\left(1+1/\omega^2\right) \to \omega/\omega^2 = 1/\omega \to 0$ and also $\tan\omega \to \left[\frac{\pi}{2} - (-\frac{\pi}{2})\right] = \pi$. Thus,

$$\int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} \ln\left(1+\frac{a^2}{\omega^2}\right) = \frac{a}{2\pi}(0+2\pi) = a\,,$$

and the full integral follows immediately.

b. With $G = \frac{G_0}{1+s/\omega_c}$ and $|G|^2(\omega) = \frac{G_0^2}{1+\omega^2/\omega_c^2}$, we have

$$\begin{aligned}\mathcal{I}(U;Y) &= \frac{1}{2}\int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} \ln\left(1+\frac{\mathrm{SNR}_0^2}{1+\omega^2/\omega_c^2}\right) \\ &= \frac{1}{2}\int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} \ln\left(\frac{\omega^2+\omega_c^2(1+\mathrm{SNR}_0^2)}{\omega^2+\omega_c^2}\right) \\ &= \frac{\omega_c}{2}\left(\sqrt{1+\mathrm{SNR}_0^2}-1\right).\end{aligned}$$

where $\mathrm{SNR}_0 = G_0\sigma_u/\xi$.

15.17 Information rates in two different limits. In Section 15.2.3, the mutual information rate for a constant-gain amplifier, sampled at T_s, is given as $\mathcal{I}(U;Y) = \frac{1}{2T_s}\log(1+\mathrm{SNR}_0^2)$. On the other hand, the rate for a continuously sampled amplifier of bandwidth ω_c is $\mathcal{I}(U;Y) = \frac{\omega_c}{2}(\sqrt{1+\mathrm{SNR}_0^2}-1)$. Calculate $\mathcal{I}(U;Y)$ for finite T_s and ω_c, and reconcile the two expressions. Assume $G(s) = \frac{G_0}{1+s/\omega_c}$.

Solution.

For a finite sampling interval T_s, the maximum frequency is the Nyquist frequency, $|\omega_N| = \pi/T_s$. With a finite bandwidth ω_c, the amplifier response is

$$|G|^2(\omega) = \frac{G_0^2}{1+\frac{\omega^2}{\omega_c^2}}\,.$$

With $\mathrm{SNR}_0 = \frac{G_0\sigma_u}{\xi}$ and $\alpha \equiv \omega_N/\omega_c$, the information rate is

$$\begin{aligned}\mathcal{I}(U;Y) &= \frac{1}{2}\int_{-\omega_N}^{\omega_N} \frac{\mathrm{d}\omega}{2\pi} \log\left(1+\frac{\mathrm{SNR}_0^2}{1+\frac{\omega^2}{\omega_c^2}}\right) \\ &= \frac{\omega_c}{2}\int_{-\alpha}^{\alpha} \frac{\mathrm{d}\omega}{2\pi} \log\left(1+\frac{\mathrm{SNR}_0^2}{1+\omega^2}\right)\end{aligned}$$

$$= \frac{\omega_c}{2} \int_{-\alpha}^{\alpha} \frac{d\omega}{2\pi} \log\left(\frac{\omega^2 + 1 + SNR_0^2}{\omega^2 + 1}\right)$$

$$= \frac{\omega_c}{2\pi}\left[\alpha \log\left(1 + \frac{SNR_0^2}{1+\alpha^2}\right) + 2\sqrt{1+SNR_0^2}\,\tan^{-1}\left(\frac{\alpha}{\sqrt{1+SNR_0^2}}\right) - 2\tan^{-1}\alpha\right],$$

where the integral is derived in Problem 15.16.

In the limit $\omega_N \gg \omega_c$, or $\alpha \to \infty$, we use the result from Problem 15.16:

$$\mathcal{I}(U;Y) = \frac{\omega_c}{2}\left[\sqrt{1+SNR_0^2} - 1\right].$$

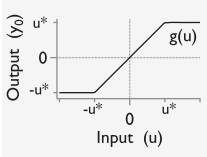
Alternatively, in the limit $\omega_c \gg \omega_N$, or $\alpha \to 0$, we find

$$\mathcal{I}(U;Y) = \frac{\omega_c}{2\pi}\left[\alpha \log\left(1+SNR_0^2\right) + 2\sqrt{1+SNR_0^2}\left(\frac{\alpha}{\sqrt{1+SNR_0^2}}\right) - 2\alpha\right]$$

$$= \frac{\omega_N}{2\pi}\log\left(1+SNR_0^2\right)$$

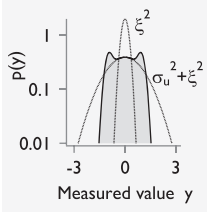
$$= \frac{1}{2T_s}\log\left(1+SNR_0^2\right).$$

Thus, we see that the two expressions depend on which infinite limit is taken first. Should we be concerned about this ambiguity? I would argue no: concepts such as infinite bandwidths are idealizations of Nature that are never precisely realized, anymore than other concepts such as perfect linear relationships or perfect geometric forms. In a real physical situation, ω_c is finite. And, while a relationship can be continuous (assuming time itself is not quantized), any measurements will have an effective ω_N, too. In a given physical situation, there will be a specific value of α. If $\alpha \gg 1$ or $\alpha \ll 1$, then we can make the appropriate approximation.



15.18 Nonlinearities can reduce information rates. Consider measuring the signal $u_k \sim \mathcal{N}(0, \sigma_u^2)$ with the nonlinear saturation function $y_0 = g(u)$ shown at left.

a. Show, for suitably defined a, that $p(y_0) = a\left[\delta(y_0 - u^*) + \delta(y_0 + u^*)\right] + \frac{1}{\sqrt{2\pi}} e^{-y_0^2/2}$ for $|y_0| \le u^*$ and 0 otherwise. (Check that $p(y_0)$ is normalized, too.)



b. Then find the full distribution for $p(y)$ for $y = g(u) + \xi$ by convoluting with the noise distribution $\xi \sim \mathcal{N}(0, \xi^2)$. Do the convolution symbolically or numerically.

c. For $u^* = \sigma_u = 1$ and $\xi = 0.2$, confirm $p(y)$, left. Confirm, too, that the dashed lines show the limiting distributions: $\mathcal{N}(0, \xi^2)$ for $u^*/\sigma_u \ll 1$ and $\mathcal{N}(0, \sigma_u^2 + \xi^2)$ for $u^*/\sigma_u \gg 1$. Confirm, too, the information-rate plot in the text.

Solution.

a. To show that

$$p(y_0) = \begin{cases} a\left[\delta\left(y_0 - u^*\right) + \delta\left(y_0 + u^*\right)\right] + \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-y_0^2/2} & |y_0| \le u^* \\ 0 & |y_0| > u^* \end{cases}$$

we change variables:

$$p(y_0) = \left.\frac{p(u)}{|g'(u)|}\right|_{u=g^{-1}(y_0)}, \qquad p(u) = \frac{1}{\sqrt{2\pi\sigma_u^2}}\, \mathrm{e}^{-u^2/2\sigma_u^2}\,.$$

The δ-functions then arise from the regions $|u| > u^*$, with weight

$$a = \int_{u^*}^{\infty} \mathrm{d}u\, \frac{\mathrm{e}^{-u^2/2\sigma_u^2}}{\sqrt{2\pi\sigma_u^2}} = \frac{1}{2}\,\mathrm{erfc}\left(\frac{u^*}{\sqrt{2}\sigma_u}\right).$$

For $-u^* < u < u^*$, $g'(u) = 1$, so that $p(y_0) = p(u)$ in that region. This gives the last term in the PDF.

We verify that

$$\int_{-u^*}^{u^*} \mathrm{d}u\, \frac{\mathrm{e}^{-u^2/2\sigma_u^2}}{\sqrt{2\pi\sigma_u^2}} = 1 - 2a\,,$$

showing that the full distribution is normalized.

b. The noisy measurement is $y = y_0 + \xi$. Thus, we find $p(y)$ by convoluting $p(y_0)$, found above with the noise distribution $p(\xi)$, with $\xi \sim \mathcal{N}(0, \xi^2)$: Thus, $p(y) = p(y_0) * p(\xi)$. This is straightforward numerically but can also be done symbolically. Let us see how to do it here (a symbolic-algebra program helps). The δ functions each give rise to a Gaussian distribution with mean $\pm u^*$. The convolution integral is

$$\int_{-\infty}^{\infty} \mathrm{d}x\, \frac{1}{\sqrt{2\pi\xi^2}}\, \mathrm{e}^{-\frac{x^2}{2\xi^2}} \begin{cases} \frac{1}{\sqrt{2\pi\sigma_u^2}}\, \mathrm{e}^{-\frac{(y-x)^2}{2\sigma_u^2}} & |y| \le u^* \\ 0 & |y| > u^* \end{cases}$$

Evaluating the integral and combining with the δ-function convolutions gives

$$p(y|u^*, \xi^2) = \frac{1}{2}\frac{a}{\sqrt{2\pi\xi^2}}\left(\mathrm{e}^{-\frac{(y-u^*)^2}{2\xi^2}} + \mathrm{e}^{-\frac{(y+u^*)^2}{2\xi^2}}\right) + \frac{1}{2}\frac{\mathrm{e}^{-\frac{y^2}{2(\sigma_u^2+\xi^2)}}\left[\mathrm{erf}\left(\frac{u^*\xi^2 + u^*\sigma_u^2 - y\sigma_u^2}{\sigma_u\xi\sqrt{2(\sigma_u^2+\xi^2)}}\right) + \mathrm{erf}\left(\frac{u^*\xi^2 + u^*\sigma_u^2 + y\sigma_u^2}{\sigma_u\xi\sqrt{2(\sigma_u^2+\xi^2)}}\right)\right]}{\sqrt{2\pi(\sigma_u^2 + \xi^2)}}\,.$$

For $u^* \gg \{\sigma_u, \xi\}$, we have $a \to 0$ and $\mathrm{erf}(\cdot) \to 1$, $\implies$ $p(y|u^*, \xi^2) \to \mathcal{N}(0, \sigma_u^2 + \xi^2)$.

For $u^* \ll \{\sigma_u, \xi\}$, we have $a \to \frac{1}{2}$ and $\mathrm{erf}(\cdot) \to 0$, $\implies$ $p(y|u^*, \xi^2) \to \mathcal{N}(0, \xi^2)$.

c. See code on book website. The mutual-information calculations are based on Eq. (15.50), which states that

$$I(U;Y) = H(Y) - H(\xi)\,.$$

The latter entropy is $\log\sqrt{2\pi e\,\xi^2}$, and the calculation reduces to finding $H(Y)$, which is based on $p(y) = p(y_0) * p(\xi)$.

15.19 Information flow in the small-noise limit. Make the arguments about the small-noise limit that lead to Eq. (15.53) more precise.

Solution.

The main issue is how to convert $p(y)$ to $p(u)$ in the low-noise limit. We proceed by formulating the joint distribution for $p(y,u,\xi)$ and then marginalizing over u and ξ to find $p(y)$. Using the definition of conditional probability and the independence of u and ξ, we have,

$$\begin{aligned} p(y) &= \int \mathrm{d}u\,\mathrm{d}\xi\; p(y,u,\xi) \\ &= \int \mathrm{d}u\,\mathrm{d}\xi\; p(y|u,\xi)\,p(u)\,p(\xi) \\ &= \int \mathrm{d}u\,\mathrm{d}\xi\; \delta[y-g(u)-\xi]\,p(u)\,p(\xi) \\ &= \int \mathrm{d}u\; p(u)\,p[y-g(u)]\,. \end{aligned}$$

Here, $p(\xi) = \mathcal{N}(0,\xi^2)$, so that $p[y-g(u)] = \mathcal{N}[y-g(u),\xi^2]$ is just a normal distribution, with mean $y-g(u)$ and variance ξ^2. Thus,

$$\begin{aligned} p(y) &= \frac{1}{\sqrt{2\pi\xi^2}}\int \mathrm{d}u\; p(u)\;\mathrm{e}^{-\frac{[y-g(u)]^2}{2\xi^2}} = \frac{1}{\sqrt{2\pi\xi^2}}\int \frac{\mathrm{d}z}{|g'(u)|}\;p(u)\;\mathrm{e}^{-\frac{[y-z]^2}{2\xi^2}} \\ &\to \int \frac{\mathrm{d}z}{|g'(u)|}\;p(u)\,\delta\,(y-z)\,, \end{aligned}$$

where we have changed variables from u to $z = g(u)$ and then taken the low-noise limit $\xi \to 0$. Evaluating the integral then gives

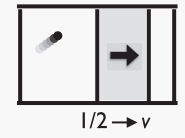
$$p(y) = \left.\frac{p(u)}{|g'(u)|}\right|_{u=g^{-1}(y)}\,.$$

Here $g(u)$ is assumed to be monotonic in u. This is equivalent to Eq. (15.53).

15.20 Classic Szilard engine with noisy measurements. For the "energy" version of the Szilard engine discussed in the text, one can extract energy up to $k_{\mathrm{B}}T\ln 2\,I(X;Y)$ for noisy measurements where the probability that the wrong state is observed is ξ. Here, we show that the same result occurs for the traditional version of the Szilard engine illustrated in Section 15.3.1. See Sagawa (2019).

a. Show that the average information gained by a single measurement y is $I(X;Y) = 1 - H_2(\xi)$, where $H_2(\xi)$ is the Shannon entropy function for two states, in bits.

b. Extract an average work $k_B T \ln 2\, I(X;Y)$ as follows: If the measurement shows the particle on the left, move the partition from the center of the box to a position v chosen to maximize the extracted work. What if the measurement indicates that the particle is on the right? Explain intuitively this optimal protocol.



Solution.

a. From the definition of mutual information between the state x and measurement y,

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)\,.$$

The two forms are equivalent because $I(X;Y) = I(Y;X)$. Here, the second form is more convenient because we are given the conditional probabilities $P(Y|X)$. Note that it is obvious, due to the symmetry of the problem, that $P(Y) = P(X) = \frac{1}{2}$, so that $H(Y) = \ln 2$. For the conditional entropy,

$$\begin{aligned} H(Y|X) &= -\sum_{\{X,Y\}=L,R} P(X,Y) \ln P(Y|X) \\ &= -P(Y) \sum_{\{X,Y\}=L,R} P(Y|X) \ln P(Y|X) \\ &= 2\left(\tfrac{1}{2}\right)[-\xi \ln \xi - (1-\xi)\ln(1-\xi)] \\ &= H_2(\xi)\,. \end{aligned}$$

Thus,

$$I(X;Y) = H(Y) - H(Y|X) = \ln 2 - H_2(\xi)\,.$$

b. To calculate the maximum extractable work, we analyze the suggested protocol. We also first calculate the average work conditioned on the observation of a particle in the left side.

$$\langle W_L \rangle = P(Y = L|X = L)P(X = L)W(v, \text{good}) + P(Y = L|X = R)P(X = R)W(v, \text{bad})\,,$$

where $W(v, \text{good})$ is the work extracted when the measurement is correct and $W(v, \text{bad})$ is the work "paid" when the measurement is in error. In units of $k_B T$, we have

$$\begin{aligned} W(v, \text{good}) &= \ln \frac{v}{1/2} = \ln 2 + \ln v \\ W(v, \text{bad}) &= \ln \frac{1-v}{1/2} = \ln 2 + \ln(1-v)\,. \end{aligned}$$

In the "bad" case, we are compressing the particle from a volume 1/2 that of the box to $(1-v)$. Substituting and using $P(X = L, R) = \frac{1}{2}$ and the conditional probabilities for observations gives

$$\langle W_L \rangle = (1-\xi)\tfrac{1}{2}[\ln 2 + \ln v] + \xi \tfrac{1}{2}[\ln 2 + \ln(1-v)]$$

$$= \frac{1}{2}\left[(1-\xi+\xi)\ln 2 + (1-\xi)\ln v + \xi\ln(1-v)\right]$$
$$= \frac{1}{2}\left[\ln 2 + (1-\xi)\ln v + \xi\ln(1-v)\right] .$$

By symmetry, it is clear that the other contribution $\langle W_R\rangle$ will have exactly the same form, with $\xi \to 1-\xi$ and $v \to 1-v$. Thus,

$$\langle W_{\text{ext}}\rangle = \frac{2}{2}\left[\ln 2 + (1-\xi)\ln v + \xi\ln(1-v)\right] .$$

The last step is to choose v to maximize $\langle W_{\text{ext}}\rangle$. Solving $d_v\langle W_{\text{ext}}\rangle(v) = 0$ gives $v = 1-\xi$ and

$$\langle W_{\text{ext}}\rangle = \left[\ln 2 + (1-\xi)\ln(1-\xi) + \xi\ln\xi\right]$$
$$= \ln 2 + H_2(\xi) = I(X;Y) .$$

Thus, $\langle W_{\text{ext}}\rangle = k_{\text{B}}TI(X;Y)$. Measurement error reduces the information acquired from the system, which then reduces by the same amount (times $k_{\text{B}}T$) the work that can be extracted. Of course, poorer protocols can extract less work.

Why move the partition only partway? When the measurement is in error, we will compress the gas. If we compress too much, the cost of bad measurements will be too high. The optimal protocol balances potential gains against losses.

15.21 Two forms of the master equation. The master equation $\mathrm{d}_t p_i = \sum_j \left(W_{ij}\, p_j - W_{ji} p_i\right)$ has transition rates W_{ij} from state j to i.

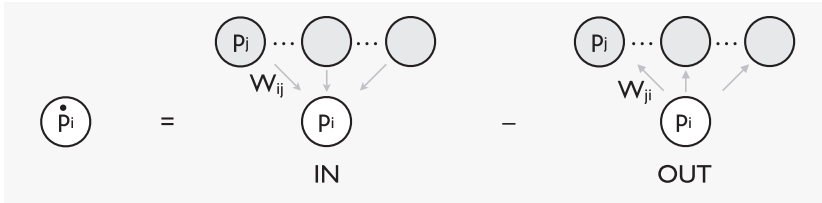
a. Interpret this form of the master equation physically in terms of in and out currents. Can you give a graphical interpretation, too? (Cf. Section 15.3.4.)
b. An equivalent form of the master equation that is more convenient mathematically is given by $\mathrm{d}_t \boldsymbol{p} = \mathbb{W}\boldsymbol{p}$. For this form, why must each column of $\mathbb{W}$ sum to zero?
c. Express the rate matrix $\mathbb{W}$ in terms of the transition rates W_{ij}.
d. Relate $\mathrm{d}_t \boldsymbol{p} = \mathbb{W}\boldsymbol{p}$ to its discrete-time counterpart $\boldsymbol{p}_{k+1} = \boldsymbol{A}\boldsymbol{p}_k$ for a protocol of duration τ. (Hint: exponentiate to find a condition between $\mathbb{W}$ and $\boldsymbol{A}$.)
e. Write Eq. (15.67) for a two-state system. What is its matrix $\mathbb{W}$?

Solution.

a. The product of probability × transition rate can be interpreted as a vector of probability currents $\boldsymbol{J} = \boldsymbol{W}\boldsymbol{p}$. That is, each component J_i is the current through the i^{th} state, so that the master equation can also be interpreted as

$$\mathrm{d}_t \boldsymbol{p} = \boldsymbol{J} = \boldsymbol{J}_{\text{in}} - \boldsymbol{J}_{\text{out}} .$$

A graphical interpretation of the first form of the master equation, identifying the in- and outgoing currents into each node p_i is given below:



b. The condition that the columns of $\mathbb{W}$ sum to zero is required by the need to keep the probabilities $p_i(t)$ normalized at all times. If we sum the master equation $\mathrm{d}_t p_i = \sum_j \mathbb{W}_{ij} p_j$ over i, we have

$$\sum_i \mathrm{d}_t p_i = \sum_i \left(\sum_j \mathbb{W}_{ij} p_j \right)$$

$$\mathrm{d}_t \left(\sum_i p_i \right) = \sum_j p_j \left(\sum_i \mathbb{W}_{ij} \right)$$

$$\mathrm{d}_t(1) = \sum_j p_j \left(\sum_i \mathbb{W}_{ij} \right)$$

$$0 = \sum_j p_j \left(\sum_i \mathbb{W}_{ij} \right).$$

The inside sum must be true for arbitrary p_j vectors (whose components sum to one), which implies that the columns of the matrix $\boldsymbol{W}$ must sum to zero. This is the continuous version of the normalization condition for discrete-time dynamics, $\sum_i \boldsymbol{A}_{ij} = 1$.

c. Let us try the following rule for constructing $\mathbb{W}$ from $\boldsymbol{W}$:

$$\mathbb{W}_{ij} = \begin{cases} W_{ij} & i \neq j \\ -\sum_k W_{kj} & i = j \end{cases}$$

We could also express this as

$$\mathbb{W}_{ij} = W_{ij} - \delta_{ij} \sum_k W_{kj}$$

using the Kronecker delta function (1 if $i = j$ and 0 otherwise). It is worth noting that $W_{ii} = 0$ for all i, as $\boldsymbol{W}$ records only the transitions between different states.

The above definition is set up so that the normalization condition is automatically enforced:

$$\sum_i \mathbb{W}_{ij} = \left(\sum_{i \neq j} W_{ij} \right) + W_{jj} = \left(\sum_{i \neq j} W_{ij} \right) - \sum_k W_{kj} = 0\,.$$

Armed with this form for $\mathbb{W}$, we show that the compact matrix form is equivalent to the original in-out form. The evolution equation for probabilities is

$$\begin{aligned} \mathrm{d}_t p_i &= \sum_{j=1}^{n} \mathbb{W}_{ij} p_j \\ &= \sum_{j\neq i} \mathbb{W}_{ij} p_j + \mathbb{W}_{ii} p_i \\ &= \sum_{j\neq i} \left(W_{ij}\, p_j - W_{ji}\, p_i \right) \\ &= \sum_{j} \left(W_{ij}\, p_j - W_{ji}\, p_i \right) . \end{aligned}$$

The first term represents the fraction of states going from any other state i into j, while the second term represents the fraction of states going out from j to any other state i. The master equation is usually written in the more intuitive in-out form. In the last line, we can remove the restriction that the sum be over $j \neq i$, changing to a simple sum over j. This step is justified because the $i = j$ term has a net right-hand side that is automatically zero.

d. Exponentiating the continuous-time equation

$$\mathrm{d}_t \boldsymbol{p} = \boldsymbol{W}\boldsymbol{p}$$

gives

$$\boldsymbol{p}(\tau) = \mathrm{e}^{\boldsymbol{W}\tau}\, p(0) \equiv \boldsymbol{A}\boldsymbol{p}(0) .$$

Thus,

$$\boldsymbol{A} = \mathrm{e}^{\tau \boldsymbol{W}} \approx \mathbb{I} + \tau \boldsymbol{W} ,$$

where the latter approximation is to first order in τ. These relations for stochastic transition matrices are analogous to the ones we derived in Chapter 5 to relate discrete- and continuous-time formulations of linear dynamics.

e. To connect to the Maxwell-demon example of Section 15.3.5, we denote the *left* and *right* states as 1 and 2. Thus, we define p_1 and $p_2 = 1 - p_1$. The equations of motion, in this language, are

$$\begin{aligned} \mathrm{d}_t p_1 &= -\omega_{21} p_1 + \omega_{12} p_2 \\ \mathrm{d}_t p_2 = -\dot{p}_1 &= +\omega_{21} p_1 - \omega_{12} p_2 , \end{aligned}$$

which implies that the matrix $\mathbb{W}$ is given by

$$\mathbb{W} = \begin{pmatrix} -\omega_{21} & \omega_{12} \\ \omega_{21} & -\omega_{12} \end{pmatrix} .$$

Each column sums to zero, as expected.

15.22 Work extraction from a finite-time protocol. We explore numerically and analytically finite-time protocols of duration τ for a two-state Szilard engine.

a. Show that the master equation reduces to a single differential equation for $p(t)$, the probability to be in the initially unoccupied state, and $\epsilon(t)$, its energy level.
b. Integrate the master equation for the protocol $\epsilon(t) = \epsilon_0(1 - t/\tau)$ and evaluate the average work over the protocol. Confirm that $\langle W \rangle$ is maximized for $\tau \to \infty$.
c. Show numerically that the average power extracted, $\langle P \rangle \equiv \langle W \rangle / \tau$, is maximized for $\tau \to 0$. Given that, show analytically that $\epsilon_0 \approx 0.3$ maximizes $\langle P \rangle$.

Solution.

a. We begin by writing the master equation for a two-state system:

$$\begin{aligned} \mathrm{d}_t p_1 &= -\omega_{21} p_1 + \omega_{12} p_2 \\ \mathrm{d}_t p_2 = -\dot{p}_1 &= +\omega_{21} p_1 - \omega_{12} p_2 \,, \end{aligned}$$

where $\omega_{21}/\omega_{12} = \mathrm{e}^{-\epsilon}$. With a scaled time $t \to \Gamma t$, we can write $\omega_{21} = \mathrm{e}^{-\epsilon}$ and $\omega_{12} = 1$. Then, using $p_2 = 1 - p_1$ and writing $p \equiv p_1$, we have, finally,

$$\mathrm{d}_t p = \mathrm{e}^{-\epsilon(t)}(1 - p) - p \,, \qquad p(0) = 0 \,,$$

where $\epsilon(t) = \epsilon_0(1 - t/\tau)$.

b. We can solve this equation numerically for a finite-time protocol $\epsilon(t) = \epsilon_0(1 - t/\tau)$ over a cycle of duration τ. The expression for the average energy extracted as work during the protocol is given by

$$W = -\int_0^\tau \mathrm{d}\epsilon(t)\, p(t, \epsilon(t)) = +\frac{\epsilon_0}{\tau} \int_0^\tau \mathrm{d}t\, p(t, \epsilon(t)) \,.$$

The sign is chosen so that $W > 0$ implies that a positive energy has been extracted from the heat bath. For numerical results, see the code on book website. You should be able to reproduce the image given in the text and to verify that curves tend to

$$W(\epsilon, \tau \to \infty) = \ln \frac{2}{1 + \mathrm{e}^{-\epsilon}} \,.$$

c. From the numerics in (b), it is clear that the average power always decreases with τ and hence is greatest at $\tau \to 0$. Similarly, you can readily check that $p(t) \ll 1$ for all t in the range $0 \le t \le \tau$, as $\tau \to 0$. The master equation for $p(t)$ is then, approximately,

$$\mathrm{d}_t p \approx \mathrm{e}^{-\epsilon_0(1 - t/\tau)} = \mathrm{e}^{-\epsilon_0}\, \mathrm{e}^{\epsilon_0 t/\tau} \,.$$

Integrating, we find

$$p(t) = \frac{\mathrm{e}^{-\epsilon_0}\, \tau}{\epsilon_0} \left(\mathrm{e}^{\epsilon_0 t/\tau} - 1 \right) .$$

and an average extracted work

$$W = \frac{\epsilon_0}{\tau}\int_0^\tau \mathrm{d}t\, p(t) = \frac{\epsilon_0}{\tau}\frac{\mathrm{e}^{-\epsilon_0}\,\tau}{\epsilon_0}\int_0^\tau \mathrm{d}t\left(\mathrm{e}^{\epsilon_0 t/\tau}-1\right) = \tau\left[\frac{1}{\epsilon_0}\left(1-\mathrm{e}^{-\epsilon_0}\right)-\mathrm{e}^{-\epsilon_0}\right].$$

The power is then

$$P = \left[\frac{1}{\epsilon_0}\left(1-\mathrm{e}^{-\epsilon_0}\right)-\mathrm{e}^{-\epsilon_0}\right],$$

which, numerically has a maximum $P_{\max} \approx 0.298$ for $\epsilon_0 \approx 1.79$. This matches the maximum found numerically from the full equations, with no approximation.

The main point is that while the maximum power per cycle is extracted from long, quasistatic protocols, the maximum power (here) is extracted for very fast cycles. Note that we have still assumed a particular protocols, $\epsilon(t) = \epsilon_0(1-t/\tau)$. Bauer et al. (2014) use the calculus of variations to find the optimal protocol function, which turns out to be to let ϵ jump from 0 to 1 and then back to 0 rather than ramping down. The maximum possible power is then $P_{\max} = 1/e \approx 0.368$, which is somewhat better than what we found with our restricted protocol.

15.23 Entropy production in stochastic thermodynamics.

a. Derive the entropy decomposition $\mathrm{d}_t S = \dot{S}_\mathrm{i} + \dot{S}_\mathrm{e}$ by showing (or identifying)

$$\mathrm{d}_t S = \frac{1}{2}\sum_{ij} J_{ij}\ln\frac{p_j}{p_i}, \quad \dot{S}_\mathrm{i} = \frac{1}{2}\sum_{ij} J_{ij}\ln\frac{W_{ij}p_j}{W_{ji}p_i}, \quad \dot{S}_\mathrm{e} = \frac{1}{2}\sum_{ij} J_{ij}\ln\frac{W_{ji}}{W_{ij}}.$$

b. Show that $\dot{Q} = T\dot{S}_\mathrm{e}$ is consistent with the definition $\dot{Q} = \sum_i(\mathrm{d}_t p_i)\epsilon_i$ used in the first law. For help on this problem, see Van den Broeck and Esposito (2015).

Solution.

a. The time derivative of the entropy (in units of k_B) is

$$\mathrm{d}_t S = -\sum_i(\mathrm{d}_t p_i)\ln p_i - \sum_i p_i(\mathrm{d}_t \ln p_i).$$

The second term vanishes because of probability normalization:

$$\begin{aligned}\sum_i p_i(\mathrm{d}_t \ln p_i) &= \sum_i p_i\left(\frac{1}{p_i}\right)\mathrm{d}_t p_i \\ &= (1)\,\mathrm{d}_t\left(\sum_i p_i\right) \\ &= \mathrm{d}_t(1) = 0.\end{aligned}$$

Thus,

$$\begin{aligned} \mathrm{d}_t S &= -\sum_i (\mathrm{d}_t p_i) \ln p_i \\ &= -\sum_{ij} \left(\mathbb{W}_{ij} p_j\right) \ln p_i \\ &= -\sum_{ij} \left(\mathbb{W}_{ij} p_j\right) \ln \frac{p_i}{p_j} . \end{aligned}$$

This last step uses the property that $\sum_i \mathbb{W}_{ij} = 0$. Explicitly, the new term is

$$\sum_{ij} \mathbb{W}_{ij} p_j \ln p_j = \sum_j p_j \ln p_j \underbrace{\left(\sum_i \mathbb{W}_{ij}\right)}_{0} = 0 .$$

Next, we recognize that indices that are summed are dummy variables and can have any name. If we switch index names $i \leftrightarrow j$, we thus get the same sum. Noting that the diagonal term then drops out, so that $\mathbb{W}_{ij} = W_{ij}$, we can thus can re-express the entropy rate as

$$\mathrm{d}_t S = +\frac{1}{2} \sum_{ij} \left(W_{ij} p_j - W_{ji} p_i\right) \ln \frac{p_j}{p_i} = \frac{1}{2} \sum_{ij} J_{ij} \ln \frac{p_j}{p_i} ,$$

where we also use $\ln\left(p_j/p_i\right) = -\ln\left(p_i/p_j\right)$ and recall that we have previously defined the current from j to i as $J_{ij} = W_{ij} p_j - W_{ji} p_i$. This is the desired expression for $\mathrm{d}_t S$.

Finally, writing

$$\ln \frac{p_j}{p_i} = \ln \frac{W_{ij} p_j}{W_{ji} p_i} + \ln \frac{W_{ji}}{W_{ij}} ,$$

we arrive at the desired result: $\mathrm{d}_t S = \dot{S}_\mathrm{e} + \dot{S}_\mathrm{i}$, with

$$\dot{S}_\mathrm{i} = \frac{1}{2} \sum_{ij} J_{ij} \ln \frac{W_{ij} p_j}{W_{ji} p_i} \quad \text{and} \quad \dot{S}_\mathrm{e} = \dot{Q}/T = \frac{1}{2} \sum_{ij} J_{ij} \ln \frac{W_{ji}}{W_{ij}} .$$

b. We start from the definition used in the first law:

$$\begin{aligned} \dot{Q} &= \sum_i (\mathrm{d}_t p_i) \epsilon_i \\ &= \sum_{ij} \left(\mathbb{W}_{ij} p_j\right) \epsilon_i \\ &= \sum_{ij} \left(\mathbb{W}_{ij} p_j\right) \frac{\epsilon_i}{\epsilon_j} \\ &= \frac{1}{2} \sum_{ij} J_{ij} \frac{\epsilon_i}{\epsilon_j} \\ &= \frac{1}{2} \sum_{ij} J_{ij} \ln \frac{W_{ji}}{W_{ij}} . \end{aligned}$$

Thus, the two definitions of $\dot{Q}$ are equivalent.

15.24 Nonequilibrium free energy

a. Show that $F(p) = -k_B T \ln Z$ when p is the equilibrium distribution $\boldsymbol{\pi}$.
b. Show that $F = F^{\text{eq}} + D(p\|\pi)$, for relative entropy $D(p\|\pi) \equiv \sum_i p_i \ln(p_i/\pi_i) \geq 0$.
c. Generalize the result in (b) to one-dimensional continuous distributions, with $p_i \to p(x)$ and $\sum_i \to \int \mathrm{d}x$. Assume overdamped dynamics, so that $\epsilon_i \to U(x)$.
d. Show that the protocol of Figure 15.7 extracts a work $W_{\text{extract}} = F - F^{\text{eq}}$.
e. Show that the irreversible protocol dissipates heat $Q = F - F^{\text{eq}}$ into the bath.

Solution.

a. The equilibrium distribution is

$$\pi_i = \frac{1}{Z} \mathrm{e}^{-\epsilon_i} ,$$

where the partition function,

$$Z = \sum_i \mathrm{e}^{-\epsilon_i} ,$$

ensures normalization: $\sum_i \pi_i = 1$. We then write the definition of nonequilibrium free energy using the equilibrium distribution:

$$\begin{aligned} F^{\text{eq}} &= \sum_i \pi_i \epsilon_i + \pi_i \ln \pi_i \\ &= \sum_i \pi_i \epsilon_i + \pi_i(-\epsilon_i - \ln Z) \\ &= -\left(\sum_i \pi_i \right) \ln Z \\ &= -\ln Z . \end{aligned}$$

Going back to unscaled (physical) units gives $F^{\text{eq}} = -k_B T \ln Z$. This is one traditional definition of the equilibrium free energy.

b. The difference between nonequilibrium and equilibrium free energy is

$$\begin{aligned} F - F^{\text{eq}} &= \sum_i \epsilon_i p_i + p_i \ln p_i - \epsilon_i \pi_i - \pi_i \ln \pi_i \\ &= \epsilon_i (p_i - \pi_i) + p_i \ln p_i - \pi_i \ln \pi_i - p_i \ln \pi_i + p_i \ln \pi_i \\ &= \epsilon_i (p_i - \pi_i) + \ln \pi_i (p_i - \pi_i) + p_i \ln \frac{p_i}{\pi_i} \\ &= \epsilon_i (p_i - \pi_i) + (-\epsilon_i - \ln Z)(p_i - \pi_i) + p_i \ln \frac{p_i}{\pi_i} \\ &= p_i \ln \frac{p_i}{\pi_i} \\ &= D(p\|\pi) . \end{aligned}$$

The terms proportional to $\ln Z$ vanish because both p and π are normalized distributions. Notice that when $p = \pi$, we immediately see that $F = F^{\text{eq}}$, since $D(\pi\|\pi) = 0$ for an arbitrary distribution π.

c. For probability density functions $p(x)$ and its equilibrium counterpart $\pi(x)$,

$$F - F^{\text{eq}} = \int \mathrm{d}x \, [U(x)p(x) + p(x) \ln p(x) - U(x)\pi(x) - \pi(x) \ln \pi(x)] \, .$$

In this expression $U(x)$ is the system energy (all potential energy) when the particle state is x. Because the system is overdamped and one dimensional, the state is characterized by the single continuous variable x. Otherwise, we would need a two-dimensional state, with the second component representing the momentum. In this case, though, the equilibrium distribution is $\pi(x) = (1/Z)\, \mathrm{e}^{-U(x)}$, or $\ln \pi = -U - \ln Z$.

The notation is cleaner if we drop the x dependence from the functions.

$$\begin{aligned} F - F^{\text{eq}} &= \int \mathrm{d}x \, [U p + p \ln p - U\pi - \pi \ln \pi] \\ &= \int \mathrm{d}x \, [U(p - \pi) + p \ln p - \pi \ln \pi - p \ln \pi + p \ln \pi] \\ &= \int \mathrm{d}x \, [U(p - \pi) + (\ln \pi)(p - \pi) + p \ln \frac{p}{\pi}] \\ &= \int \mathrm{d}x \, [U(p - \pi) + (-U - \ln Z)(p - \pi) + p \ln \frac{p}{\pi}] \\ &= \int \mathrm{d}x \, p \ln \frac{p}{\pi} \, . \end{aligned}$$

We again use the normalization conditions, $\int \mathrm{d}x \, p(x) = \int \mathrm{d}x \, \pi(x) = 1$. Thus, the derivation proceeds entirely analogously to the discrete case, and again we find

$$F - F^{\text{eq}} = D(p \| \pi) \geq 0 \, ,$$

for the continuous probability density functions $p(x)$ and $\pi(x)$.

d. The trick is to define notation carefully. Let $U_0(x)$ be the potential shown at the top of Figure 15.7. Its equilibrium distribution is $\pi(x)$. Similarly, let $U_p(x)$ be the potential after the quench, shown at bottom left. It is chosen to be the potential whose equilibrium distribution is $p(x)$, the initial nonequilibrium state in the protocol. The quench thus requires work

$$W_{\text{quench}} = \langle U_p \rangle_p - \langle U_0 \rangle_p \, ,$$

where the angle brackets $\langle \cdot \rangle_p$ denote the ensemble average with respect to the distribution $p(x)$. In particular, $\langle U_p \rangle_p = \int \mathrm{d}x \, p(x) U_p(x)$, and $\langle U_0 \rangle_p = \int \mathrm{d}x \, p(x) U_0(x)$.

In the second step of the protocol, we make a quasistatic transformation between the two equilibrium states. The work done is the difference in equilibrium free energies:

$$F_0^{\text{eq}} - F_p^{\text{eq}} = F_0^{\text{eq}} - \langle U_p \rangle_p + TS(p) \, ,$$

where the second term uses the definition of the equilibrium free energy for $p(x)$. Putting these two contributions together, the work required to carry out the protocol is

$$\begin{aligned} W &= \left(\langle U_p \rangle_p - \langle U_0 \rangle_p\right) + \left(F_0^{\text{eq}} - \langle U_p \rangle_p + TS(p)\right) \\ &= F_0^{\text{eq}} - \langle U_0 \rangle_p + TS(p) \\ &= F_0^{\text{eq}} - F_0(p) \,. \end{aligned}$$

More intuitively, the work *extracted* from the heat bath is the negative of this:

$$W_{\text{extract}} = -W = F - F^{\text{eq}} \,,$$

where we have simplified the notation since both free energies are referenced to the equilibrium distribution of the start (and end) potential $U_0(x)$. If the second step of the protocol is carried out in a finite time, then we would conclude that $W_{\text{extract}} < F - F^{\text{eq}}$. Some of the energy is returned to the heat bath.

e. In Figure 15.7, the "do nothing" protocol simply lets the nonequilibrium distribution $p(x)$ relax to the equilibrium $\pi(x)$. Since the potential is unchanged, no work is done on the system. The heat transfer is just $T\Delta S_{\text{tot}}$, where ΔS_{tot} is the total entropy change, which includes the change to the system (particle in potential) and the surrounding heat bath. This is given by

$$\begin{aligned} Q = T\Delta S_{\text{tot}} &= \underbrace{T\left[S(\pi) - S(p)\right]}_{\text{system}} + Q_{\text{relax}} \\ &= \underbrace{T\left[S(\pi) - S(p)\right]}_{\text{system}} - \underbrace{\left[\langle U \rangle_\pi + \langle U \rangle_p\right]}_{\text{bath}} \\ &= F - F^{\text{eq}} \,. \end{aligned}$$

In the second line, $-Q_{\text{relax}} = \langle U \rangle_\pi - \langle U \rangle_p$ results from the first law. Remember that the unchanging potential means no work is done during the relaxation and also that the first law $\Delta E = W + Q$ refers all energies to the system. We thus have $Q_{\text{relax}} = -Q$ because we need to compute the heat transferred *to the bath*.

15.25 Bipartite system. Analyze aspects of their dynamics and thermodynamics.

a. Show that the information flow is given by Eq. (15.90).
b. Derive the decomposition of entropy given in Eq. (15.91). Give explicit expressions for all components and prove that $\dot{S}_{\text{i}}^X$ and $\dot{S}_{\text{i}}^Y$ are each nonnegative.
c. Imagine that, not knowing about System Y, you tried to define an X-only "entropy-production rate" $\sigma_{\text{i}}^X = \frac{1}{2}\sum_{xy,x'} J_{xx'}^y \frac{W_{xx'}^y p_{x'}}{W_{x'x}^y p_x}$. Explain mathematically and physically how σ_{i}^X can be negative.

Solution.

a. The mutual information between states x and y is given (in nats) by

$$I^{XY} = \sum_{xy} p_{xy} \ln \frac{p_{xy}}{p_x p_y} \geq 0 \,.$$

We then use the master equation $\mathrm{d}_t p_{xy} = \sum_{x'y'} J^{yy'}_{xx'}$ to write the time derivative as

$$\begin{aligned} \mathrm{d}_t I^{XY} &= \sum_{xy} \left(\mathrm{d}_t p_{xy}\right) \ln \frac{p_{xy}}{p_x p_y} + \sum_{xy} p_{xy} \left(\frac{1}{p_{xy}} \mathrm{d}_t p_{xy} - \frac{1}{p_x} \mathrm{d}_t p_x - \frac{1}{p_y} \mathrm{d}_t p_y \right) \\ &= \sum_{xy} \left(\mathrm{d}_t p_{xy}\right) \ln \frac{p_{xy}}{p_x p_y} \\ &= \sum_{xy,x'y'} J^{yy'}_{xx'} \ln \frac{p_{xy}}{p_x p_y} \,. \end{aligned}$$

In the first step, we perform the sums before the time derivative and use the normalization of the probability distributions, noting that $\sum_{xy} p_{xy} = 1$. Similarly, $\sum_y p_{xy}/p_x = \sum_y p_{y|x} = 1$ and $\sum_x p_{xy}/p_y = \sum_x p_{x|y} = 1$, where we have introduced the conditional probabilities: $p_{xy} = p_{y|x} p_x = p_{x|y} p_y$.

Following Eq. (15.88), we use the bipartite structure of J to write

$$\sum_{x'y'} J^{yy'}_{xx'} = \sum_{x'} J^{y}_{xx'} + \sum_{y'} J^{yy'}_{x}$$

so that

$$\mathrm{d}_t I^{XY} = \left(\sum_{xy,x'} J^{y}_{xx'} + \sum_{xy,y'} J^{yy'}_{x} \right) \ln \frac{p_{xy}}{p_x p_y} \,.$$

From the definition $J^y_{xx'} \equiv W^y_{xx'} p_{x'y} - W^y_{x'x} p_{xy}$, we see that $J^y_{xx'} = -J^y_{x'x}$. Thus,

$$\begin{aligned} \sum_{xy,x'} J^y_{xx'} \ln \frac{p_{xy}}{p_x p_y} &= \frac{1}{2} \sum_{xy,x'} \left(J^y_{xx'} - J^y_{x'x} \right) \ln \frac{p_{xy}}{p_x p_y} \\ &= \left(\frac{1}{2} \sum_{xy,x'} J^y_{xx'} \ln \frac{p_{xy}}{p_x p_y} \right) - \left(\frac{1}{2} \sum_{x'y,x} J^y_{xx'} \ln \frac{p_{x'y}}{p_{x'} p_y} \right) \quad \text{(swap } x \leftrightarrow x' \text{ in 2nd sum)} \\ &= \left(\frac{1}{2} \sum_{xy,x'} J^y_{xx'} \ln \frac{p_{y|x}}{p_y} \right) - \left(\frac{1}{2} \sum_{xy,x'} J^y_{xx'} \ln \frac{p_{y|x'}}{p_y} \right) \\ &= \left(\frac{1}{2} \sum_{xy,x'} J^y_{xx'} \ln \frac{p_{y|x}}{p_{y|x'}} \right) \\ &\equiv \dot{I}^X \,. \end{aligned}$$

Similarly, from the definition $J_x^{yy'} \equiv W_x^{yy'} p_{xy'} - W_x^{y'y} p_{xy}$, we see that $J_x^{yy'} = -J_x^{y'y}$ and thus

$$\sum_{xy,y'} J_x^{yy'} \ln \frac{p_{xy}}{p_x p_y} = \cdots = \frac{1}{2} \sum_{xy,y'} J_x^{yy'} \ln \frac{p_{x|y}}{p_{x|y'}} \equiv \dot{I}^Y .$$

Putting both terms together gives $\mathrm{d}_t I^{XY} = \dot{I}^X + \dot{I}^Y$.

b. The entropy production of the joint system is

$$\dot{S}_{\mathrm{i}}^{XY} = \frac{1}{2} \sum_{xy,x'y'} J_{xx'}^{yy'} \ln \frac{W_{xx'}^{yy'} p_{x'y'}}{W_{x'x}^{y'y} p_{xy}} \geq 0 ,$$

non-negative because each term in the sum is of the form $(x - y)\ln(x/y) \geq 0$. Then, using the bipartite structure of the currents and rates, we write

$$\begin{aligned}
\dot{S}_{\mathrm{i}}^{XY} &= \frac{1}{2} \sum_{xy,x'y'} \left[J_{xx'}^{yy'} \left(\ln W_{xx'}^{yy'} p_{x'y'} - \ln W_{x'x}^{y'y} p_{xy} \right) \right] \\
&= \frac{1}{2} \left[\sum_{xy,x'} J_{xx'}^{y} \left(\ln W_{xx'}^{y} p_{x'y} - \ln W_{x'x}^{y} p_{xy} \right) + \sum_{xy,y'} J_{x}^{yy'} \left(\ln W_{x}^{yy'} p_{xy'} - \ln W_{x}^{y'y} p_{xy} \right) \right] \\
&= \underbrace{\frac{1}{2} \sum_{xy,x'} J_{xx'}^{y} \ln \frac{W_{xx'}^{y} p_{x'y}}{W_{x'x}^{y} p_{xy}}}_{\dot{S}_{\mathrm{i}}^{X}} + \underbrace{\frac{1}{2} \sum_{xy,y'} J_{x}^{yy'} \ln \frac{W_{x}^{yy'} p_{xy'}}{W_{x}^{y'y} p_{xy}}}_{\dot{S}_{\mathrm{i}}^{Y}} .
\end{aligned}$$

Recalling that

$$J_{xx'}^{y} \equiv W_{xx'}^{y} p_{x'y} - W_{x'x}^{y} p_{xy} \quad \text{and} \quad J_{x}^{yy'} \equiv W_{x}^{yy'} p_{xy'} - W_{x}^{y'y} p_{xy} ,$$

we see that $\dot{S}_{\mathrm{i}}^{X} \geq 0$ and $\dot{S}_{\mathrm{i}}^{Y} \geq 0$ for the same reason that $\dot{S}_{\mathrm{i}}^{XY} \geq 0$. Thus, the overall entropy production can be split into X and Y contributions that are *each* separately non-negative.

We now rewrite the log term in $\dot{S}_{\mathrm{i}}^{X}$:

$$\ln \frac{W_{xx'}^{y} p_{x'y}}{W_{x'x}^{y} p_{xy}} = \ln \frac{W_{xx'}^{y} p_{x'|y}}{W_{x'x}^{y} p_{x|y}} = \ln \frac{W_{xx'}^{y} p_{x'}}{W_{x'x}^{y} p_{x}} - \ln \frac{p_{x|y} p_{x'}}{p_{x'|y} p_{x}} .$$

Next, we use Bayes' Theorem to write

$$\frac{p_{x|y} p_{x'}}{p_{x'|y} p_{x}} = \frac{p_{y|x} p_{x} p_{x'} p_{y}}{p_{y|x'} p_{x'} p_{x} p_{y}} = \frac{p_{y|x}}{p_{y|x'}} .$$

Then,

$$\begin{aligned}
\dot{S}_{\mathrm{i}}^{X} &= \frac{1}{2} \sum_{xy,x'} J_{xx'}^{y} \ln \frac{W_{xx'}^{y} p_{x'}}{W_{x'x}^{y} p_{x}} - \frac{1}{2} \sum_{xy,x'} J_{xx'}^{y} \ln \frac{p_{y|x}}{p_{x'|y}} \\
&= \underbrace{\frac{1}{2} \sum_{xy,x'} J_{xx'}^{y} \ln \frac{p_{x'}}{p_{x}}}_{\mathrm{d}_t S^X} - \underbrace{\frac{1}{2} \sum_{xy,x'} J_{xx'}^{y} \ln \frac{W_{x'x}^{y}}{W_{xx'}^{y}}}_{\dot{S}_{\mathrm{e}}^{X}} - \underbrace{\frac{1}{2} \sum_{xy,x'} J_{xx'}^{y} \ln \frac{p_{y|x}}{p_{y|x'}}}_{\dot{I}^X} .
\end{aligned}$$

Collecting the terms gives

$$\dot{S}^X_{\rm i} = {\rm d}_t S^X - \dot{S}^X_{\rm e} - \dot{I}^X \geq 0\,,$$

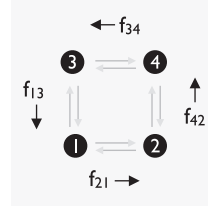
which is what we set out to prove. The Y term is handled analogously. Notice that the X terms involve an average over Y and vice versa. When we define X-only terms such as ${\rm d}_t S^X$, etc., we always implicitly assume that all "latent" (unknown, from the point of view of X) terms are averaged over.

c. The imagined "entropy production rate" $\sigma^X_{\rm i} \equiv {\rm d}_t S^X - \dot{S}^X_{\rm e}$ is

$$\begin{aligned}\sigma^X_{\rm i} &= \frac{1}{2}\sum_{xy,x'} J^y_{xx'}\frac{W^y_{xx'}p_{x'}}{W^y_{x'x}p_x} \\ &= \frac{1}{2}\sum_{xy,x'}\left(W^y_{xx'}p_{x'y} - W^y_{x'x}p_{xy}\right)\frac{W^y_{xx'}p_{x'}}{W^y_{x'x}p_x}\,.\end{aligned}$$

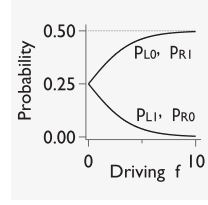
Mathematically, there is no reason for this sum to be positive, as, in order to be of the canonical form $(x-y)\ln(x/y)$, the probabilities in the log term would have to be $p_{x'y}$ and p_{xy}. These are simply different from $p_{x'}$ and p_x. Perhaps a more intuitive expression is to use $p_{x'y}/p_{xy} = p_{x'|y}/p_{x|y}$ and to note that p_x is different from $p_{x|y}$, etc. Physically, we have given the reason in the text: the true entropy production rate has an additional term representing flows of information to and from the hidden system.

15.26 Four-state bipartite system. The system illustrated at right copies that of Figure 15.8 but relabels the states as $\{1,2,3,4\}$, to simplify the analysis as a single joint system. In addition, we add up to four nonequilibrium driving potentials f_{21}, f_{42}, f_{34}, f_{13} (all in units of $k_{\rm B}T$) going from $2 \to 1$, etc. The $W_{ij} = 1$ for the "base" rates (light forward-backward arrows). When nonequilibrium driving is present, the rates are modified to $W_{21} = {\rm e}^{f_{21}/2}$ and $W_{12} = {\rm e}^{-f_{21}/2}$, so that $W_{21}/W_{12} = {\rm e}^{f_{21}}$, etc.

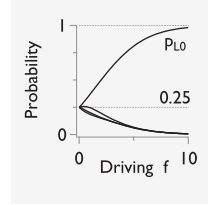


a. Write down the master equation for the joint bipartite system. Show that the steady-state solution has $p_i = \frac{1}{4}$ when all driving terms $f_{ij} = 0$.

b. When the driving terms are present, show that the entropy production rate is $\dot{S}^{XY}_{\rm i} = J(f_{21} + f_{42} + f_{34} + f_{13})$, where J is the current around the loop. Similarly, show that $\dot{S}^X_{\rm i} = J(f_{21} + f_{34})$ and $\dot{S}^Y_{\rm i} = J(f_{42} + f_{13})$, thus confirming $\dot{S}^{XY}_{\rm i} = \dot{S}^X_{\rm i} + \dot{S}^Y_{\rm i}$. Finally, show that the information flow is $\dot{I}^X = -\dot{I}^Y = J\ln(p_{R0}p_{L1}\,/\,p_{L0}p_{R1})$.



c. Consider the sensor case where $f_{13} = f_{42} = f$ and $f_{21} = f_{34} = 0$. Show that the steady-state probabilities states are $\frac{1}{4}(1 \pm \tanh\frac{1}{4}f)$, as plotted at right. Show, too, that the steady-state current around the loop is $J = \frac{1}{2}\tanh(\frac{1}{4}f)$. The total dissipation rate to run the sensor is then $\dot{S}^{XY}_{\rm i} = 2fJ \approx \frac{1}{4}f^2$ for $f \ll 1$ and $\approx f$ for $f \gg 1$. Show that the information flow $\dot{I}^Y = Jf \geq 0$, as expected for a sensor.



d. Consider the regulator case, where $f_{13} = f_{42} = f_{34} = f$ and $f_{21} = 0$. Show that the steady-state probabilities are as shown at right. Find J. Use the latter

to find $\dot{S}_{\mathrm{i}}^{XY}$ and the information flow. Show that for $f \gg 1$, $\dot{S}_{\mathrm{i}}^{XY} \approx 3f$ and $\dot{I}^{X} \approx -f/2$.

Solution.

a. The rate matrix $\mathbb{W}$ has the following form:

$$\begin{pmatrix} -\mathrm{e}^{-f_{13}/2} - \mathrm{e}^{-f_{21}/2} & \mathrm{e}^{f_{21}/2} & \mathrm{e}^{f_{13}/2} & 0 \\ \mathrm{e}^{-f_{21}/2} & -\mathrm{e}^{f_{21}/2} - \mathrm{e}^{f_{42}/2} & 0 & \mathrm{e}^{-f_{42}/2} \\ \mathrm{e}^{-f_{13}/2} & 0 & -\mathrm{e}^{f_{13}/2} - \mathrm{e}^{-f_{34}/2} & \mathrm{e}^{f_{34}/2} \\ 0 & \mathrm{e}^{f_{42}/2} & \mathrm{e}^{-f_{34}/2} & -\mathrm{e}^{f_{34}/2} - \mathrm{e}^{-f_{42}/2} \end{pmatrix}$$

We note some general points:

i. The cross-diagonal terms are all zero because of the bipartite structure.
ii. The diagonal terms are simply minus the sum of the other column components. They are chosen so that each column sums to zero, conserving probability.

With all $f_{ij} = 0$, the rate matrix simplifies to

$$\mathbb{W} = \begin{pmatrix} -2 & 1 & 1 & 0 \\ 1 & -2 & 0 & 1 \\ 1 & 0 & -2 & 1 \\ 0 & 1 & 1 & -2 \end{pmatrix}.$$

By inspection, the $\mathbb{W}\boldsymbol{p}_{\mathrm{eq}} = \mathbf{0}$ for

$$\boldsymbol{p}_{\mathrm{eq}} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix},$$

meaning that $p_i = \frac{1}{4}$ after normalization.

b. In single-index notation, the entropy production is

$$\dot{S}_{\mathrm{i}} = \sum_{i<j} J_{ij} \ln \frac{W_{ij} p_j}{W_{ji} p_i},$$

where we revert to the restricted sum so that we do not double-count terms. Because there is a single loop, the non-zero currents are $J_{12} = J_{42} = J_{34} = J_{13} = J$, which can be factored out of the sum. We then have

$$\dot{S}_{\mathrm{i}} = J \sum_{i<j} \left(\ln \frac{W_{ij}}{W_{ji}} + \ln \frac{p_j}{p_i} \right),$$

For the first term in the sum, each one gives f_{ij} for each non-zero contribution. That is, we have $f_{12} + f_{42} + f_{34} + f_{13}$. As for the other term, it is

$$\ln \frac{p_2 p_4 p_3 p_1}{p_1 p_2 p_4 p_3} = \ln 1 = 0.$$

In this simple case, the decomposition into X and Y means simply summing the current times the forces in X and Y, respectively, which immediately gives the result for this problem.

For the information flow, we write the index sum explicitly for the X subsystem:

$$\begin{aligned}\dot{I}^X &= \frac{1}{2}\sum_{xy,x'} J^y_{xx'} \ln \frac{p_{y|x}}{p_{y|x'}} \\ &= J^0_{LR} \ln \frac{p_{0|L}}{p_{0|R}} + J^1_{LR} \ln \frac{p_{1|L}}{p_{1|R}} \\ &= (-J) \ln \frac{p_{0|L}}{p_{0|R}} + J \ln \frac{p_{1|L}}{p_{1|R}} \\ &= J \ln \frac{p_{1|L} p_{0|R}}{p_{1|R} p_{0|L}} \\ &= J \ln \frac{p_{L1} p_{R0}}{p_{R1} p_{L0}} .\end{aligned}$$

Conversely,

$$\dot{I}^Y = -\dot{I}^X = J \ln \frac{p_{R1} p_{L0}}{p_{L1} p_{R0}} .$$

A few comments: The sum over x, x' has but one term (LR here). The terms LL and RR are zero because the corresponding J is always zero. The term RL is equal (and compensates for the $1/2$ factor). In going from conditional to joint probabilities, we multiply and divide by factors of p_L and p_R, as needed. Those factors all cancel in the end, if you have your signs correct! Notice, in particular, that the summation convention combined with the sign convention for currents means that J enters with opposite sign in the two terms.

c. By symmetry, it is clear that $p_{L0} = p_{R1}$ or $p_1 = p_4$ in the single-index notation. Likewise, $p_{L1} = p_{R0}$, or $p_2 = p_3$. With this condition and setting $f_{13} = f_{42} = f$ and $f_{21} = f_{34} = 0$, the steady-state equations are

$$\begin{pmatrix} -1 - \mathrm{e}^{-f/2} & 1 & \mathrm{e}^{f/2} & 0 \\ 1 & -1 - \mathrm{e}^{-f/2} & 0 & \mathrm{e}^{-f/2} \\ \mathrm{e}^{-f/2} & 0 & -1 - \mathrm{e}^{-f/2} & 1 \\ 0 & \mathrm{e}^{f/2} & -1 - \mathrm{e}^{-f/2} & \end{pmatrix} \begin{pmatrix} p \\ 1-p \\ 1-p \\ p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} .$$

We can solve, for example, the first equation (and divide p and $1 - p$ by 2 for normalization).

To find the steady-state current J around the loop, we note that every edge has the same current. It will be easier here to use single-index notation. We calculate the flux through the "bottom" edge, from $1 \to 2$:

$$\begin{aligned} J = J_{21} &= W_{21} p_1 - W_{12} p_2 \\ &= p_1 - p_2 \end{aligned}$$

$$= \tfrac{1}{4}\left(1 + \tanh \tfrac{1}{4}f\right) - \tfrac{1}{4}\left(1 \pm \tanh \tfrac{1}{4}f\right)$$
$$= \tfrac{1}{2}\tanh\left(\tfrac{1}{4}f\right).$$

For the information flow, we calculate the Y term as

$$\begin{aligned} \dot{I}^Y &= J \ln \frac{p_{R1}p_{L0}}{p_{L1}p_{R0}} \\ &= 2J \ln \frac{1 + \tanh \frac{1}{4}f}{1 - \tanh \frac{1}{4}f} \\ &= 2J\tfrac{1}{2}f = Jf\,. \end{aligned}$$

The factor of 2 in the second line results from $p_{R1} = p_{L0}$ and similarly in the denominator. The square in the log becomes the 2. Going from line 2 to 3 requires writing out the tanh in terms of exponentials.

15.27 Voltage fluctuations in an RCR circuit.

a. By finding $V(t)$ for arbitrary driving function $\eta(t)$, evaluate $\langle V(t)\,V(0)\rangle$ for an RC circuit and thereby derive Eq. (15.94).
b. Find $\langle V^2\rangle$ for an RC circuit using Eq. (8.50) for the evolution of the variance.
c. Generalize to the two-resistor case, Eq. (15.97).

Solution.

a. The equation of motion for the current in the single-resistor case is

$$C\dot{V} = -\frac{V}{R} + I + \sqrt{\frac{2k_BT}{R}}\,\eta(t)\,,$$

where the fluctuating term is the Johnson-Nyquist noise at temperature T. Treating $\eta(t)$ as "just a function" and neglecting initial conditions (which die away exponentially and are irrelevant for long-time statistics), we integrate to find

$$V(t) = \frac{\sqrt{2k_BTR}}{\tau}\,\mathrm{e}^{-t/\tau}\int_{\infty}^{t} \mathrm{d}t'\,\mathrm{e}^{t'/\tau}\,\eta(t')\,,$$

with $\tau = RC$. Multiplying by $V(0)$ and ensemble averaging over the noise gives

$$\begin{aligned} \langle V(t)\,V(0)\rangle &= \frac{2k_BTR}{\tau^2}\,\mathrm{e}^{-t/\tau}\int_{-\infty}^{t}\mathrm{d}t'\int_{-\infty}^{0}\mathrm{d}t''\,\mathrm{e}^{(t'+t'')/\tau}\,\underbrace{\langle\eta(t')\,\eta(t'')\rangle}_{\delta(t'-t'')} \\ &= \frac{2k_BTR}{\tau^2}\,\mathrm{e}^{-t/\tau}\int_{-\infty}^{0}\mathrm{d}t''\,\mathrm{e}^{2t''/\tau} \\ &= \frac{k_BTR}{\tau}\,\mathrm{e}^{-t/\tau} \\ &= \frac{k_BT}{C}\,\mathrm{e}^{-t/\tau}\,. \end{aligned}$$

Setting $t = 0$ then gives $\langle V^2 \rangle = k_B T/C$. Thus, in accordance with the Equipartition Theorem, the average energy stored in the capacitor is

$$\frac{1}{2} C \left\langle V^2 \right\rangle = \frac{1}{2} k_B T \,.$$

b. Equation (8.50) for the evolution of the variance in a general linear system gives a much faster derivation. Since there are no observations (here),

$$\frac{\mathrm{d}\boldsymbol{P}}{\mathrm{d}t} = \underbrace{\boldsymbol{A}_c \boldsymbol{P} + \boldsymbol{P} \boldsymbol{A}_c^{\mathsf{T}}}_{\text{dynamics}} + \underbrace{\boldsymbol{Q}_v^c}_{\text{disturbances}} \,.$$

Here, we translate $\boldsymbol{P} \to \langle V^2 \rangle$, $\boldsymbol{A}_c \to -1/(RC) = -1/\tau$, and $\boldsymbol{Q}_v^c \to \frac{2 k_B T R}{\tau^2}$. Then the stationary variance equation is

$$-2 \frac{\langle V^2 \rangle}{\tau} + \frac{2 k_B T R}{\tau^2} = 0 \,,$$

which implies, using $\tau = RC$,

$$\langle V^2 \rangle = \frac{k_B T R}{\tau} = \frac{k_B T}{C} \,.$$

Physically, a typical power is $P = \langle V^2 \rangle / R = k_B T/\tau$, meaning that energies of order $k_B T$ slosh in and out of the capacitor on time scales of order $\tau = RC$.

c. For the two-resistor case, the current obeys,

$$C \dot{V} = -\frac{V}{R} - \frac{V}{R'} + \sqrt{\frac{2 k_B T}{R}}\, \eta(t) + \sqrt{\frac{2 k_B T'}{R'}}\, \eta'(t) \,,$$

with independent noise sources $\eta(t)$ and $\eta'(t)$. Rewriting, we have

$$\dot{V} = -\frac{V}{\tau_{\text{eff}}} + \sqrt{\frac{2 k_B T R}{\tau}}\, \eta(t) + \sqrt{\frac{2 k_B T' R'}{\tau'}}\, \eta'(t) \,,$$

with $\tau_{\text{eff}}^{-1} = \tau^{-1} + \tau'^{-1}$, $\tau = RC$, and $\tau' = R'C$.

The simplest solution is to realize that the independence of the noise sources implies that we can simply add the variances of the two contributions. Thus,

$$\begin{aligned} \langle V^2 \rangle &= \tau_{\text{eff}} \left(\frac{k_B T R}{\tau^2} + \frac{k_B T' R'}{\tau'^2} \right) \\ &= \frac{\tau \tau'}{\tau + \tau'} \left(\frac{k_B T R}{\tau^2} + \frac{k_B T' R'}{\tau'^2} \right) \\ &= \frac{k_B}{C} \left(\frac{R'}{R + R'} T + \frac{R}{R + R'} T' \right) . \end{aligned}$$