Appendix Mathematics

In this appendix are the various mathematical concepts and results that are drawn upon in the body of the text. The intent is not to teach the mathematics but to define notation, list results that I assume you know, and distinguish between math and control theory.

To spell the last item out more fully, I want to distinguish between standard mathematical identities and ideas that are specific to control theory. Section A.10, on information theory, is longer than the others because the topic is still rarely taught in the physics curriculum. We have drawn our material on information theory from Cover and Thomas (2006), MacKay (2003), Gibson (2014), and Bialek (2012). These all give a much more complete and balanced description than I can give here.

If you are unfamiliar with a topic, please consult one of the references for the full story. Many good books cover "mathematics for physicists," including, at the undergraduate level, Shankar (1995) and Boas (2005) and, at the graduate level, Stone and Goldbart (2009). The brief overview of probability theory largely follows Wasserman (2004). For background on Bayesian inference, see Sivia and Skilling (2006) for an elementary introduction and von der Linden et al. (2014) for a thorough exposition. The proof showing that the conditional mean minimizes the mean square estimation error comes from von der Linden et al. (2014). For decision theory and its implications for estimating parameters and states, see Van Trees et al. (2013) (especially Chapter 4).

A.1 Linear Algebra and Calculus

A.1.1 Vector and Matrix Notation and Basics

We recall notation and basic facts concerning real vectors and matrices.

1. *Transpose*:

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \Longrightarrow \quad \boldsymbol{x}^{\mathsf{T}} = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix}. \tag{A.1}$$

For matrices, $\left(\boldsymbol{A}^{\mathsf{T}}\right)_{ij} = A_{ji}$, or,

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \quad \Longrightarrow \quad \boldsymbol{A}^{\mathsf{T}} = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix} \tag{A.2}$$

Note that $(\boldsymbol{ABC})^{\mathsf{T}} = \boldsymbol{C}^{\mathsf{T}}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{A}^{\mathsf{T}}$.

2. *Symmetric matrix*: $\boldsymbol{A} = \boldsymbol{A}^{\mathsf{T}}$ and must be square. An $n \times n$ matrix has n^2 components. If symmetric, it has at most $\frac{1}{2}n(n+1)$ independent coefficients.
3. *Antisymmetric matrix*:$\boldsymbol{A} = -\boldsymbol{A}^{\mathsf{T}}$ and must be square. There are $\frac{1}{2}n(n-1)$ independent coefficients, and the diagonal elements are zero.
4. *Trace*:

$$\mathrm{Tr}\ \boldsymbol{A} = \mathrm{Tr} \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} = \sum_{i=1}^{n} a_{ii}\,. \tag{A.3}$$

- The trace is invariant under cyclic permutation: $\mathrm{Tr}\ \boldsymbol{ABC} = \mathrm{Tr}\ \boldsymbol{CAB} = \mathrm{Tr}\ \boldsymbol{BCA}$.
- $\mathrm{Tr}\ \boldsymbol{A}$ is coordinate invariant: $\mathrm{Tr}\ \boldsymbol{A} = \mathrm{Tr}\ \boldsymbol{RA'R}^{-1} = \mathrm{Tr}\ \boldsymbol{A'R}^{-1}\boldsymbol{R} = \mathrm{Tr}\ \boldsymbol{A'}$.
- If $\boldsymbol{A}$ has n eigenvalues λ_i, $\mathrm{Tr}\ \boldsymbol{A} = \sum_{i=1}^{n} \lambda_i$.

5. *Determinant*: Let $\boldsymbol{A}$ be an $n \times n$ matrix of real elements a_{ij}. The determinant, $\det \boldsymbol{A}$, is a real number defined as

$$\det \boldsymbol{A} \equiv \sum_{i_1,i_2,\ldots,i_n} \varepsilon_{i_1,i_2,\ldots,i_n}\, a_{1,i_1} \cdots a_{n,i_n}\,, \tag{A.4}$$

where the *Levi-Civita symbol* $\varepsilon_{i_1,i_2,\ldots,i_n}$ is $+1$ for even permutations of $(1,2,\ldots,n)$ and -1 for odd permutations. For example, the determinant of a 2×2 matrix is

$$\underbrace{a_{11}a_{22}}_{\varepsilon_{12}} - \underbrace{a_{12}a_{21}}_{\varepsilon_{21}} \tag{A.5}$$

and of a 3×3 matrix is

$$\underbrace{a_{11}a_{22}a_{33}}_{\varepsilon_{123}} - \underbrace{a_{11}a_{23}a_{32}}_{\varepsilon_{132}} - \underbrace{a_{12}a_{21}a_{33}}_{\varepsilon_{213}} + \underbrace{a_{12}a_{23}a_{31}}_{\varepsilon_{231}} + \underbrace{a_{13}a_{21}a_{32}}_{\varepsilon_{312}} - \underbrace{a_{13}a_{22}a_{31}}_{\varepsilon_{321}}\,. \tag{A.6}$$

Some useful properties include

- $\det(\boldsymbol{AB}) = (\det \boldsymbol{A})(\det \boldsymbol{B})$. A corollary is that $\det \boldsymbol{A}^{-1} = \frac{1}{\det \boldsymbol{A}}$.
- $\det$ is invariant under coordinate transformation. $\boldsymbol{A} = \boldsymbol{RA'R}^{-1} \implies \det \boldsymbol{A} = \det \boldsymbol{A'}$.
- If $\boldsymbol{A}$ has n eigenvalues λ_i, $\det \boldsymbol{A} = \prod_{i=1}^{n} \lambda_i$. (Let $\boldsymbol{A'} = \boldsymbol{D}$, the diagonal matrix of λ_i.)
- $\det \boldsymbol{A}$ is the oriented n-dimensional volume defined by the vector product of the n eigenvectors (assumed nondegenerate).

6. *Tensor product*: The tensor product of two n-element vectors is an $n \times n$ matrix:

$$\boldsymbol{x} \otimes \boldsymbol{y} \equiv \boldsymbol{x}\boldsymbol{y}^{\mathsf{T}} = x_i y_j = \begin{pmatrix} x_1 y_1 & \cdots & x_1 y_n \\ \vdots & \ddots & \vdots \\ x_n y_1 & \cdots & x_n y_n \end{pmatrix} \tag{A.7}$$

7. *Scalar product*:

$$\boldsymbol{x} \cdot \boldsymbol{y} \equiv \boldsymbol{x}^{\mathsf{T}}\boldsymbol{y} = \boldsymbol{y}^{\mathsf{T}}\boldsymbol{x} = \operatorname{Tr} \boldsymbol{x}\boldsymbol{y}^{\mathsf{T}} = \sum_{i=1}^{n} x_i y_i \,. \tag{A.8}$$

8. *Inverse*:

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \mathbb{I}, \tag{A.9}$$

where $\mathbb{I}$ s the *identity matrix* (zeros, except for ones along the diagonal). The inverse of an $n \times n$ matrix is another $n \times n$ matrix. If $\boldsymbol{A}$ is symmetric, then so is $\boldsymbol{A}^{-1}$ (Problem A.1.14). Also, $\det \boldsymbol{A}^{-1} = (\det \boldsymbol{A})^{-1}$.

9. *Norm*: The norm of a vector $\boldsymbol{x}$ that is a member of a vector space V is a measure of its size. Abstractly, the norm is a function $\|\boldsymbol{x}\| : V \to \mathcal{R}$ that satisfies

- $\|\boldsymbol{x}\| = 0 \implies \boldsymbol{x} = \boldsymbol{0}$ (positive definiteness);
- $\|a\boldsymbol{x}\| = |a| \, \|\boldsymbol{x}\|$ (homogeneity);
- $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ (triangle inequality).

As a consequence, $\|\boldsymbol{x}\| \geq 0$ (positivity). If the vector space is over the field of complex numbers, then $|a|$ refers to the magnitude of the complex amplitude.

There are many possible functions that serve as norms. Popular ones include

$$\underbrace{\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}}_{\text{Euclidean norm}} , \qquad \underbrace{\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|}_{\text{absolute-value norm}} , \qquad \underbrace{\|\boldsymbol{x}\|_\infty = \sup_{1 \leq i \leq n} |x_i|}_{\text{sup norm}} . \tag{A.10}$$

The sup norm is the least-upper bound (maximum) of all the elements. All of these are particular cases of the p-norm

$$\|\boldsymbol{x}\|_p \equiv \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \tag{A.11}$$

with the sup norm corresponding to $p \to \infty$. We can generalize the notion of norm to define the size of matrices (Section A.1.7) and functions (Section A.3).

10. *Quadratic forms*. For the symmetric matrix $\boldsymbol{A}$, we define the quadratic form $\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x}$. A quadratic form is *positive definite* if $\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x} > 0$ for all $\|\boldsymbol{x}\| > 0$. The form is positive definite if and only if the eigenvalues λ_i of $\boldsymbol{A}$ are positive: $\lambda_i > 0$. The quadratic form is *positive semidefinite* if $\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x} \geq 0$ for all $\|\boldsymbol{x}\| \geq 0$, which is equivalent to the statement that the eigenvalues of $\boldsymbol{A}$ satisfy $\lambda_i \geq 0$.

11. *Real, symmetric, positive-definite matrices*. The study of quadratic forms lead us to consider the properties of real, symmetric, positive-definite $n \times n$ matrices $\boldsymbol{A}$. We quote several properties (proofs left as an exercise): The eigenvalues are real and positive, and the eigenvectors can be chosen to be real and to form an orthonormal basis in $\mathcal{R}^n$. As a result, we can decompose $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\mathsf{T}}$, where $\boldsymbol{D}$ is the diagonal matrix of eigenvalues and $\boldsymbol{U}$ is a *unitary transformation* made from the eigenvectors and satisfies $\boldsymbol{U}^{\mathsf{T}} = \boldsymbol{U}^{-1}$.

A.1.2 Matrix Rank

The notion of *rank* helps to generalize the idea that a square matrix must have nonzero determinant to be invertible. For a rectangular matrix, the rank is the size of the largest nonzero subdeterminant. It equals the dimension of the largest invertible subspace and also the number of nonzero singular values of the matrix (see Section A.1.7). For a rectangular $m \times n$ matrix, the maximum value of the rank = $\min(m, n)$, giving a *full rank* matrix.

Example A.1 The 3×4 matrix $\boldsymbol{M}$

$$\boldsymbol{M} = \begin{pmatrix} 1 & 3 & 1 & 2 \\ 0 & 4 & 1 & 0 \\ 2 & 2 & 1 & 4 \end{pmatrix} \tag{A.12}$$

could have a rank as large as 3. However, all the 3×3 subdeterminants are zero. For example,

$$\det \begin{pmatrix} 1 & 3 & 1 \\ 0 & 4 & 1 \\ 2 & 2 & 1 \end{pmatrix} = 1 \cdot (4-2) - 3 \cdot (0-2) + 1 \cdot (0-8) = 2 + 6 - 8 = 0\,. \tag{A.13}$$

But there are 2×2 matrices with nonzero determinant, implying that rank($\boldsymbol{M}$) = 2:

$$\det \begin{pmatrix} 4 & 1 \\ 2 & 1 \end{pmatrix} = 4 - 2 = 2 \neq 0\,. \tag{A.14}$$

Many analysis programs have a command to compute the rank of a given matrix.

A.1.3 Matrix Inversion

There are a number of helpful formulae for inverting matrices. In the discussion of recursive least squares (Problem 10.10), we use the *Sherman–Morrison* formula:

$$\left(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}\right)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}}{1 + \boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{u}}\,, \tag{A.15}$$

where $\boldsymbol{A}$ is invertible and $\boldsymbol{u}$, $\boldsymbol{v}$ are vectors. See Problem A.1.15, below.

A.1.4 Matrix Exponential

In solving linear equations, we often use the matrix exponential, defined as

$$\mathrm{e}^{\boldsymbol{A}t} \equiv \left(\mathbb{I} + t\,\boldsymbol{A} + \frac{t^2}{2}\boldsymbol{A}^2 + \cdots \right). \tag{A.16}$$

Note that $\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{e}^{\boldsymbol{A}t} = \boldsymbol{A}\,\mathrm{e}^{\boldsymbol{A}t} = \mathrm{e}^{\boldsymbol{A}t}\,\boldsymbol{A}$.

To compute $\mathrm{e}^{\boldsymbol{A}t}$, we change coordinates by the matrix $\boldsymbol{R}$ (so that $\boldsymbol{x} \to \boldsymbol{R}\boldsymbol{x}$) to diagonalize the matrix: $\boldsymbol{A} = \boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^{-1}$, where $\boldsymbol{D}$ is diagonal with entries equal to the eigenvalues. (We ignore potential complications due to repeated eigenvalues). Then

$$\boldsymbol{A}^n = \left(\boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^{-1}\right)^n = \underbrace{\boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^{-1} \cdot \boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^{-1} \cdots}_{n \text{ times}} = \boldsymbol{R}\,\boldsymbol{D}^n\,\boldsymbol{R}^{-1} \tag{A.17}$$

and

$$\begin{aligned} \mathrm{e}^{\boldsymbol{A}t} = \mathrm{e}^{\boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^{-1}t} &= \boldsymbol{R}\boldsymbol{R}^{-1} + t\boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^{-1} + \frac{t^2}{2!}(\boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^{-1})^2 + \cdots + \frac{t^n}{n!}(\boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^{-1})^n + \cdots \\ &= \boldsymbol{R}\left(\mathbb{I} + t\,\boldsymbol{D} + \frac{t^2}{2!}\boldsymbol{D}^2 + \cdots + \frac{t^n}{n!}\boldsymbol{D}^n + \cdots \right)\boldsymbol{R}^{-1} = \boldsymbol{R}\left(\mathrm{e}^{\boldsymbol{D}t}\right)\boldsymbol{R}^{-1}. \end{aligned} \tag{A.18}$$

The exponential of the diagonal matrix is just

$$\boldsymbol{D} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix} \quad \Longrightarrow \quad \mathrm{e}^{\boldsymbol{D}t} = \begin{pmatrix} \mathrm{e}^{\lambda_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathrm{e}^{\lambda_n t} \end{pmatrix}. \tag{A.19}$$

Further properties of the matrix exponential and of the analogously defined matrix logarithm are given in the problems at the end of the section.

A.1.5 Matrix Square Root (Cholesky Decomposition)

If a square, symmetric, positive-definite matrix $\boldsymbol{A}$ can be written as $\boldsymbol{A} = \boldsymbol{B}\,\boldsymbol{B}^{\mathsf{T}}$, then $\boldsymbol{B}$ is a kind of "matrix square root" of $\boldsymbol{A}$. The transpose in the second $\boldsymbol{B}$ factor is motivated by the symmetry of $\boldsymbol{A}$, as $\boldsymbol{A}^{\mathsf{T}} = (\boldsymbol{B}^{\mathsf{T}})^{\mathsf{T}}\,\boldsymbol{B}^{\mathsf{T}} = \boldsymbol{A}$.

But just as a real number has two square roots ($4 = 2 \times 2$ and -2×-2), so, too, the matrix $\boldsymbol{B}$ is not unique. One way to specify a unique $\boldsymbol{B}$ is to demand that it be *lower triangular* – all elements above the diagonal are zero. Then $\boldsymbol{B}^{\mathsf{T}}$ is *upper triangular*, with all elements below the diagonal equal to zero. $\boldsymbol{B}$ and $\boldsymbol{B}^{\mathsf{T}}$ have the same diagonal elements. This choice of $\boldsymbol{B}$ is known as *Cholesky decomposition* (Press et al., 2007, Section 2.9). Lower (or upper) triangular matrices have many attractive features. Their determinant is just the product of the diagonal elements, and the inverse is correspondingly easy to compute, as well.

Example A.2 The Cholesky decomposition of $\boldsymbol{A} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ is

$$\underbrace{\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}}_{\boldsymbol{A}} = \underbrace{\begin{pmatrix} \sqrt{2} & 0 \\ \frac{1}{\sqrt{2}} & \sqrt{\frac{3}{2}} \end{pmatrix}}_{\boldsymbol{B}} \underbrace{\begin{pmatrix} \sqrt{2} & \frac{1}{\sqrt{2}} \\ 0 & \sqrt{\frac{3}{2}} \end{pmatrix}}_{\boldsymbol{B}^{\mathsf{T}}} . \tag{A.20}$$

Note that the diagonal elements of $\boldsymbol{B}$ are *not* equal to the square root of the eigenvalues of $\boldsymbol{A}$, which are (3,1). Notice, too, that $\det \boldsymbol{B} = \det \boldsymbol{B}^{\mathsf{T}} = \sqrt{\det A} = \sqrt{3}$.

The Cholesky decomposition is used to generate correlated random variables from independent random variables, essentially by rotating coordinates. It also plays a role in the Kalman filter and its modifications (UKF, etc.) presented in Chapter 8. Finally, you can use it to solve linear equations $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ when $\boldsymbol{A}$ is symmetric and positive definite.

A.1.6 Cayley–Hamilton Theorem

The Cayley–Hamilton (C-H) theorem states that a matrix satisfies its own *characteristic equation*, the polynomial equation that determines the eigenvalues.

Example A.3

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad \Longrightarrow \quad \begin{vmatrix} \lambda & -1 \\ 1 & \lambda \end{vmatrix} = \lambda^2 + 1 = 0 \tag{A.21}$$

has a characteristic equation $\lambda^2 + 1 = 0$. C-H claims that $\boldsymbol{A}$ also satisfies the equation:

$$\boldsymbol{A}^2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -\mathbb{I} \quad \Longrightarrow \quad \boldsymbol{A}^2 + \mathbb{I} = \boldsymbol{0} . \tag{A.22}$$

We can prove C-H easily if we assume that $\boldsymbol{A}$ can be diagonalized as $\boldsymbol{R}\,\boldsymbol{D}\,\boldsymbol{R}^{-1}$, with $\boldsymbol{D}$ the diagonal matrix of eigenvalues. Then we write the characteristic equation as

$$\begin{aligned} \boldsymbol{A}^n + a_1 \boldsymbol{A}^{n-1} + a_2 \boldsymbol{A}^{n-2} + \cdots + a_n \mathbb{I} = 0 \\ \boldsymbol{R} \left(\boldsymbol{D}^n + a_1 \boldsymbol{D}^{n-1} + a_2 \boldsymbol{D}^{n-2} + \cdots + a_n \mathbb{I} \right) \boldsymbol{R}^{-1} = 0 . \end{aligned} \tag{A.23}$$

But,

$$\boldsymbol{D}^n + a_1 \boldsymbol{D}^{n-1} + a_2 \boldsymbol{D}^{n-2} + \cdots + a_n \mathbb{I} = \boldsymbol{0} \tag{A.24}$$

is just the characteristic equation, copied n times. There are some useful consequences:

- The highest "independent" power of $\boldsymbol{A}$ is $\boldsymbol{A}^{n-1}$. For example, Cayley–Hamilton states that $\boldsymbol{A}^n = -a_1 \boldsymbol{A}^{n-1} - a_2 \boldsymbol{A}^{n-2} - \cdots - a_n \mathbb{I}$. We can reexpress $\boldsymbol{A}^n$ as a linear combination of $\{\mathbb{I}, \boldsymbol{A}, \boldsymbol{A}^2, \cdots, \boldsymbol{A}^{n-1}\}$. Then we do the same for $\boldsymbol{A}^{n+1} = \boldsymbol{A}\boldsymbol{A}^n$ and for higher powers.

- The matrix exponential $e^{\boldsymbol{A}t} = \mathbb{I} + t\boldsymbol{A} + \frac{1}{2}t^2\boldsymbol{A}^2 + \cdots$ can therefore also be written in terms of $\{\mathbb{I}, \boldsymbol{A}, \boldsymbol{A}^2, \cdots, \boldsymbol{A}^{n-1}\}$:

$$e^{\boldsymbol{A}t} = \sum_{j=0}^{\infty} \frac{t^j}{j!}\,\boldsymbol{A}^j = \sum_{j=0}^{n-1} \alpha_j(t)\,\boldsymbol{A}^j\,. \tag{A.25}$$

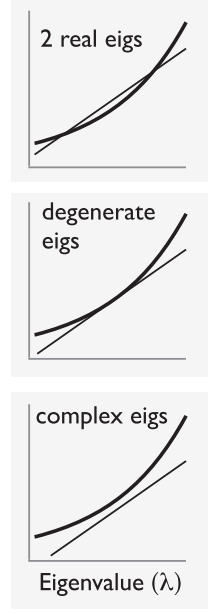
  Note that the functions $\alpha_j(t)$ are not proportional to t^j. In expanding the matrix exponential, it is the matrices $\boldsymbol{A}^j$ that are rewritten in terms of lower powers, not the prefactors.

- The Cayley–Hamilton theorem gives a way to evaluate a matrix exponential that is useful for small matrices. From Eq. (A.24), the finite expansion, Eq. (A.25), is satisfied by *each* eigenvalue λ, giving n linear equations for the n functions $\alpha_j(t)$. Here, we illustrate the $n = 2$ case. Higher orders are perhaps better evaluated numerically or by a computer-algebra program. For $n = 2$ and eigenvalues, $\lambda_\pm$,

$$e^{\boldsymbol{A}t} = \alpha_0(t)\mathbb{I} + \alpha_1(t)\boldsymbol{A} \quad \Longrightarrow \quad e^{\lambda_\pm t} = \alpha_0(t) + \alpha_1(t)\lambda_\pm\,, \tag{A.26}$$

  If the eigenvalues are distinct, there are two independent equations for $\alpha_0(t)$ and $\alpha_1(t)$.

- If the eigenvalues are degenerate, we find the second equation by differentiating both sides of the equation with respect to λ. To see why, consider the three graphs at left, which plot the left- and right-hand sides of Eq. (A.26) against λ. The heavy curve is the exponential, the lighter line $\alpha_0 + \alpha_1\lambda$. At top, there are two real, distinct eigenvalues; at bottom, two complex-conjugate eigenvalues. The middle case corresponds to the degenerate case, where Eq. (A.26) and its derivative with respect to λ coincide. The second, independent equation is then $\alpha_1 = t\,e^{\lambda t}$.



Example A.4

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad \Longrightarrow \quad e^{\boldsymbol{A}t} = \alpha_0(t)\,\mathbb{I} + \alpha_1(t)\,\boldsymbol{A}\,. \tag{A.27}$$

This expansion for $\boldsymbol{A}$ is also satisfied by each of its eigenvalues, which here are $\pm i$. Thus,

$$e^{it} = \alpha_0 + i\alpha_1 \tag{A.28a}$$

$$e^{-it} = \alpha_0 - i\alpha_1\,. \tag{A.28b}$$

Adding and subtracting the equations gives $\alpha_0(t) = \cos t$ and $\alpha_1(t) = \sin t$, so that

$$e^{\boldsymbol{A}t} = \cos t\,\mathbb{I} + \sin t\,\boldsymbol{A} = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}. \tag{A.29}$$

A.1.7 Singular Value Decomposition

The singular value decomposition (SVD) is introduced in Section 3.8.3, in the discussion of the frequency response of MIMO systems. We recap some of its properties here:

- *Principal gains*: $\overline{\sigma}(\boldsymbol{A})\,\|x\|_2 \geq \|\boldsymbol{A}\boldsymbol{x}\|_2 \geq \underline{\sigma}(\boldsymbol{A})\,\|\boldsymbol{x}\|_2\,,\ \forall \boldsymbol{x}$. That is, the operation of $\boldsymbol{A}$ on any vector $\boldsymbol{x}$ changes its length by a factor that ranges from $\underline{\sigma}(\boldsymbol{A})$ up to $\overline{\sigma}(\boldsymbol{A})$.
- *Induced 2-norm*: $\overline{\sigma}(\boldsymbol{A})\,\|\boldsymbol{A}\|_2 = \sup_{\boldsymbol{x}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}$.
- If $\boldsymbol{A}^{-1}$ exists, $\overline{\sigma}(\boldsymbol{A}) = 1/\underline{\sigma}(\boldsymbol{A}^{-1})$.
- *Condition number*: $\kappa = \overline{\sigma}(\boldsymbol{A})/\underline{\sigma}(\boldsymbol{A})$. It reduces to $|\lambda|_{\max}/|\lambda|_{\min}$ when A is normal.

  The condition number plays an important role in numerical analysis.

Example A.5 (Condition number and numerical precision) Consider

$$\boldsymbol{A}\boldsymbol{x}_0 = \boldsymbol{y}_0 \quad \Longrightarrow \quad \boldsymbol{y}_0 = \boldsymbol{A}^{-1}\boldsymbol{x}_0\,. \tag{A.30}$$

Assume that the matrix $\boldsymbol{A}$ is known but that there is some uncertainty in the "data" $\boldsymbol{y}_0$, so that we actually solve for $\boldsymbol{y} = \boldsymbol{y}_0 + \delta\boldsymbol{y}$. Then, expanding $\boldsymbol{A}(\boldsymbol{x}_0 + \delta\boldsymbol{x}) = (\boldsymbol{y}_0 + \delta\boldsymbol{y})$ gives

$$\delta\boldsymbol{x} = \boldsymbol{A}^{-1}\delta\boldsymbol{y} \quad \Longrightarrow \quad \|\delta\boldsymbol{x}\| \leq \|\boldsymbol{A}^{-1}\|\,\|\delta\boldsymbol{y}\| = \|\delta\boldsymbol{y}\|/\underline{\sigma}(\boldsymbol{A})\,, \tag{A.31}$$

using the Cauchy-Schwartz inequality and $\|\boldsymbol{A}^{-1}\| = \overline{\sigma}(\boldsymbol{A}^{-1}) = 1/\underline{\sigma}(\boldsymbol{A})$. Also, we have, from $\boldsymbol{A}\boldsymbol{x}_0 = \boldsymbol{y}_0$, that $\overline{\sigma}(\boldsymbol{A})\,\|\boldsymbol{x}_0\| \geq \|\boldsymbol{y}_0\|$. Putting everything together gives

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}_0\|} \leq \left(\frac{\|\delta\boldsymbol{y}\|}{\underline{\sigma}(\boldsymbol{A})}\right)\left(\frac{\overline{\sigma}(\boldsymbol{A})}{\|\boldsymbol{y}_0\|}\right) = \kappa\,\frac{\|\delta\boldsymbol{y}\|}{\|\boldsymbol{y}_0\|}\,. \tag{A.32}$$

Thus, the fractional uncertainty in the solution is κ times the fractional uncertainty in the data. Matrices $\boldsymbol{A}$ with large condition number can magnify the uncertainties in data so much that the solution becomes unusable. Thus, numerical methods place a premium on formulating linear equations with reasonable condition numbers. In particular, unitary matrices have $\kappa = 1$, which motivates their use in many algorithms.

A.1.8 Complex Numbers and Linear Algebra

While most of the book uses linear algebra over real numbers, we sometimes consider complex vectors and matrices. Here are a few extensions that are needed for that case.

1 *Scalar product*: $\boldsymbol{A} \cdot \boldsymbol{B} = \sum_i A_i\,B_i^*$.
2 *Hermitian transpose*: $\boldsymbol{A}^\dagger_{ij} = \boldsymbol{A}^*_{ji}$.
3 *Normal matrices*: $\boldsymbol{A}^\dagger\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^\dagger$.
4 *Unitary matrices*: $\boldsymbol{U}^{-1} = \boldsymbol{U}^\dagger$, implying that $\boldsymbol{U}\boldsymbol{U}^\dagger = \mathbb{I}$.
5 *Singular Value Decomposition (SVD)*: $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\dagger$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary. When $\boldsymbol{A}$ is an $n \times n$ matrix, $\boldsymbol{\Sigma}$ is an $n \times n$ diagonal matrix of singular values.

Polar Decomposition

In our discussion of quantum control in Chapter 13, we will need to use the polar decomposition of a matrix. The decomposition generalizes the polar form of a complex number:

$$z = x + \mathrm{i}y = r\,\mathrm{e}^{\mathrm{i}\theta}, \qquad r = |z|\,; \theta = \tan^{-1}\left(\tfrac{y}{x}\right). \tag{A.33}$$

The analogous relation for a matrix $\boldsymbol{A}$ with complex elements is

$$\boldsymbol{A} = \boldsymbol{U}_0\boldsymbol{P}_0\,, \tag{A.34}$$

where $\boldsymbol{U}$ is unitary and $\boldsymbol{P}$ is positive semidefinite.

We can use the singular-value decomposition to find this decomposition explicitly:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\dagger = \underbrace{\boldsymbol{U}\boldsymbol{V}^\dagger}_{\boldsymbol{U}_0}\,\underbrace{\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\dagger}_{\boldsymbol{P}_0} \tag{A.35}$$

We then see immediately that $\boldsymbol{U}_0$ is unitary. Since the elements of $\boldsymbol{\Sigma}$ are nonnegative, so must be the matrix $\boldsymbol{P}_0$. Indeed, since the singular values are also the square root of the eigenvalues of $\boldsymbol{A}\boldsymbol{A}^\dagger$, we can see that they play the role of r in a polar decomposition. The values $\sigma_i = \sqrt{\lambda\lambda^*} = r_i$. The decomposition in Eq. (A.35) is unique, but there is a second decomposition of the form

$$\boldsymbol{A} = \underbrace{\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\dagger}_{\boldsymbol{P}_1}\,\underbrace{\boldsymbol{U}\boldsymbol{V}^\dagger}_{\boldsymbol{U}_1}\,. \tag{A.36}$$

Alternatively, the polar-decomposition theorem implies the singular-value-decomposition theorem for the case of a square matrix:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{P} = \boldsymbol{U}\,\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^{\mathsf{T}} = (\boldsymbol{U}\boldsymbol{R})\,\boldsymbol{\Sigma}\,(\boldsymbol{R}^{\mathsf{T}}) \tag{A.37}$$

In general, singular-value decomposition works for rectangular as well as square matrices.

Example A.6 Consider the complex matrix $\boldsymbol{A} = \left(\begin{smallmatrix} 1 & 2\mathrm{i} \\ -2\mathrm{i} & 1 \end{smallmatrix}\right)$. It has eigenvalues 3, −1, but its singular value decomposition is

$$\boldsymbol{A} = \underbrace{\frac{1}{\sqrt{2}}\begin{pmatrix} \mathrm{i} & \mathrm{i} \\ 1 & -1 \end{pmatrix}}_{\boldsymbol{U}}\underbrace{\begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}}_{\boldsymbol{\Sigma}}\underbrace{\frac{1}{\sqrt{2}}\begin{pmatrix} -\mathrm{i} & 1 \\ \mathrm{i} & 1 \end{pmatrix}}_{\boldsymbol{V}^\dagger} = \underbrace{\begin{pmatrix} 0 & \mathrm{i} \\ -\mathrm{i} & 1 \end{pmatrix}}_{\boldsymbol{U}_0}\underbrace{\begin{pmatrix} 2 & \mathrm{i} \\ -\mathrm{i} & 2 \end{pmatrix}}_{\boldsymbol{P}_0}. \tag{A.38}$$

Verify that $\boldsymbol{U}_0\boldsymbol{U}_0^\dagger = \mathbb{I}$ and that $\boldsymbol{P}_0$ has eigenvalues of 3 and 1, making it positive definite.

A.1.9 Matrix Calculus

We review expressions from multivariable calculus that include vectors and matrices and derive a few identities that are used in the main text. In general, it is better to do

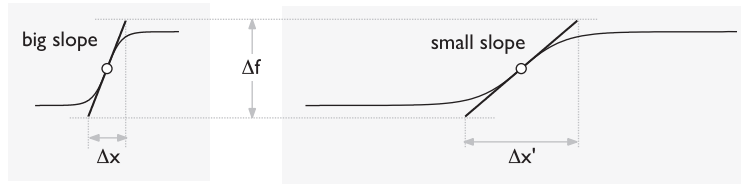


Fig. A.1 Scaling increases Δx but decreases the gradient $\frac{\partial}{\partial x}$ and keeps Δf constant.

such calculations using tensors (or differential forms), particularly if third- and higher-order tensors are needed. For this book, the matrix-calculus notation is simpler and suffices.

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be n-element column vectors, with $\boldsymbol{x}^\mathsf{T}$ and $\boldsymbol{y}^\mathsf{T}$ the corresponding row vectors. Then, if we represent a position $\boldsymbol{x}$ by a column vector, we should represent the *gradient* of a scalar $\boldsymbol{\nabla} = \frac{\partial}{\partial \boldsymbol{x}}$ by a row vector, $(\frac{\partial}{\partial x_1}, \cdots, \frac{\partial}{\partial x_n})$. Figure A.1 justifies this choice by showing the rate of change in a scalar function $f(x)$ in two coordinate systems, x and $x' = \lambda x$, where λ is a scale factor. The guiding principle is that the change in the physical quantity f due to a displacement should be independent of the units we use to describe that displacement. Thus, $\Delta f = \frac{\partial f}{\partial x}\Delta x = \frac{\partial f}{\partial x'}\Delta x'$. Since $\Delta x' = \lambda \Delta x$, we must have $\frac{\partial}{\partial x'} = \frac{1}{\lambda}\frac{\partial}{\partial x}$. Generalizing to a scalar function of a vector $\boldsymbol{x}$, the same logic gives $\boldsymbol{x}' = \boldsymbol{\Lambda x}$ implies $\boldsymbol{\nabla}' = \boldsymbol{\Lambda}^{-1}\boldsymbol{\nabla}$, where $\boldsymbol{\Lambda}$ is a matrix that transforms the $\boldsymbol{x}$ coordinates into $\boldsymbol{x}'$ coordinates. In the fancier language of tensors, $\boldsymbol{x}$ is a *contravariant* vector, whereas $\frac{\partial}{\partial \boldsymbol{x}}$ is a *covariant* vector. Here, they are just column and row vectors, respectively. For the deeper story involving *calculus on manifolds*, see Stone and Goldbart (2009).

Example A.7 Let $f = \boldsymbol{x}^\mathsf{T}\boldsymbol{A}\boldsymbol{x}$, with $\boldsymbol{x}$ an n-dimensional vector and $\boldsymbol{A}$ an $n \times n$ matrix.

$$\begin{aligned}\text{Then}\quad \frac{\partial f}{\partial \boldsymbol{x}} &= \left(\tfrac{\partial}{\partial x_1}, \cdots, \tfrac{\partial}{\partial x_n}\right)\left(A_{11}x_1^2 + A_{12}x_1x_2 + A_{21}x_2x_1 + \cdots\right) \\ &= [2A_{11}x_1 + (A_{12}+A_{21})x_2 + \cdots, (A_{12}+A_{21})x_1 + 2A_{22}x_2 + \cdots,] \\ &= \boldsymbol{x}^\mathsf{T}(\boldsymbol{A} + \boldsymbol{A}^\mathsf{T}). \end{aligned} \tag{A.39}$$

Problem A.1.1 Show that if $\boldsymbol{A}$ is symmetric, then so is $\boldsymbol{A}^{-1}$.

Problem A.1.2 Verify the Sherman–Morrison formula, Eq. (A.15).

Problem A.1.3 Show $\frac{\partial^2}{\partial \boldsymbol{x}^\mathsf{T}\partial \boldsymbol{x}}\left(\boldsymbol{x}^\mathsf{T}\boldsymbol{A}\boldsymbol{x}\right) = \boldsymbol{A} + \boldsymbol{A}^\mathsf{T}$; $\frac{\partial}{\partial \boldsymbol{x}}\mathrm{Tr}\left(\boldsymbol{x}\boldsymbol{x}^\mathsf{T}\right) = 2\boldsymbol{x}^\mathsf{T}$, and $\frac{\partial}{\partial \boldsymbol{x}}\left(\boldsymbol{x}^\mathsf{T}\boldsymbol{y}\right) = \boldsymbol{y}^\mathsf{T}$.

Problem A.1.4 Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be $n \times m$ matrices. Show $\partial_{\boldsymbol{A}}\left(\mathrm{Tr}\,\boldsymbol{A}\boldsymbol{B}^\mathsf{T}\right) = \partial_{\boldsymbol{A}}\left(\mathrm{Tr}\,\boldsymbol{B}\boldsymbol{A}^\mathsf{T}\right) = \boldsymbol{B}^\mathsf{T}$. Hint: Make sure your definition of derivative with respect to a matrix is consistent with the previously defined limiting case $m = 1$ for a vector.

Problem A.1.5 Let $\boldsymbol{A}$ be an $n \times m$ matrix and let $\boldsymbol{B}$ be an $m \times m$ matrix. Show that $\partial_{\boldsymbol{A}}\left(\mathrm{Tr}\,\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^\mathsf{T}\right) = \left(\boldsymbol{B} + \boldsymbol{B}^\mathsf{T}\right)\boldsymbol{A}^\mathsf{T}$.

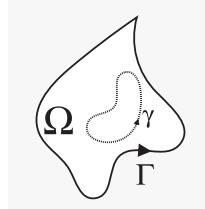
Problem A.1.6 Show that if $\boldsymbol{AB} = \boldsymbol{BA}$, then $\mathrm{e}^{\boldsymbol{A}+\boldsymbol{B}} = \mathrm{e}^{\boldsymbol{A}}\,\mathrm{e}^{\boldsymbol{B}}$. If you are clever, no calculations are required! This identity does *not* hold when $\boldsymbol{A}$ and $\boldsymbol{B}$ do not commute.

Problem A.1.7 In analogy with the matrix exponential, we can define a matrix logarithm via the identity $\ln(\mathbb{I} + \boldsymbol{A}) = \boldsymbol{A} - \frac{1}{2}\boldsymbol{A}^2 + \frac{1}{3}\boldsymbol{A}^3 - \cdots$. Use the previous problem to show that if $\boldsymbol{AB} = \boldsymbol{BA}$, then $\log \boldsymbol{AB} = \ln \boldsymbol{A} + \log \boldsymbol{B}$.

Problem A.1.8 Show $\ln \det \boldsymbol{A} = \mathrm{Tr} \ln \boldsymbol{A}$, for symmetric, positive-definite matrices $\boldsymbol{A}$. The identity holds for more general $\boldsymbol{A}$ using the complex logarithm.

A.2 Complex Analysis

In this section, we collect a few results from complex analysis that we draw on in the text. Many of the results are stated in terms of a *domain* in the complex plane. Domains are open sets that are connected (each pair of points can be joined by a polygonal path that consists of a finite number of line segments joined end to end). A *region* is a domain plus its boundary and is *closed* if it includes all its boundary points. By convention, the boundary is oriented counterclockwise. Left, we show a domain Ω, its oriented boundary $\Gamma = \partial\Omega$, and an interior oriented path γ.



1. *Cauchy-Riemann equations*. Let $f = u(x, y) + \mathrm{i}\, v(x, y)$ be a complex-valued function defined in the xy plane. In general, $f = f(x, y)$, but let us consider the case where $f = f(z)$ is a function of the single complex variable $z = x + \mathrm{i}y$. We then implicitly assert that f has no dependence on the complex conjugate $\bar{z} = x - \mathrm{i}y$. In other words,

$$\frac{\partial f}{\partial \bar{z}} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial \bar{z}} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial \bar{z}} = \left(\frac{\partial u}{\partial x} + \mathrm{i}\frac{\partial v}{\partial x}\right)\frac{1}{2} + \left(\frac{\partial u}{\partial y} + \mathrm{i}\frac{\partial v}{\partial y}\right)\frac{\mathrm{i}}{2} = 0\,, \tag{A.40}$$

   which implies, after separating real and imaginary parts into two equations, that

$$\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} = 0\,, \qquad \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} = 0\,, \tag{A.41}$$

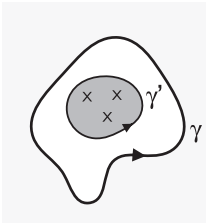
   which are the Cauchy-Riemann equations. If u and v derivatives exist with respect to x and y, then a function $f(z)$ that satisfies the Cauchy-Riemann equations is differentiable in the sense that the limit $[f(z) - f(z_0)]/(z - z_0)$ as $z \to z_0$ exists and is independent of the path used by z to approach z_0. Functions that are differentiable in an open domain Ω are *analytic*. Analytic functions are infinitely differentiable and have a power-series expansion about every point in their domain. Taking further derivatives of the Cauchy-Riemann equations implies $\nabla^2 u = \nabla^2 v = 0$, where $\nabla^2 = \partial_x^2 + \partial_y^2$. That is, u and v are *harmonic* functions of x and y.
2. *Cauchy Integral Theorem*. Let $f(z)$ be analytic in a domain Ω enclosed by the oriented boundary curve $\Gamma = \partial\Omega$. Then

$$\oint_\Gamma \mathrm{d}z\, f(z) = 0\,. \tag{A.42}$$

The proof is based on Green's theorem: Let $z = x + \mathrm{i}y$ and $f = u + \mathrm{i}v$. Then

$$\oint_\Gamma \mathrm{d}z\, f(z) = \oint_\Gamma (\mathrm{d}x + \mathrm{i}\mathrm{d}y)(u + iv) = \oint_\Gamma (\mathrm{d}x\, u - \mathrm{d}y\, v) + \mathrm{i}\oint_\Gamma (\mathrm{d}x\, v + \mathrm{d}y\, u)$$
$$= -\iint_\Omega \mathrm{d}x\,\mathrm{d}y \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right) + \mathrm{i}\iint_\Omega \mathrm{d}x\,\mathrm{d}y \left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}\right), \tag{A.43}$$

which vanishes by the Cauchy-Riemann equations. One application of this theorem is to deform the path of a contour integration without changing the value of the integral. Specifically, if two paths γ and γ' define a region where $f(z)$ is analytic, then $\oint_\gamma = \oint_{\gamma'}$, as illustrated at right. (The × marks denote poles, where $f(z)$ is not analytic.)



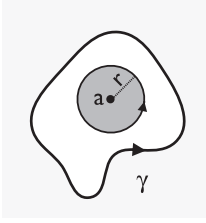
3. *Maximum Modulus Principle*. If $f(z)$ is analytic in Ω, then its modulus $|f(z)|$ does not have a maximum on Ω. More simply, either $f(z)$ is constant or its maximum is on the boundary $\Gamma = \partial\Omega$. In lieu of a proof, consider this physical argument: We have $\log f(z) = \log|f| + \mathrm{i}\arg f(z)$. Since f is analytic, so is $\log f$, implying that $\log|f|$ is harmonic. We interpret $\log|f|$ as a steady-state solution of a heat flow obeying the diffusion equation. But if such a flow has a local maximum, then there would be a heat current away in all directions and the temperature at that point could not be stationary, since there are no sources (whose singularity would make f nonanalytic). Since the log function is monotonic, $|f(z)|$ also cannot have a local maximum.

   We will use this principle to calculate sup norms in Chapter 9 (Robust Control).

4. *Cauchy's Integral Formula*. The value of an analytic function $f(z)$ at $z = a$ can be determined by an integral along a closed path γ that encircles a:

$$f(a) = \frac{1}{2\pi\,\mathrm{i}} \oint_\gamma \mathrm{d}z\, \frac{f(z)}{z - a}\,. \tag{A.44}$$

   To see this, use the Cauchy Integral theorem to contract the contour γ to a small circle of radius r about a, as shown at right. Substitute $z = a + r\,\mathrm{e}^{\mathrm{i}\theta}$ and $dz = \mathrm{i}\,r\,\mathrm{e}^{\mathrm{i}\theta}\,\mathrm{d}\theta$ and take the limit $r \to 0$:



$$\oint_\gamma \mathrm{d}z\, \frac{f(z)}{z - a} = \lim_{r\to 0} \int_0^{2\pi} \mathrm{d}\theta\, \frac{f(a + r\,\mathrm{e}^{\mathrm{i}\theta})}{r\,\mathrm{e}^{\mathrm{i}\theta}}\, \mathrm{i}\,r\,\mathrm{e}^{\mathrm{i}\theta} = f(a)\ \mathrm{i}2\pi\,. \tag{A.45}$$

5. *Residue Theorem*. Let $f(z)$ be a complex function that is analytic in the domain Ω, except at isolated points α_k. Let γ be a closed path within this open set that is traversed counterclockwise. Then,

$$\oint_\gamma \mathrm{d}z\, f(z) = 2\pi\,\mathrm{i} \sum_k \mathrm{Res}_k\,, \tag{A.46}$$

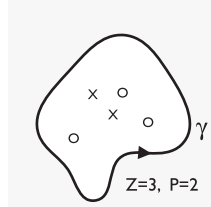
   where Res_k is the residue of $f(z)$ associated with point α_k. The *residue* is defined to be the a_{-1} coefficient of the *Laurent* expansion. If the function f has a simple pole at a point $z = \alpha$ and can be expressed as $f(z) = \frac{g(z)}{h(z)}$,

$$\mathrm{Res}(z) = \lim_{z\to\alpha} (z - \alpha)\, f(z) = \frac{g(\alpha)}{h'(\alpha)}\,, \tag{A.47}$$

which we can see by noting that if f has a simple pole at $z = \alpha$, then $h(z)$ has a simple zero and can be Taylor expanded as $h(z) = \cancelto{0}{h(\alpha)} + (z-\alpha)\,h'(\alpha) + \cdots$.

6. *Principle of the Argument*. Let $f(z)$ be a meromorphic function (i.e., analytic except at isolated points). Then, for a closed contour γ

$$\frac{1}{2\pi\,\mathrm{i}} \oint_\gamma \mathrm{d}z\, \frac{f'(z)}{f(z)} = Z - P\,, \tag{A.48}$$



where Z and P are the number of zeros and poles inside the contour γ. To prove this relation, we follow Stone and Goldbart (2009) and write $f(z) = (z-a)^m\,h(z)$, where $h(z)$ is analytic near a and m can be positive or negative (zero or pole of order m). Then

$$\frac{f'(z)}{f(z)} = \frac{m}{z-a} + \frac{h'(z)}{h(z)}\,, \tag{A.49}$$

which has a simple pole at a with residue m. Summing the contributions from each singularity then gives the result. To interpret Eq. (A.48), we note that

$$\oint_\gamma \mathrm{d}z\, \frac{f'(z)}{f(z)} = \oint_\gamma \mathrm{d}z\, \frac{\mathrm{d}}{\mathrm{d}z}[\ln f(z)] = \Delta_\gamma \ln f = \Delta_\gamma \left[\ln\left(r\,\mathrm{e}^{\mathrm{i}\theta}\right)\right] = \mathrm{i}\,\Delta_\gamma\theta\,, \tag{A.50}$$

where $\Delta_\gamma\theta$ is the change in the argument of $f(z)$ along γ.

7. *Jensen's formula*. Let $f(z)$ be analytic inside and on the unit circle, except for poles at $p_1, p_2, \ldots, p_m$ and zeros at $z_1, z_2, \ldots, z_n$. The poles and zeros are inside the unit circle. Then

$$\int_{-\pi}^{\pi} \frac{\mathrm{d}\theta}{2\pi}\, \ln|f(\mathrm{e}^{\mathrm{i}\theta})| = \ln|f(0)| + \sum_{k=1}^{m} \ln|p_k| - \sum_{k=1}^{n} \ln|z_k|\,. \tag{A.51}$$

To prove this, we first recall that $\mathrm{Re}\,(\log z) = \log|z|$ and write

$$\begin{aligned} \ln|f(\mathrm{e}^{\mathrm{i}\theta})| = \mathrm{Re}\,[\log f(\mathrm{e}^{\mathrm{i}\theta})| &= \mathrm{Re}\,\left\{\log f(0) + \int_0^1 \mathrm{d}r\, \frac{\mathrm{d}}{\mathrm{d}r}\left[\log f(r\,\mathrm{e}^{\mathrm{i}\theta})\right]\right\} \\ &= \log|f(0)| + \mathrm{Re}\int_0^1 \mathrm{d}r\, \frac{f'(r\,\mathrm{e}^{\mathrm{i}\theta})\,\mathrm{e}^{\mathrm{i}\theta}}{f(r\,\mathrm{e}^{\mathrm{i}\theta})}\,. \end{aligned} \tag{A.52}$$

We substitute into our integral and interchange the r and θ integrals:

$$\begin{aligned} \int_{-\pi}^{\pi} \frac{\mathrm{d}\theta}{2\pi}\, \ln|f(\mathrm{e}^{\mathrm{i}\theta})| &= \log|f(0)| + \mathrm{Re}\int_{-\pi}^{\pi} \frac{\mathrm{d}\theta}{2\pi} \int_0^1 \mathrm{d}r\, \frac{f'(r\,\mathrm{e}^{\mathrm{i}\theta})\,\mathrm{e}^{\mathrm{i}\theta}}{f(r\,\mathrm{e}^{\mathrm{i}\theta})} \\ &= \log|f(0)| + \mathrm{Re}\int_0^1 \frac{\mathrm{d}r}{2\pi\,\mathrm{i}\,r} \int_{-\pi}^{\pi} \mathrm{d}\theta\,(\mathrm{i}r\,\mathrm{e}^{\mathrm{i}\theta}) \frac{f'(r\,\mathrm{e}^{\mathrm{i}\theta})}{f(r\,\mathrm{e}^{\mathrm{i}\theta})} \\ &= \log|f(0)| + \mathrm{Re}\int_0^1 \frac{\mathrm{d}r}{2\pi\,\mathrm{i}\,r} \oint_{\gamma_r} \mathrm{d}z\, \frac{f'(z)}{f(z)} \\ &= \log|f(0)| + \int_0^1 \mathrm{d}r\, \frac{Z(r) - P(r)}{r}\,. \end{aligned} \tag{A.53}$$

The last step is based on the Principle of the Argument, Eq. (A.48), with $Z(r)$ the number of zeros within a circle γ_r of radius r and $P(r)$ the corresponding number of poles.

We can do the radial integral explicitly. Since the zeros have magnitude $|z_k|$, we write

$$\begin{aligned}\int_0^1 \mathrm{d}r\,\frac{Z(r)}{r} &= \int_0^{|z_1|}\mathrm{d}r\,\frac{0}{r} + \int_{|z_1|}^{|z_2|}\mathrm{d}r\,\frac{1}{r} + \int_{|z_2|}^{|z_3|}\mathrm{d}r\,\frac{2}{r} + \cdots \int_{|z_n|}^{1}\mathrm{d}r\,\frac{n}{r} \\ &= (\ln|z_2| - \ln|z_1|) + 2(\ln|z_3| - \ln|z_2|) + \cdots n(\ln 1 - \ln|z_n|) \\ &= -\ln|z_1| - \ln|z_2| - \cdots \ln|z_n| = -\sum_{k=1}^{n}\ln|z_k|\,. \end{aligned} \tag{A.54}$$

Repeating the argument for the poles, at magnitude $|p_k|$, leads to Eq. (A.51). Jensen's formula is used to derive the discrete version of Bode's waterbed theorem (Problem 15.6).

A.3 Functional Analysis

Function spaces can generalize the notion of vector space to infinite dimensions. Their indices may be *countable* (one-to-one correspondence with the integers) or *uncountable* (one-to-one correspondence with real numbers). Many of the notions of finite-dimensional spaces have their infinite-dimensional counterparts. The mathematics is considerably more subtle because it is harder to prove completeness. For example, a set of N linearly independent vectors forms a basis for an N-dimensional space. For an infinite-dimensional space, an infinite set of linearly independent functions may or may not form a complete basis. The idea of expanding a function over an infinite set of basis functions leads to notions of Fourier series and transforms, which we explore below in Section A.4.

The notion of a norm, defined for finite-dimensional vector spaces in Section A.1.1, can be extended to functions. The $\mathcal{L}_p$ norm of a function $f(t)$ is

$$\|f\|_p = \left[\int_{-\infty}^{\infty}\mathrm{d}t\,|f(t)|^p\right]^{1/p}. \tag{A.55}$$

The function space $\mathcal{L}_p$ (*Lebesgue space*) is then the set of functions with $\|f\|_p < \infty$. The norm and functions can also be defined over a restricted range (t_0, t_1). The $p = 2$ case gives the square norm. The $p \to \infty$ case picks out the maximum value of $f(t)$ on the interval and is known as the *sup norm* (for *supremum*). Strictly speaking, $\|\cdot\|_p$ is not quite a norm, as the integral of $|f|^p$ may vanish even if f itself is not everywhere 0. Stone and Goldbart (2009) discuss how to finesse this complication. (See their Section 2.2.2.)

For complex, meromorphic functions such as transfer functions, the notion of a *Hardy space* and its associated norm is useful, too. For $p \geq 1$, the $\mathcal{H}_p$ norm of $f(s)$ is

$$\|f\|_p = \left(\int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} |f(\mathrm{i}\omega)|^p \right)^{1/p} . \tag{A.56}$$

The square ($\mathcal{H}_2$) and sup ($\mathcal{H}_\infty$) norms play a role in the theory of robust control (Chapter 9).

A.4 Laplace and Fourier Transforms

Because linear systems obey the law of superposition of solutions, expanding a general solution as an infinite sum of simple solutions is a useful technique. This motivates the study of various transform methods. Here, we review some basic properties of the Fourier series and transform. Then we introduce the Laplace transform, which we use extensively, as it is the predominant transform tool for control systems.

A.4.1 Fourier Series

Expansion in Sines and Cosines

A *periodic* function satisfies $f(t + T) = f(t)$. Under fairly weak constraints on $f(t)$, we can express a periodic function as a sum of sines and cosines. Explicitly,

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos n\omega t + b_n \sin n\omega t) , \tag{A.57}$$

where $\omega = 2\pi/T$ and $f(t)$ need only be square integrable over the period T. In particular, $f(t)$ can be C^0, with jump discontinuities. The coefficients $\{a_n\}$ and $\{b_n\}$ are given by

$$a_n = \frac{2}{T} \int_{-T/2}^{T/2} \mathrm{d}t\, f(t) \cos n\omega t , \quad b_n = \frac{2}{T} \int_{-T/2}^{T/2} \mathrm{d}t\, f(t) \sin n\omega t , \tag{A.58}$$

You can verify Eqs. (A.58) by directly substituting the relation for $f(t)$ in Eq. (A.57). The key idea is that sines and cosines of different frequencies that are rationally related are *orthogonal* when integrated over their common period:

$$\frac{2}{T} \int_{-T/2}^{T/2} \mathrm{d}t \cos \ell\omega t \, \cos n\omega t = \frac{2}{T} \int_{-T/2}^{T/2} \mathrm{d}t \sin \ell\omega t \, \sin n\omega t = \delta_{\ell n} ,$$
$$\int_{-T/2}^{T/2} \mathrm{d}t \cos \ell\omega t \, \sin n\omega t = 0 . \tag{A.59}$$

Using Eqs. (A.58), you can show, for example (Problem A.4.1), that a periodic square wave ⊓⊔ that is symmetric about $t = 0$ can be expanded as

$$\mathrm{sq}(t) = \frac{1}{2} + \frac{2}{\pi} \left(\cos \omega t - \frac{1}{3} \cos 3\omega t + \frac{1}{5} \cos 5\omega t - \ldots \right) , \tag{A.60}$$

where sq(t) is defined to be an even function of period T, satisfying $f(t) = -f(-t)$. Because the function sq(t) is symmetric about $t = 0$, there are only cosine terms. The terms decrease as i/n (that is, $a_n \sim \mathrm{i}/n$). This turns out to result from the jump discontinuities. If you integrate a square wave, you get a triangle wave, ⩘. If you integrate the terms of the square wave expansion, the new terms of the triangle expansion go as $(\mathrm{i}/n)^2$. Thus, the Fourier series of a triangle wave converges more quickly than that of a square wave. Note that a triangle wave is a C^1 function, with a discontinuity in its first derivative. Continuing the argument, a C^k function (discontinuity in the nth derivative) has coefficients that are $\sim (1/n)^k$ – the smoother the function, the faster its Fourier series converges. Finally, note the absence of even-order terms in Eq. (A.58). This is not usual and traces back to the 50% "duty cycle" of the square wave (Problem A.4.1). Any other duty cycle – ratio of the interval where it is 1 to where it is 0 – would have both odd and even coefficients.

Expansion in Complex Exponentials

In Eq. (A.57), a function $f(t)$ is expanded in terms of sines and cosines. Recalling that

$$\cos\,\omega t = \tfrac{1}{2}\left(\mathrm{e}^{\mathrm{i}\omega t} + \mathrm{e}^{-\mathrm{i}\omega t}\right) \qquad \sin\,\omega t = \tfrac{1}{2\mathrm{i}}\left(\mathrm{e}^{\mathrm{i}\omega t} - \mathrm{e}^{-\mathrm{i}\omega t}\right), \tag{A.61}$$

we can rewrite a Fourier series in terms of complex exponential functions:

$$f(t) = \sum_{n=-\infty}^{\infty} c_n\, \mathrm{e}^{+\mathrm{i}n\omega t}, \tag{A.62}$$

with Fourier coefficients

$$c_n = \frac{1}{T}\int_{-T/2}^{T/2} \mathrm{d}t\, f(t)\, \mathrm{e}^{-\mathrm{i}n\omega t}\,. \tag{A.63}$$

Note that the coefficients can be simplified in special cases:

$$c_n = \begin{cases} c_{-n} & \text{if } f(t) \text{ is even}, \\ -c_{-n} & \text{odd}, \\ c_{-n}^* & \text{real}. \end{cases} \tag{A.64}$$

Connection with Linear Algebra

As noted above, the key step in deriving the coefficients of a Fourier series expansion is to use the orthogonality of the trigonometric functions (or complex exponentials). We can understand better *why* these expansions exist by recognizing a connection to the concept of basis expansion of a vector in terms of a basis. Recall that we can write a vector $\mathbf{v}$ that is an element of an N-dimensional vector space as

$$\mathbf{v} = \sum_{\ell=1}^{N} v_\ell\, \hat{\boldsymbol{e}}_\ell\,, \qquad \hat{\boldsymbol{e}}_\ell \cdot \hat{\boldsymbol{e}}_n = \delta_{\ell n}\,, \qquad v_n = \mathbf{v} \cdot \hat{\boldsymbol{e}}_n\,. \tag{A.65}$$

In Eq. (A.65), the set of basis vectors $\{\hat{\boldsymbol{e}}_\ell\,;\ \ell \in \{1,\ldots,N\}\}$ is orthonormal (middle relation). We can prove the last relationship by noting

$$\mathbf{v}\cdot\hat{\boldsymbol{e}}_n = \left(\sum_\ell v_\ell \hat{\boldsymbol{e}}_\ell\right)\cdot\hat{\boldsymbol{e}}_n = \sum_\ell v_\ell\,(\hat{\boldsymbol{e}}_\ell\cdot\hat{\boldsymbol{e}}_n) = \sum_\ell v_\ell\,\delta_{\ell n} = v_n\,. \tag{A.66}$$

In a finite-dimensional vector space, the expansion in terms of basis vectors is familiar. For example, in two dimensions ($N = 2$),

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = v_1\begin{pmatrix} 1 \\ 0 \end{pmatrix} + v_2\begin{pmatrix} 0 \\ 1 \end{pmatrix} \equiv v_1\hat{\boldsymbol{e}}_1 + v_2\hat{\boldsymbol{e}}_2\,. \tag{A.67}$$

In the present case of Fourier series, our vector space is a (countably) infinite-dimensional function space of square-integrable functions defined on the interval $[0, T)$. The scalar (or "dot") product is defined by generalizing the definition in Section A.1.8 to

$$\mathbf{f}\cdot\mathbf{g} \equiv \frac{1}{T}\int_{-T/2}^{T/2} \mathrm{d}t\, f(t)\, g^*(t)\,, \tag{A.68}$$

where $g^*(t)$ denotes the complex conjugate of the function $g(t)$. The basis vectors are

$$\hat{\boldsymbol{e}}_\ell \equiv \mathrm{e}^{+\,\mathrm{i}\ell\omega t}\,, \qquad -\infty < \ell < \infty\,, \tag{A.69}$$

with orthogonality conditions

$$\hat{\boldsymbol{e}}_\ell\cdot\hat{\boldsymbol{e}}_n = \frac{1}{T}\int_{-T/2}^{T/2}\mathrm{d}t\,\mathrm{e}^{+\,\mathrm{i}\ell\omega t}\,\mathrm{e}^{-\,\mathrm{i}n\omega t} = \frac{1}{T}\int_{-T/2}^{T/2}\mathrm{d}t\,\mathrm{e}^{\mathrm{i}(\ell-n)\omega t} = \delta_{\ell n}\,. \tag{A.70}$$

Given this notation, the Fourier series is merely a basis expansion:

$$\mathbf{f} = \sum_\ell c_\ell\,\hat{\boldsymbol{e}}_\ell = \sum_\ell c_\ell\,\mathrm{e}^{+\,\mathrm{i}\ell\omega t}\,, \qquad c_n = \mathbf{f}\cdot\hat{\boldsymbol{e}}_n = \frac{1}{T}\int_{-T/2}^{T/2}\mathrm{d}t\, f(t)\,\mathrm{e}^{-\,\mathrm{i}n\omega t}\ . \tag{A.71}$$

The subtle point is to show that the set $\{\hat{\boldsymbol{e}}_\ell\,;\ \ell \in \mathbb{Z}\}$, with $\mathbb{Z}$ the set of all integers, is *complete*. In other words, one needs to prove that the set contains all the basis vectors. Finite-dimensional settings are simple: an N-dimensional vector space is spanned by N independent basis vectors. Any N linearly independent vectors can form a complete basis. Here, the set $\hat{\boldsymbol{e}}_\ell$ has a countable infinity of elements, as does the dimension of the function space. But does our set *span* the function space? Are any basis vectors missing? See the delightful book by Boyd (2000) for the full story on orthogonal function expansions.

A.4.2 Fourier Transforms

If we view a general continuous function as a periodic function with $T \to \infty$, then the Fourier series sum becomes an integral. To see this, let us rewrite the Fourier series expansion explicitly in terms of T:

$$f(t) = \sum_{n=-\infty}^{\infty} c_n\,\mathrm{e}^{+2\pi\,\mathrm{i}nt/T}\,, \qquad c_n = \frac{1}{T}\int_{-T/2}^{T/2}\mathrm{d}t\, f(t)\,\mathrm{e}^{-2\pi\,\mathrm{i}nt/T}\ . \tag{A.72}$$

Then, rewriting in terms of $\Delta\omega \equiv 2\pi/T$ gives

$$f(t) = \sum_{n=-\infty}^{\infty} \frac{\Delta\omega}{2\pi}\, T c_n\, \mathrm{e}^{+\mathrm{i}(n\,\Delta\omega)t}\,, \qquad T c_n = \int_{-T/2}^{T/2} \mathrm{d}t\, f(t)\, \mathrm{e}^{-\mathrm{i}(n\,\Delta\omega)t}\,. \tag{A.73}$$

In the limit $(n,\, T) \to \infty$, we have $\Delta\omega \to d\omega$ and $n\,\Delta\omega \to \omega$ and $T c_n \to F_f(\omega)$. Then,

$$f(t) = \int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi}\, F_f(\omega)\, \mathrm{e}^{+\mathrm{i}\omega t}\,, \qquad F_f(\omega) = \int_{-\infty}^{\infty} \mathrm{d}t\, f(t)\, \mathrm{e}^{-\mathrm{i}\omega t}\,. \tag{A.74}$$

Equation (A.74) defines the inverse and forward Fourier transforms, respectively.[1] In the main text of the book, we use the notation $f(\omega)$ rather than the more explicit $F_f(\omega)$.

The connection to linear algebra – expansion of a vector over a complete set of orthonormal basis vectors – also generalizes, but with subtleties. We define the basis element

$$\hat{\boldsymbol{e}}_\omega \equiv \mathrm{e}^{\mathrm{i}\omega t}\,, \qquad \mathbf{f}\cdot\mathbf{g} \equiv \int_{-\infty}^{\infty} \mathrm{d}t\, f(t)\, g^*(t)\,, \tag{A.75}$$

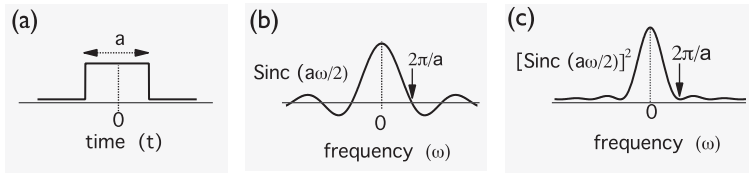
with ω now serving as a *continuous* index, in contrast to the discrete index ℓ that we defined for Fourier series. This vector (function) space has a continuous infinity of dimensions, with orthogonality relations

$$\hat{\boldsymbol{e}}_\omega \cdot \hat{\boldsymbol{e}}_{\omega'} = \int_{-\infty}^{\infty} \mathrm{d}t\, \mathrm{e}^{\mathrm{i}(\omega-\omega')t} = 2\pi\,\delta\,(\omega-\omega')\,, \tag{A.76}$$

where $\delta\,(\omega-\omega')$ is the *Dirac delta function*, which is the continuous-space equivalent of the Kronecker delta $\delta_{\ell n}$ used above. Loosely, the delta function is zero except at the origin $(\omega = \omega')$, where it is infinite, with unit "area." Interchanging $t \leftrightarrow \omega$ gives the time-domain delta function:

$$\delta\,(t-t') = \int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi}\, \mathrm{e}^{\mathrm{i}\omega(t-t')} = 1\,. \tag{A.77}$$

[1] There are many conventions for the Fourier transform. We follow a long tradition in physics (e.g., Stone and Goldbart [2009]) and write $\frac{\mathrm{d}\omega}{2\pi} = \mathrm{d}f$ to remind us that the expansion is physically over real, and not angular frequencies. Indeed, some authors write the Fourier integrals in terms of the frequency, including the exponentials, $\mathrm{e}^{\pm 2\pi\,\mathrm{i}\,ft}$. Using ω minimizes factors of 2π and is common in the physics literature. By contrast, we follow engineering conventions and use $\mathrm{e}^{-\mathrm{i}\omega t}$ for the forward and hence $\mathrm{e}^{+\mathrm{i}\omega t}$ for the inverse transform. In the physics literature, for example, Jackson (1999), Press et al. (2007), and Stone and Goldbart (2009), the opposite sign convention is more typical. No matter which convention you choose, be consistent!



(a) Rectangular "pulse" function; (b) its Fourier transform; (c) its power spectrum. **Fig. A.2**

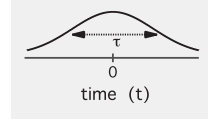
The delta "function" is pathological and is more properly specified as a *distribution* that is defined by its properties under integration. A better definition is thus

$$\int_{-\infty}^{\infty} \mathrm{d}t\, f(t)\, \delta\,(t-a) = f(a)\,, \tag{A.78}$$

for all suitable *test functions* $f(t)$. Thus, the delta function "picks out" the value of $f(t)$ at the zero of the argument of the delta function [$t = a$ in Eq. (A.78)].

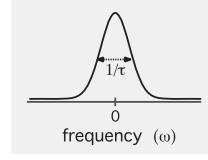
Example A.8 (Pulse) A standard, important example of Fourier analysis is the transform of a pulse of width a, or rect() function, shown in Figure A.2, below. The Fourier transform of the time-domain function $f(t)$ is $F_f(\omega) = a\ \mathrm{sinc}(a\omega/2)$, where $\mathrm{sinc}(x) \equiv \sin(x)/x$. In Figure A.2, we also show the *power spectrum*, given by the magnitude squared of the transform, $|F_f(\omega)|^2$, which here is just $a^2\ \mathrm{sinc}^2(\omega a/2)$. The power spectrum gives the power in a signal between frequencies ω and $\omega + \mathrm{d}\omega$.

Note that the first zero of $\mathrm{sinc}(\frac{1}{2}a\omega)$ occurs when the argument = π, which happens for $\omega = \frac{2\pi}{a}$. This means that the power of the signal is contained within a set of frequencies ("bandwidth") of roughly $\frac{2\pi}{a}$. Thus, the longer the pulse (the bigger a), the narrower the range of frequencies in the signal. In the limit $a \to \infty$, the power spectrum becomes a Dirac delta function, $\delta\,(\omega)$.



Example A.9 (Gaussian) Another important example is the Gaussian. In Eq. (A.79) and at left, we see that a pulse of width τ is another Gaussian, of width $1/\tau$.

$$f(t) = \mathrm{e}^{-\frac{1}{2}(t/\tau)^2} \qquad \Longrightarrow \qquad F_f(\omega) = \tau\sqrt{\pi}\ \mathrm{e}^{-\frac{1}{2}(\omega\tau)^2}\ . \tag{A.79}$$



Example A.10 (Convolution) One of the most useful Fourier properties involves the *convolution* of two functions $f(t)$ and $g(t)$:

$$[f * g](t) \equiv \int_{-\infty}^{\infty} \mathrm{d}t'\, f(t')\, g(t-t')\,. \tag{A.80}$$

The *Convolution Theorem* then states that

$$\mathcal{F}_f[f * g](\omega) = F_f(\omega)\, G_f(\omega)\,, \tag{A.81}$$

where $F_f(\omega)$ and $G_f(\omega)$ are the Fourier transforms of $f(t)$ and $g(t)$. We show this by writing

$$\begin{aligned}\mathcal{F}_f[f * g](\omega) &= \int_{-\infty}^{\infty} \mathrm{d}t \int_{-\infty}^{\infty} \mathrm{d}t'\, f(t')\, g(t-t')\, \mathrm{e}^{-\mathrm{i}\omega t} = \int_{-\infty}^{\infty} \mathrm{d}t'\, f(t') \int_{-\infty}^{\infty} \mathrm{d}t\, g(t-t')\, \mathrm{e}^{-\mathrm{i}\omega t} \\ &= \int_{-\infty}^{\infty} \mathrm{d}t'\, f(t')\, \mathrm{e}^{-\mathrm{i}\omega t'}\; G_f(\omega) = F_f(\omega)\, G_f(\omega)\,, \end{aligned} \tag{A.82}$$

where we use the Shift Theorem in evaluating the Fourier transform of $g(t)$. As an exercise, verify that the Fourier transform of the product $f(t)\, g(t)$ is given by

$$\mathcal{F}_f[fg](\omega) = \frac{1}{2\pi}[F_f * G_f](\omega)\,. \tag{A.83}$$

A.4.3 Laplace Transforms

For reasons described in the text (Section 2.3.1), we will mostly use the Laplace transform rather than the Fourier transform to analyze linear dynamical systems. The Laplace transform of a function $f(t)$ is given by

$$F(s) = \int_0^{\infty} \mathrm{d}t\, f(t)\, \mathrm{e}^{-st} \quad \equiv \mathcal{L}[f]\,. \tag{A.84}$$

Note that, unlike the main text, we use a different symbol, $F(s)$, to denote the Laplace transform of $f(t)$, in order to be clear. In the text, it is assumed that you are comfortable enough with the concept of transforms to use $f(s)$ to denote the Laplace transform of $f(t)$. But remember that they are different functions! Despite the potential confusion, it can be simpler to think of one "object" f, which has representations either in the time domain or the Laplace domain (or the Fourier domain).

We quickly list a few properties of the Laplace transform:

$$\begin{aligned} \mathcal{L}[af + bg] &= a\, F(s) + b\, G(s) && \text{linearity} \\ \mathcal{L}[\mathrm{e}^{at}\, f(t)] &= F(s-a) \\ \mathcal{L}[f(t-\tau)\theta(t-\tau)] &= \mathrm{e}^{-\tau s}\, F(s) && \text{delay} \\ \mathcal{L}[f(t/a)] &= a\, F(as) \\ \mathcal{L}\left[\frac{\mathrm{d}f}{\mathrm{d}t}\right] &= s\, F(s) - f(0) \\ \mathcal{L}\left[\frac{\mathrm{d}^2 f}{\mathrm{d}t^2}\right] &= s^2\, F(s) - s\, f(0) - \left.\frac{\mathrm{d}f}{\mathrm{d}t}\right|_{t=0} \end{aligned} \tag{A.85}$$

$$\mathcal{L}\left[\frac{\mathrm{d}^n f}{\mathrm{d}t^n}\right] = s^n F(s) - s^{n-1} f(0) - s^{n-2} \left.\frac{\mathrm{d}f}{\mathrm{d}t}\right|_{t=0} \cdots - \left.\frac{\mathrm{d}^{n-1} f}{\mathrm{d}t^{n-1}}\right|_{t=0}$$

$$\mathcal{L}\left[\int_0^t \mathrm{d}t'\, f(t')\right] = \frac{1}{s} F(s)$$

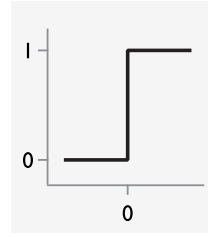
$$\mathcal{L}[f * g] = F(s)\, G(s)\,, \qquad \text{Convolution Theorem}$$

$$f(0) = \lim_{s\to\infty} [s\, F(s)]\,, \qquad \text{Initial-Value theorem}$$

$$f(\infty) = \lim_{s\to 0} [s\, F(s)]\,, \qquad \text{Final-Value theorem}\,. \tag{A.86}$$

In this case, the convolution of two functions ($h = f * g$) is defined as

$$h(t) = \int_0^\infty \mathrm{d}t'\, f(t')\, g(t - t')\,, \tag{A.87}$$



and $\theta(t)$ is the *Heaviside* step function (zero for $t < 0$, one for $t \geq 0$; see left). Note that the Laplace transform of a time derivative just multiplies the Laplace transform of the function by s. Conversely, integration corresponds to multiplying by $1/s$. In the Laplace domain, integration and differentiation are literally inverse operations. Finally, the last property, the convolution theorem, will be useful in exploring the dynamics of two systems in series.

There are similar relations for Fourier transforms. One difference is the presence of initial conditions in the transforms of derivatives. This is a useful feature if one wants to solve an initial-value problem. If not, they are a nuisance, and one usually assumes zero-initial conditions.

We can quickly prove the relationships for differentiation and integration, since we shall use them so often.

$$\begin{aligned}\int_0^\infty \mathrm{d}t\, \frac{\mathrm{d}f}{\mathrm{d}t}\, \mathrm{e}^{-st} &= f(t)\, \mathrm{e}^{-st}\Big|_0^\infty - \int_0^\infty \mathrm{d}t\, f(t)\,(-s)\, \mathrm{e}^{-st} \\ &= 0 - f(0)\cdot 1 + s\int_0^\infty \mathrm{d}t\, f(t)\, \mathrm{e}^{-st} = -f(0) + s\, F(s)\,.\end{aligned} \tag{A.88}$$

For zero initial conditions, this is just $\mathcal{L}[\frac{\mathrm{d}f}{\mathrm{d}t}] = s\, F(s)$.

To show that $\mathcal{L}[\int \mathrm{d}t'\, f(t')] = F(s)/s$, we write

$$\begin{aligned}\mathcal{L}\left[\int_0^t \mathrm{d}t'\, f(t')\right] &= \int_0^\infty \mathrm{d}t \left(\mathrm{e}^{-st}\int_0^t f(t')dt'\right) = \cancelto{0}{\left(-\frac{\mathrm{e}^{-st}}{s}\right)\int_0^t \mathrm{d}t'\, f(t')\Big|_0^\infty} + \int_0^\infty \mathrm{d}t\, \frac{\mathrm{e}^{-st}}{s} f(t) \\ &= \frac{1}{s}\int_0^\infty \mathrm{d}t\, \mathrm{e}^{-st}\, f(t) = \frac{F(s)}{s}\,.\end{aligned} \tag{A.89}$$

Let us also prove the initial- and final-value theorems.

$$\begin{aligned}s\, F(s) &= \int_0^\infty \mathrm{d}t\, f(t) s\, \mathrm{e}^{-st} = -\int_0^\infty \mathrm{d}t\, f(t) \frac{\mathrm{d}}{\mathrm{d}t}\, \mathrm{e}^{-st} = -f(t)\, \mathrm{e}^{-st}\Big|_0^\infty + \int_0^\infty \mathrm{d}t\, \frac{\mathrm{d}f}{\mathrm{d}t}\, \mathrm{e}^{-st} \\ &= f(0)\cdot 1 + \int_0^\infty \mathrm{d}t\, \frac{\mathrm{d}f}{\mathrm{d}t}\, \mathrm{e}^{-st} = \begin{cases} f(0) & s \to \infty\,, \\ f(\infty) & s \to 0\,. \end{cases}\end{aligned} \tag{A.90}$$

Example A.11 (Initial- and Final-Value Theorems)

$$F(s) = \frac{1}{s+a} \qquad \Longrightarrow \qquad f(t) = \mathrm{e}^{-at}$$

$$s\,F(s) = \frac{s}{s+a} \qquad \Longrightarrow \qquad \begin{cases} 1 = f(0) & s \to \infty\,, \\ 0 = f(\infty) & s \to 0\,. \end{cases}$$

The Final-Value Theorem is particularly useful to check whether the function $f(t)$ whose Laplace transform is $F(s)$ is bounded as $t \to \infty$. For example, if $F = 1/s^2$, then we immediately see that as $s \to 0$, then $f(\infty) = sF(s) \to \infty$. Since $f(t) = t$ in this case, the conclusion is not surprising. The nice feature of the theorem is that we can know $f(0)$ and $f(\infty)$ even if we cannot transform back from $F(s)$.

One useful application of the Final-Value Theorem is to compute the DC (zero frequency) gain of a transfer function $G(s)$. The DC gain is the ratio, after transients have decayed, of the output $y(t)$ to the input $u(t)$. If we let $u(t)$ be a step function, then $u(s) = \frac{1}{s}$ and $y(t \to 0) = \lim_{s\to 0}[s\,G(s)\,\frac{1}{s}]$. Thus, DC gain $= \lim_{s\to 0} G(s)$.

One caveat about the Final-Value Theorem is that a final value must be well-defined, in the sense that the initial value of the signal must decay to zero. That is, the signal $F(s)$ should have all its poles in the left-hand side of the complex s-plane. Recall that poles with Re $s > 0$ imply an exponentially growing amplitude, and poles with Re $s = 0$ oscillate without decay. In both cases, we cannot define a meaningful final value. To see an explicit example of such a failure, consider

$$F(s) = \frac{1}{s-a} \qquad \Longrightarrow \qquad f(t) = \mathrm{e}^{+at}\ .$$

The Final-Value theorem predicts $f(t \to \infty) = \lim_{s=0} \frac{s}{s-a} = 0$, instead of ∞. The theorem implicitly depends on having initial values decay to zero, which is not the case here.

It is also useful to collect a few common transforms:

$$\begin{aligned}
\mathcal{L}[\,\delta(t)] &= \int_0^\infty \mathrm{d}t\ \delta(t)\ \mathrm{e}^{-st} = 1 \\
\mathcal{L}[\theta(t)] &= \int_0^\infty \mathrm{d}t\ 1 \cdot \mathrm{e}^{-st} = \frac{1}{s} \\
\mathcal{L}[t] &= \int_0^\infty \mathrm{d}t\ t\ \mathrm{e}^{-st} = -\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{1}{s}\right) = \frac{1}{s^2} \\
\mathcal{L}[t^n] &= \int_0^\infty \mathrm{d}t\ t^n\ \mathrm{e}^{-st} = (-1)^n \frac{\mathrm{d}^n}{\mathrm{d}s^n}\left(\frac{1}{s}\right) = \frac{n!}{s^{n+1}} \\
\mathcal{L}[\mathrm{e}^{-at}] &= \frac{1}{s+a}\,, && \text{pole at } s = -a \\
\mathcal{L}[\sin \omega t] &= \frac{\omega}{s^2+\omega^2}\,, && \text{pole at } s = \pm\,\mathrm{i}\,\omega \\
\mathcal{L}[\mathrm{e}^{\mathrm{i}\omega t}] &= \frac{1}{s-\mathrm{i}\,\omega} && \qquad (A.91)
\end{aligned}$$

In writing down these identities, we see that it is useful to consider complex values of s. Of course, when $s = \mathrm{i}\omega$, we "almost" have a Fourier transform. (The limits are 0 to ∞, not $-\infty$ to $+\infty$.) In many cases, we see that the complex function of s has *poles* at specific points in the complex s-plane. For example, $\mathcal{L}[\mathrm{e}^{-at}] = 1/(s+a)$ has a pole at $s = -a$, while $\mathcal{L}[\sin \omega t] = \omega/(s^2 + \omega^2)$ has poles at $s = \pm \mathrm{i}\omega$ on the negative real axis. Similarly, a transform may vanish at *zeros* in the complex plane, as well.

A.4.4 Partial-Fraction Decomposition

A commonly used trick that can improve numerical robustness is to replace a rational polynomial $G(s) = N(s)/D(s)$ by a sum of simple fractions. Although this can be done for very general kinds of functions, we are interested in cases where N and D are polynomials over the reals. We start with the case where the poles of $G(s)$ (zeros of $D(s)$) are all real and nondegenerate. Then, we seek to write

$$G(s) = \frac{(s+z_1)(s+z_2)\ldots(s+z_m)}{(s+p_1)(s+p_2)\ldots(s+p_n)} = \frac{a_1}{s+p_1} + \frac{a_2}{s+p_2} + \cdots + \frac{a_n}{s+p_n}\,. \tag{A.92}$$

By substituting into the right-hand side of Eq. (A.92), it is easy to see that we should write

$$a_j = G(s)(s-p_j)\Big|_{s=p_j}\,. \tag{A.93}$$

The basic idea is that all terms in the sum vanish, except the one proportional to $(s-p_j)^{-1}$.

Example A.12 We seek to write

$$G(s) = \frac{s+2}{(s+3)(s+4)} = \frac{a_1}{s+3} + \frac{a_2}{s+4} \tag{A.94}$$

Equation (A.93) then implies

$$\begin{aligned} a_1 &= \frac{(s+2)(s+3)}{(s+3)(s+4)}\bigg|_{s=-3} = \frac{s+2}{s+4}\bigg|_{s=-3} = \frac{-1}{1} = -1 \\ a_2 &= \frac{(s+2)(s+4)}{(s+3)(s+4)}\bigg|_{s=-4} = \frac{s+2}{s+3}\bigg|_{s=-4} = \frac{-2}{-1} = 2\,, \end{aligned} \tag{A.95}$$

so that

$$G(s) = \frac{s+2}{(s+3)(s+4)} = \frac{-1}{s+3} + \frac{2}{s+4}\,, \tag{A.96}$$

which can be verified directly. (Once you get the hang of it, you can write down the coefficients very quickly; or use a computer-algebra program to help.)

If the denominator has a pair of complex-conjugate poles, the denominator $\sim (s^2 + p^2)$, and the partial-fraction decomposition should include a term of the form

$$G(s) = \cdots + \frac{a+bs}{s^2+p^2} + \cdots\,, \tag{A.97}$$

where a and b can be determined by evaluating $G(s)(s^2 + p^2)$ at $s = p$ or p^*. (You get the same result no matter which pole you choose for the evaluation.)

Example A.13 Find the decomposition

$$G(s) = \frac{s}{(s^2+1)(s+2)} = \frac{a_1 + b_1 s}{s^2+1} + \frac{a_2}{s+2}\,. \tag{A.98}$$

Using the above principles, we have

$$\begin{aligned} a_1 + b_1 s|_{s=\mathrm{i}} &= G(s)(s^2+1)\big|_{s=\mathrm{i}} = \frac{s}{s+2}\bigg|_{s=\mathrm{i}} = \frac{\mathrm{i}}{\mathrm{i}+2} = \frac{1+2\,\mathrm{i}}{5} \\ a_2 &= G(s)(s+2)|_{s=-2} = \frac{s}{s^2+1}\bigg|_{s=-2} = \frac{-2}{5}\,, \end{aligned} \tag{A.99}$$

so that

$$G(s) = \frac{s}{(s^2+1)(s+2)} = \frac{1}{5}\left(\frac{1+2s}{s^2+1} - \frac{2}{s+2}\right). \tag{A.100}$$

A.4.5 Solving Linear ODEs with Initial Conditions

An important application of Laplace transforms is the solution of linear ordinary differential equations with constant coefficients. For example, consider

$$\ddot{x} + 5\dot{x} + 6x = 0\,, \quad x(0) = \dot{x}(0) = 1\,. \tag{A.101}$$

Take Laplace transforms of each element of the equation:

$$[s^2 X(s) - sx(0) - \dot{x}(0)] + 5[sX(s) - x(0)] + 6X(s) = 0\,, \tag{A.102}$$

where $X(s) = \mathcal{L}[x(t)]$. Then

$$(s^2 + 5s + 6)X(s) = (s+5)x(0) + \dot{x}(0) = s + 6\,, \tag{A.103}$$

and

$$X(s) = \frac{s+6}{(s+2)(s+3)} = \frac{4}{s+2} - \frac{3}{s+3}\,, \tag{A.104}$$

using the partial-fraction decomposition from Section A.4.4.

We can now refer to our table of Laplace transforms to transform back to $x(t)$:

$$x(t) = 4\,\mathrm{e}^{-2t} - 3\,\mathrm{e}^{-3t}\,. \tag{A.105}$$

You can verify that $x(0) = \dot{x}(0) = 1$.

Notice that we have avoided writing down the explicit expression for the inverse Laplace transform. If manipulating the transform into a common form that exists in tables is possible, do it. Otherwise, the general expression is an integral in the complex s-plane:

$$\mathcal{L}^{-1}[f(s)] = \frac{1}{2\pi\,\mathrm{i}} \int_{\sigma-\mathrm{i}\infty}^{\sigma+\mathrm{i}\infty} \mathrm{d}s\, F(s)\, \mathrm{e}^{st}\,, \qquad \sigma > \text{Re of all poles}\,. \tag{A.106}$$

If the poles are all in the left-hand side of the complex s-plane, one generally uses $\sigma = 0$. Equation A.106 is generally evaluated using the residue theorem.

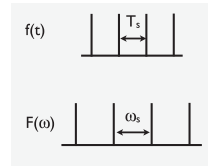
Problem A.4.1 Let us calculate some simple Fourier series.

a. *Square wave*. Verify the coefficients given in Eq. (A.60) of the Fourier series for the square wave ⎍⎍, defining it to be an even function about $t = 0$.
b. *Square wave with variable duty cycle*. Find the coefficients of an even, asymmetric square wave that equals 1 for a quarter period and 0 for the rest, ⎍⎍.
c. An even function satisfies $f(t) = -f(-t)$. Show that if the function also satisfies $f(t) = -f(t + T/2)$, the even cosine terms will vanish in the Fourier series.
d. *Triangle wave*. Find the coefficients of the Fourier series for the even triangle wave ⩘. Show, in particular, that $a_n \sim O(\frac{1}{n^2})$.

Problem A.4.2 **Poisson summation formula**. Prove the following version of the Poisson summation formula, which relates Fourier coefficients to Fourier transforms for a periodic function $f(t) = f(t + T)$ built out of non-periodic "basis" functions $g(t)$. Show that $f(t) = \sum_{k=-\infty}^{\infty} g(t + kT) = \frac{1}{T} \sum_{n=-\infty}^{\infty} G(n\omega)\, \mathrm{e}^{\mathrm{i}n\omega t}$, where $\omega = \frac{2\pi}{T}$ and the Fourier transform $G(\omega) = \int_{-\infty}^{\infty} \mathrm{d}t\, g(t)\, \mathrm{e}^{-\mathrm{i}\omega t}$.

Problem A.4.3 **Parseval's theorem**. Show that $\int_{-\infty}^{\infty} \mathrm{d}t\, [f(t)]^2 = \int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} \left|F_f(\omega)\right|^2$. Check the relation explicitly for $f(t) = \theta(t)\, \mathrm{e}^{-t}$, with $\theta(t)$ the Heaviside step function.

Problem A.4.4 **Fourier transform of a comb function**. By applying the Poisson summation formula to the delta function, $g(t) = \delta(t)$, show that the Fourier transform of the time-domain comb function, $f(t) = \sum_k \delta(t - kT_\mathrm{s})$, is the frequency-domain comb function $F(\omega) = \frac{2\pi}{T_\mathrm{s}} \sum_n \delta(\omega - n\omega_\mathrm{s})$.



Problem A.4.5 **Laplace transform of integral**. Show that $\mathcal{L}[\int_0^t \mathrm{d}t'\, f(t')] = \frac{1}{s} F(s)$.

A.5 Optimization

We first encounter the elementary notion of *unconstrained optimization* and *convex functions*. A function is convex if the graph of every line segment drawn between two points lies above the function itself (Figure A.3a). To find the minimum of a convex function, take its derivative to find its *critical point*. For example, it is easy to check that $f(x) = \frac{1}{2}x^2$ is convex. Its critical point is determined by taking a derivative and setting it to zero:

$$\frac{\mathrm{d}f}{\mathrm{d}x} = 0 \quad \Longrightarrow \quad x = 0\,. \tag{A.107}$$

If a function is convex, there is at most one critical point, and if it exists, it corresponds to a *global* minimum. If a function is not convex, setting the derivative of a function equal to zero is a necessary but not sufficient condition to have a minimum.

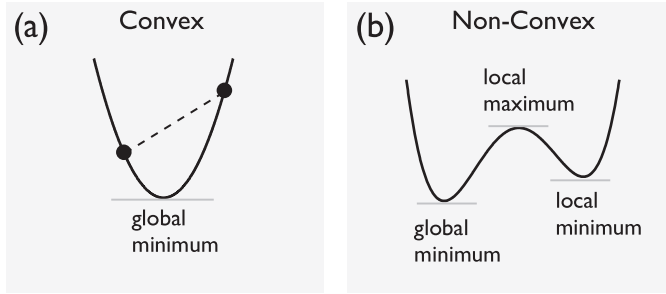


Fig. A.3 Optimization of one-dimensional, continuous functions. (a) A convex function has at most one minimum. (b) A nonconvex function may have several critical points, which can be local minima, maxima, or saddle points (not shown).

In particular, functions may have multiple critical points (Figure A.3b). Examining the second derivative at a critical point, we can determine whether a critical point is a *local* minimum ($f'' > 0$), a local maximum ($f'' < 0$), or a point of inflection ($f'' = 0$). If a function has several local minima, we may need to evaluate the function at each one to determine the global minimum. Note that we can always convert a maximization problem into a minimization problem, as maximizing f is equivalent to minimizing $-f$.

With more variables, the procedure is similar. For example, $f(x, y) = \frac{1}{2}\left(x^2 + y^2\right)$ is also convex, and we can find its critical point by equating its *gradient* to zero:

$$\nabla f = \begin{pmatrix} \partial_x f & \partial_y f \end{pmatrix} = \begin{pmatrix} x & y \end{pmatrix} = \begin{pmatrix} 0 & 0 \end{pmatrix} \quad \Longrightarrow \quad x = y = 0\,, \tag{A.108}$$

where we use the notation $\partial_x f \equiv \frac{\partial f}{\partial x}$ for clarity. The gradient of f is a (row) vector that points in the direction of increasing f.

As we add variables, the number of local extrema will increase. In general, extrema include maxima, minima, and *saddle points*, where the curvature has opposite signs in different directions. Optimization is much easier if you can make sure that the function you want to optimize is convex. Then, since there is a unique critical point, we can usually solve for it directly. For problems with more complicated functions, we distinguish *local* algorithms, which try to "walk downhill," from *global* algorithms, which try many different candidate starting points. The former can usually find a local minimum, whereas the latter have a chance of finding the global minimum.

A.5.1 Constrained Optimization and Lagrange Multipliers

Often, optimization is done under constraints. Instead of optimizing over all possible values of variables (e.g., x and y in the example above), we wish to find the optimum when only a subset of x–y combinations is allowed. In Figure A.4, the surface represents a function $f(x, y)$ and the constraint is a curve defined by $g(x, y) = 0$. The goal is to minimize f along the curve defined by $g = 0$.

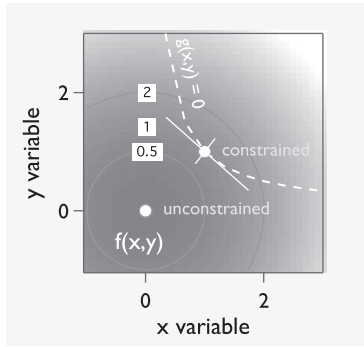


Fig. A.4

Constrained optimization: We minimize the function $f(x, y)$ subject to the constraint $g(x, y) = 0$. The background intensity indicates the value of $f(x, y)$. Light circles and tags indicate contours of $f(x, y)$ at $f = 0.5$, $f = 1$, and $f = 2$, and the big dot at $(x, y) = (0, 0)$ is the global minimum. Note that both the contours and the gradients of $f(x, y)$ and $g(x, y)$ are parallel at the constrained minimum at $(1, 1)$.

Equality Constraints

If the constraint is simple enough that we can rewrite $g(x, y) = 0$ as a function $y(x)$, then a straightforward way of incorporating the constraint into the minimization of $f(x, y)$ is to substitute for y and write $f(x, y) = f[x, y(x)] \equiv f_1(x)$ and then to minimize f_1 with respect to x, as normal. In general, though, it may not be easy to find the explicit solution for $y(x)$. The method of Lagrange multipliers, described below, works in all cases.

Figure A.4 illustrates the solution to this problem geometrically. We see that the critical point occurs where the curve g is tangent to a level-set contour of f. The algebra is easier if we work with gradients: the critical point occurs where the gradient of f (which is perpendicular to the contours) is parallel to the gradient of g:

$$\boldsymbol{\nabla} f = -\lambda \boldsymbol{\nabla} g \,, \tag{A.109}$$

where the proportionality constant is known as a *Lagrange multiplier*. (The minus sign is a historical convention.) Intuitively, if we walk along $g = 0$, then we are at the local minimum when the gradient of f is perpendicular to the local tangent vector along $g = 0$. Otherwise, continuing along $g = 0$ in the right direction could reduce the value of f. The local gradient of f is thus perpendicular to the tangent along $g = 0$. We can also look at the gradient of the function $g(x, y)$, which is perpendicular to the level sets defined by $g(x, y) =$ constant, implying that the gradients of the functions f and g are parallel to each other at the extremum of f constrained to $g = 0$.

A standard trick is then to define a new variable λ and create a new function,

$$L(x, y, \lambda) = f(x, y) + \lambda g(x, y) \,. \tag{A.110}$$

The function L is known as the *Lagrangian*. In applications to classical mechanics and other physical problems, it often is the Lagrangian function in that context as well, but it need not have a direct physical interpretation. We then solve for the *unconstrained*

critical point of L by imposing that the gradient of L taken with respect to the variables x and y and the Lagrange multiplier λ be zero:

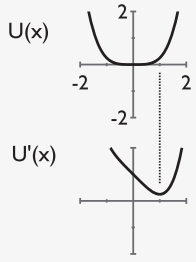
$$\frac{\partial L}{\partial x} = \frac{\partial f}{\partial x} + \lambda \frac{\partial g}{\partial x} = 0\,, \qquad \frac{\partial L}{\partial y} = \frac{\partial f}{\partial y} + \lambda \frac{\partial g}{\partial y} = 0\,, \qquad \frac{\partial L}{\partial \lambda} = g = 0\,. \tag{A.111}$$

The first two equations are the components of $\boldsymbol{\nabla} f = -\lambda \boldsymbol{\nabla} g$, and the last expresses the constraint. Taking derivatives of L with respect to all the variables, including the Lagrange multiplier λ, gives the solution to the constrained-optimization problem. The constrained optimization of $f(x, y)$ goes to an equivalent unconstrained optimization of $L(x, y, \lambda)$.

The above picture generalizes easily to more constraints, each with its own Lagrange multiplier λ_i:

$$L = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i\, g_i(\boldsymbol{x}) = f + \boldsymbol{\lambda}^{\mathsf{T}} \boldsymbol{g}\,, \tag{A.112}$$

where $\boldsymbol{\lambda}$ is now an m-dimensional vector and $\boldsymbol{g}$ is the vector of m constraint functions $g_i(\boldsymbol{x}) = 0$. (Here, $\boldsymbol{x}$ is a n-dimensional vector that, in this book, will typically be the state vector of a dynamical system.) If a constraint needs to be enforced at each time, the Lagrange multiplier vector $\boldsymbol{\lambda}(t)$ will be a continuous function of time. An example is treated in Section 7.2 on optimal control, where the dynamics act as the constraint.



Example A.14 As a trivial but instructive example, consider minimizing a one-variable function $U(x)$ subject to the constraint $x = x_0$. Obviously, the minimum value is $U(x_0)$, but let us see how this works with Lagrange multipliers:

$$U'(x) = U(x) + \lambda(x - x_0) \quad \Longrightarrow \quad \frac{\partial U'}{\partial x} = \frac{\partial U}{\partial x} + \lambda = 0 \quad \Longrightarrow \quad \lambda = -\frac{\partial U}{\partial x}\,. \tag{A.113}$$

That $x = x_0$ then follows from taking $\partial_\lambda U' = 0$. If we interpret $U(x)$ as a potential – that is, the reason for changing notation – then λ is just the force that the constraint needs to supply to "counteract" the force that the potential exerts on the "particle" at position x_0.

We illustrate this idea at right, where $U(x) = \frac{1}{4}x^4$ and $x_0 = 1$. The Lagrange multiplier is $\lambda = -1$, and the modified potential $U'(x) = \frac{1}{4}x^4 - x + 1$ does indeed have a minimum at $x = 1$. In effect, the constraint alters the potential so that the new system has a global minimum at the desired value of x.

Next, we present a more typical example, using a quadratic function of two variables subject to a linear constraint.

Example A.15 Let $f = \frac{1}{2}(x^2 + y^2)$ and look for the minimum value along the curve $y - 1/x = 0$ for $x > 0$. We define L as

$$L = \frac{1}{2}\left(x^2 + y^2\right) + \lambda\left(y - \frac{1}{x}\right), \tag{A.114}$$

which implies

$$\frac{\partial L}{\partial x} = x + \frac{\lambda}{x^2} = 0\,, \qquad \frac{\partial L}{\partial y} = y + \lambda = 0\,, \qquad \frac{\partial L}{\partial \lambda} = y - \frac{1}{x} = 0\,. \tag{A.115}$$

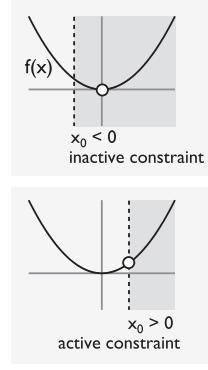
Solving the three equations gives $x^* = y^* = 1$, $\lambda = -1$, and $f(x^*, y^*) = 1$ (Figure A.4).

Alternately, we can express the constraint as $y = \frac{1}{x}$ and write $f[x, y(x)] = f_1(x)$

$$f_1(x) = \frac{1}{2}\left(x^2 + \frac{1}{x^2}\right) \quad \Longrightarrow \quad \frac{\mathrm{d}f_1}{\mathrm{d}x} = x - \frac{1}{x^3} = 0 \quad \Longrightarrow \quad x^* = \pm 1\,. \tag{A.116}$$

The two methods thus give the same solution, as the constraint curve satisfies $x > 0$.

Inequality Constraints



In control theory, we will often want to optimize a problem that has an inequality constraint. A typical goal (Section 7.5) is to maximize the performance (or minimize a cost function) of a controller while requiring the control inputs $u(t)$ stay within upper and lower bounds.

To understand the basic idea, consider minimizing $f(x) = \frac{1}{2}x^2$ subject to the constraint $x > x_0$, represented as the dark shaded area at left. There are two cases: If $x_0 < 0$, then the minimum occurs at the unconstrained value, $x^* = 0$, and the constraint is *inactive*. If $x_0 > 0$, then the minimum is the constraint border, $x^* = x_0$, and the constraint is *active*.

We can connect this simple example to our Lagrange-multiplier formalism by the following trick: We convert the constraint $g(\boldsymbol{x}) \leq 0$ into an equality constraint by introducing a *slack variable* s^2 such that $g(\boldsymbol{x}) + s^2 = 0$. That is, if the inequality constraint is satisfied, then there will exist some (real) s that makes $g + s^2 = 0$. If not, there won't. We thus write a Lagrangian of the form $L = f(\boldsymbol{x}) + \lambda[g(\boldsymbol{x}) + s^2]$.

We then proceed to take derivatives of L with respect to $\boldsymbol{x}$, λ, and s:

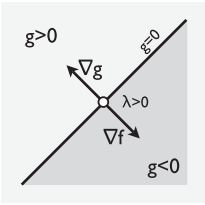
$$\boldsymbol{\nabla} L = \boldsymbol{\nabla} f + \lambda \boldsymbol{\nabla} g = \mathbf{0}\,, \qquad \frac{\partial L}{\partial \lambda} = g + s^2 = 0\,, \qquad \frac{\partial L}{\partial s} = 2\lambda s = 0\,. \tag{A.117}$$

The last relation in Eq. (A.117), $\lambda s = 0$, implies that either $\lambda = 0$ or $s = 0$. (We ignore the special case where both are zero.) If $\lambda = 0$ and $s \neq 0$, then the inequality constraint is satisfied at an interior point, and the constraint is *inactive*. We can then find the critical point as an unconstrained optimization problem, by solving $\boldsymbol{\nabla} f = \mathbf{0}$. Alternatively, if $s = 0$, then the constraint is *active*: we are at the boundary of the inequality constraint, which can now be treated as an equality constraint and solved by the method of Lagrange multipliers.

In the example above in the margin, $L = \frac{1}{2}x^2 + \lambda(-x + x_0 + s^2)$. Taking derivatives,

$$\frac{\partial L}{\partial x} = x - \lambda = 0\,, \qquad \frac{\partial L}{\partial \lambda} = -x + x_0 + s^2 = 0\,, \qquad \frac{\partial L}{\partial s} = 2\lambda s = 0\,,$$

$$x^* = \lambda\,, \qquad s^2 = x^* - x_0 \qquad \lambda = 0 \text{ or } s = 0\,, \tag{A.118}$$



For the two cases, we have

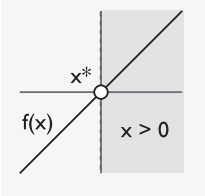
1 $\lambda = 0$ (inactive constraint): Then $x^* = 0$ and $s^2 = -x_0$.
2 $s = 0$ (active constraint): Then $x^* = x_0$ and $\lambda = x^* = x_0$, as we saw above.

Notice that when the constraint is active, we have $x_0 > 0$ and hence that $\lambda > 0$. As illustrated at right, this last result is no accident. Since the optimization is constrained, $\boldsymbol{\nabla} f = -\lambda \boldsymbol{\nabla} g$ at the extremum. For a hard constraint, that is the end of the story. For an inequality constraint, a move *inside* the constraint region may lower f. Since g is zero on the boundary and negative inside it, we know that $\boldsymbol{\nabla} g$ points away from the constraint region. If $\lambda > 0$, then $\boldsymbol{\nabla} f$ points into the constraint region, meaning that any move in that direction increases f. Thus, we add the condition $\lambda > 0$.

We can thus summarize the necessary requirements for a point $\boldsymbol{x}^*$ to be an extremum in what are known as the *Karush-Kuhn-Tucker* (KKT) conditions:

$$\boldsymbol{\nabla} f(\boldsymbol{x}^*) = -\sum_{i=1}^{m} \lambda_i \boldsymbol{\nabla} g_i(\boldsymbol{x}^*)\,, \qquad \lambda_i \geq 0\,, \qquad g_i(\boldsymbol{x}^*) \leq 0 \qquad \lambda_i g_i(\boldsymbol{x}^*) = 0 \tag{A.119}$$

These requirements cover all combinations of active and inactive constraints and are necessary but not sufficient: the second derivatives show whether the extremum corresponds to a local minimum, maximum, or saddle point. Note that for equality constraints, the Lagrange multipliers λ can have either sign, whereas for inequality constraints, $\lambda \geq 0$.



Finally, an important corollary to our discussion is that $f(\boldsymbol{x})$ need not have any critical points in order for there to be a solution of the constrained-inequality optimization. At right, $f(x) = x$ is a straight line and never satisfies $f'(x) = 0$. However, if we add the constraint $x > 0$, then f is minimized at zero, as shown.

A.5.2 Calculus of Variations

So far, we have considered optimization over the variation of a finite number of variables. In discussing continuous dynamics, we will want to optimize a *functional* (function of a function) over a set of possible curves. If we think of each point on a curve $x(t)$ as a variable to optimize, then we have a continuous infinity of optimization variables, one for each point in time. The optimization problem is then to minimize a functional $J[x(t)]$,

$$J = \int_0^\tau \mathrm{d}t\, L(x, \dot{x}, t)\,. \tag{A.120}$$

In Eq. (A.120), we choose a function $x(t)$, with fixed endpoints $x(0) = x_0$ and $x(\tau) = x_\tau$. The basic strategy is to assume that there is some $x^*(t)$ that minimizes J and to require that J be stationary with respect to infinitesimal variations of x about x^*. Of course, as with all the optimization problems discussed in this section, we need to check second

derivatives to make sure that we have a local minimum, as opposed to a maximum or saddle point.

The first trick is to realize that we can write an arbitrary $x(t)$ in terms $x^*(t)$ by defining $x(t) = x^*(t) + \varepsilon\delta x(t)$, where $\delta x(t)$ is an *arbitrary* function that vanishes at the boundaries, $\delta x(0) = \delta x(\tau) = 0$. We need δx to obey these boundary conditions to respect the boundary conditions for $x(t)$. Also, we can take a time derivative of x, which gives $\dot{x} = \dot{x}^* + \varepsilon\delta\dot{x}$. The cost J then becomes an ordinary function of ε:

$$J(\varepsilon) = \int_0^\tau \mathrm{d}t\, L(x^* + \varepsilon\delta x, \dot{x}^* + \varepsilon\delta\dot{x}, t) = J(0) + \varepsilon \int_0^\tau \mathrm{d}t \left(\frac{\partial L}{\partial x}\delta x + \frac{\partial L}{\partial \dot{x}}\delta\dot{x}\right) + O(\varepsilon^2)\,. \tag{A.121}$$

In Eq. (A.121), the variations δx and $\delta\dot{x}$ are not independent. We then use a second trick, which is to integrate the time-derivative term by parts:

$$\int_0^\tau \mathrm{d}t \left(\frac{\partial L}{\partial \dot{x}}\,\delta\dot{x}\right) = \frac{\partial L}{\partial \dot{x}}\,\delta x\Big|_0^\tau - \int_0^\tau \mathrm{d}t \left(\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{x}}\,\delta x\right) \tag{A.122}$$

Substituting Eq. (A.122) back into Eq. (A.121) and requiring $\frac{\mathrm{d}J}{\mathrm{d}\varepsilon} = 0$ gives

$$\int_0^\tau \mathrm{d}t \left(\frac{\partial L}{\partial x} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{x}}\right)\delta x(t) = 0\,. \tag{A.123}$$

Since Eq. (A.123) must hold for arbitrary $\delta x(t)$, we deduce the Euler–Lagrange equation,

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{x}} - \frac{\partial L}{\partial x} = 0\,, \tag{A.124}$$

In our application of Eq. (A.124) to problems in optimal control, we typically need to generalize in several ways: First, we add more variables. Since each can be varied independently, we find one Euler–Lagrange equation for each variable. If the variables are gathered into a state vector $\boldsymbol{x}$, then the Euler–Lagrange equations are the same as in Eq. (A.124), with the simple subsitution $x \to \boldsymbol{x}$. Next, the equations of motion are typically imposed as constraints at each time t via a Lagrange multiplier $\boldsymbol{\lambda}(t)$, as explained in Section 7.2.

The last generalization is to allow boundary variations. Thus, instead of requiring $x(\tau) = x_\tau$, we let its value be free and then penalize deviations from a preferred state. Consider a one-dimensional cost function

$$J = \varphi[x(\tau), \tau] + \int_0^\tau \mathrm{d}t\, L(x, \dot{x}, t)\,, \tag{A.125}$$

where the *endpoint cost* φ penalizes, but does not forbid, deviations of the end point $x(\tau)$ from a desired state x_τ. As a specific example, we could have $\varphi = \frac{1}{2}S\,[x(\tau) - x_\tau]^2$. Show that the usual Euler–Lagrange equations apply, but the boundary condition at $t = \tau$ becomes

$$\frac{\partial \varphi}{\partial x} + \frac{\partial L}{\partial \dot{x}} = 0\,. \tag{A.126}$$

We have

$$J(\varepsilon) = \varphi[x^* + \varepsilon\delta x, \tau] + \int_0^\tau \mathrm{d}t\, L(x^* + \varepsilon\delta x, \dot{x}^* + \varepsilon\delta\dot{x}, t)$$
$$= J(0) + \varepsilon\left\{ \frac{\partial\varphi}{\partial x}\delta x\bigg|_\tau + \int_0^\tau \mathrm{d}t \left(\frac{\partial L}{\partial x}\delta x + \frac{\partial L}{\partial \dot{x}}\delta\dot{x} \right) + O(\varepsilon^2) \right\} .$$

Then, we integrate by parts, as above, noting that while $\delta x(0) = 0$, the other endpoint, $\delta x(\tau)$, is free to vary and thus does *not* vanish. Then, including the boundary term $\partial_{\dot{x}} L(\tau)\, \delta x(\tau)$ from the integration by parts, we have

$$J(\varepsilon) = J(0) + \varepsilon\left\{ \left(\frac{\partial\varphi}{\partial x} + \frac{\partial L}{\partial \dot{x}} \right)\delta x\bigg|_\tau + \int_0^\tau \mathrm{d}t \left(\frac{\partial L}{\partial x} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{x}} \right) \delta x(t) \right\} .$$

Taking $\frac{\mathrm{d}J}{\mathrm{d}\varepsilon} = 0$ and noting that the equation must hold for all suitable functions $\delta x(t)$ gives

$$\left(\frac{\partial\varphi}{\partial x} + \frac{\partial L}{\partial \dot{x}} \right)\bigg|_\tau = 0\,, \qquad \frac{\partial L}{\partial x} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{x}} = 0\,.$$

The first of these becomes a boundary condition at $t = \tau$ for the differential equation (Euler–Lagrange) for $x(t)$. Because there is also an initial condition $x(0) = x_0$, we will have a two-point boundary-value problem on the interval $[0, \tau]$.

A.6 Probability Theory

We start by defining some elementary notions and terminology. Let the *sample space* Ω be the set of all possible outcomes x (of an experiment, measurement, card hand, etc.). For example, if a coin is tossed once, the sample space is heads or tails, { H,T }. If it is tossed N times, Ω is the set of 2^N possible outcomes. For two six-sided dice tossed once, the sample space consists of 36 possible outcomes, { (1,1), (2,1), …, (5,6), (6,6) }.

Subsets $E \subset \Omega$ are known as *events*. For a two six-sided dice toss, an event could be "you rolled a seven" and corresponds to the subset of outcomes that sum to eight (6+2, 5+3, 4+4, 3+5, 2+6), which has *cardinality* (size) equal to five. The *intersection* of two events $x_1 \cap x_2$ denotes the set of outcomes ω where $\omega \in x_1$ and $\omega \in x_2$. Two events x_1 and x_2 are *disjoint*, or *mutually exclusive*, if their intersection, $x_1 \cap x_2$, is the empty set, $\emptyset$.

A function $P(\cdot)$ that assigns a real number $P(x)$ to each event x is a *probability distribution* if it satisfies three simple axioms:[2] Let $E_1, E_2, \ldots$ be disjoint events. Then

$$P(E) \geq 0 \text{ for every} E\,, \qquad P(\Omega) = 1\,, \qquad P\left(\cup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)\,. \tag{A.127}$$

[2] One person's theorem can be another's axiom. An interesting reformulation of probability theory starts from axioms about expectation values and then derives the rules of probability as *theorems* (Whittle, 2000).

Below, we distinguish cases where Ω is a finite or countable set from cases where it is not countable.

A.6.1 Interpreting Probability

Although the laws of probability follow from a few simple axioms, we also need to *interpret* probability – to assign numerical values to probabilities. While there are at least half a dozen ways to interpret probability theory (Zalta, 2012), the two most popular are probabilities as *relative frequencies* and as *degrees of rational belief*.

Frequentist Interpretation

Much of probability's roots lie in sixteenth- and seventeenth-century efforts to understand gambling odds – how much to bet? Major "players" include Gerolamo Cardano, Pierre de Fermat, Blaise Pascal, and Christiaan Huygens. In this "classical" development of probability, the probability of an event was intimately linked to how frequently it occurs. In the nineteenth and twentieth centuries, mathematicians went further and tried to *define* probability via frequency. These developments are associated with John Venn, Karl Pearson, Jerzy Neyman, and Sir Ronald Fisher. To be more precise, we can define probability as a limiting relative frequency. For simplicity, we consider only the case where each outcome $\boldsymbol{x}$ is an "elementary" event $E = \boldsymbol{x}$. We write $\boldsymbol{x}$ as a vector to allow for outcomes characterized by multiple numbers. Given n trials, we assign the number $P(\boldsymbol{x})$ as, loosely, the number of times $\boldsymbol{x}$ is drawn (number of "successes") divided by the number of draws, or *trials* N_t:

$$P(\boldsymbol{x}) \equiv \lim_{N_t \to \infty} \frac{(\text{Number of trials resulting in event } \boldsymbol{x})}{N_t} . \tag{A.128}$$

For example, in flipping a coin, there are two events (outcomes), $x = H$ and $x = T$. We then assign

$$P(H) = \lim_{N_t \to \infty} \frac{(\text{Number of heads observed in } N_t \text{ coin flips})}{N_t} . \tag{A.129}$$

Implicit in this notion is that multiple trials are possible. While such an assumption works well for coin flips, it is less easily adapted to phenomena that are hard, or impossible, to repeat. The frequentist interpretation may not apply as widely as desired.

Subjective (Bayesian) Interpretation

This alternative view of probability has eighteenth-century origins in the work of Jakob Bernoulli, the Reverend Thomas Bayes, and Pierre Simon de Laplace. Eclipsed by the frequentist interpretation during most of the twentieth century, it was repeatedly rediscovered, often by people outside of statistics with a tough problem to solve

that frequentist statistics could not handle. It was the theory that "would not die" (McGrayne, 2011). Alan Turing famously created Bayesian algorithms for decoding German submarine communications during World War II, although the methods were classified until the early 1970s. In the more open literature, the efforts of Harold Jeffreys and, in physics, E. T. Jaynes were important. In many fields (such as machine learning, neuroscience, medical diagnosis, cosmology), the Bayesian interpretation now dominates. In the Bayesian interpretation of probability, we define

$P(\boldsymbol{x}|I)$ = The degree of rational belief that $\boldsymbol{x}$ is true *given* background information I.

In other words, if you assign $P(\boldsymbol{x}) \approx 1$, then you are confident that $\boldsymbol{x}$ is true, while $P(\boldsymbol{x}) \approx 0$ implies the converse. The probabilities are *subjective* in the sense that they depend on background information I, which may differ for two different people. The probabilities are *objective* in the sense that the rules of probability, developed below, will oblige two people with identical background information to arrive at the same value for $P(\boldsymbol{x})$. This is what we mean by *rational belief*. In many cases, the number assigned to $P(\boldsymbol{x})$ in the Bayesian approach will equal that assigned in the frequentist approach. But a Bayesian approach can lead to probability assignments in cases where the frequentist approach does not apply. Another, perhaps less-loaded term for $P(\boldsymbol{x})$ is *state-of-knowledge*: a probability distribution summarizes all that we know about the quantity $\boldsymbol{x}$.

Example A.16 Two balls in a bag. Here is a quick example to motivate "subjective" probability: You have a bag with two balls – one white, one black.

If you pick one, $P(W|I) = \frac{1}{2}$. Here, I = "there are 2 balls; 1 white, 1 black." Next, you pick one ball but do not look at it and then pick the second ball and see that it is black. Now, $P(W|I') = 1$, where I' adds to I that you picked a ball without looking and then another that was black. Thus, the assignment of P depends on an event that happened *after* the draw, showing that the timing of events is not necessarily important. Probability considers the logical integration of information, not just the temporal sequence.

A.6.2 Probability with Discrete Event Sets

We first develop the rules of probability for the case where the event set X is countable. From the axioms given in Eq. (A.127), one can derive many theorems, such as

$$P(\boldsymbol{x}|I) + P(\overline{\boldsymbol{x}}|I) = 1\,. \tag{A.130}$$

Here, $\overline{\boldsymbol{x}}$ means *not* $\boldsymbol{x}$, the set of outcomes not contained within $\boldsymbol{x}$ (e.g., its complement). For example, in a three-element, single-component event space with $X = \{x_1, x_2, x_3\}$, the complement $\overline{x}_1 = \{x_2, x_3\}$. The vertical bar | means *given*, or *conditional upon*. In

words, Eq. (A.130) states that the probability that $\boldsymbol{x}$ occurs or does not occur is one. To prove this statement, note that $\Omega = \boldsymbol{x} \cup \overline{\boldsymbol{x}}$ and that $\boldsymbol{x} \cap \overline{\boldsymbol{x}} = \emptyset$. Then, applying the second and third probability axioms, $1 = P(\Omega) = P(\boldsymbol{x} \cup \overline{\boldsymbol{x}}) = P(\boldsymbol{x}) + P(\overline{\boldsymbol{x}})$, which is Eq. (A.130).

Above, in the spirit of the Bayesian interpretation of probability, we explicitly condition all probabilities on the background information I, as there is not a unique assignment of probability: What you know, what you assume as background information I, affects your degree of belief about events.[3] For example, if we know that a coin has two heads, then we expect $P(H|I_1) = 1$. If we know nothing about the coin, then we should assume $P(H|I_2) = \frac{1}{2}$. (Indeed, assigning a probability for heads that differs from $\frac{1}{2}$ is rational only when something particular is known about the coin.) Thus, in this case,

- I_1 = "I know this coin is a fake."
- I_2 = "I just saw this coin for the first time."

Although we will often write $P(\boldsymbol{x}|I)$ as $P(\boldsymbol{x})$, we must remember that the assignment of probabilities is always conditioned on outside information.

Applying Eq. (A.130) to an n-element event set, we can iterate and derive that if $P(x_i) = p_i$, then there is a normalization condition that

$$\sum_{i=1}^{n} p_i = 1\,, \tag{A.131}$$

Next, we define the notion of *conditional probability* $P(\boldsymbol{x}|\boldsymbol{y})$:

$$P(\boldsymbol{x},\boldsymbol{y}) = P(\boldsymbol{x}|\boldsymbol{y})P(\boldsymbol{y}) \quad \Longrightarrow \quad P(\boldsymbol{x}|\boldsymbol{y}) = \frac{P(\boldsymbol{x},\boldsymbol{y})}{P(\boldsymbol{y})}\,. \tag{A.132}$$

Loosely, $P(\boldsymbol{x}|\boldsymbol{y})$ is the probability that event $\boldsymbol{x}$ occurs given that event $\boldsymbol{y}$ has occurred. Note that the temporal sequence is for intuition only. The statement $\boldsymbol{y}$ can be thought of as a condition rather than an occurrence: if $\boldsymbol{y}$ holds, then what is the probability of $\boldsymbol{x}$?

Using the notion of conditional probability, we can derive two essential theorems:

1. *Bayes' theorem* [4]

$$P(\boldsymbol{x}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{x})\,P(\boldsymbol{x})}{P(\boldsymbol{y})}\,. \tag{A.133}$$

[3] Private insurers are big-time Bayesians. The more they know about you, the better they can calculate the conditional probabilities and risks. Take health insurance: your age, recreational habits, genetic makeup, family history, state of mind – anything – can allow a better estimate of the probability of your getting sick and hence a better estimate of the amount of money they stand to gain or lose from you. In this and many other fields, "Better Bayesians make bigger bucks." Balancing privacy against efficiency is delicate.

[4] Is it Bayes' theorem or Bayes's Theorem? Traditional English favors the latter, but modern usage tends to the former (e.g., McGrayne, 2011). Although I am sympathetic with sticklers, I would argue that since people tend to pronounce only a single s, we should follow our instincts and drop the possessive s. On the other hand, when we actually pronounce the s twice – think of Gauss's Law – follow the sticklers.

Proof:

$$\begin{aligned} P(\boldsymbol{x},\boldsymbol{y}) &= P(\boldsymbol{x}|\boldsymbol{y})\,P(\boldsymbol{y}) \\ P(\boldsymbol{y},\boldsymbol{x}) &= P(\boldsymbol{y}|\boldsymbol{x})\,P(\boldsymbol{x}) \\ \text{but}\quad P(\boldsymbol{x},\boldsymbol{y}) &= P(\boldsymbol{y},\boldsymbol{x}) \qquad \Longrightarrow \qquad \text{Bayes}\,. \end{aligned}$$

2. *Marginalization*

$$P(\boldsymbol{x}) = \sum_i P(\boldsymbol{x},\boldsymbol{y}_i)\,. \tag{A.134}$$

Sketch of proof: To start, assume $\boldsymbol{Y}$ has but two states, $\boldsymbol{y}$ and $\bar{\boldsymbol{y}}$.

$$\begin{aligned} P(\boldsymbol{x},\boldsymbol{y}) &= P(\boldsymbol{y},\boldsymbol{x}) = P(\boldsymbol{y}|\boldsymbol{x})\,P(\boldsymbol{x}) \\ P(\boldsymbol{x},\bar{\boldsymbol{y}}) &= P(\bar{\boldsymbol{y}},\boldsymbol{x}) = P(\bar{\boldsymbol{y}}|\boldsymbol{x})\,P(\boldsymbol{x}) \\ \text{Add: } P(\boldsymbol{x},\boldsymbol{y}) + P(\boldsymbol{x},\bar{\boldsymbol{y}}) &= \underbrace{[P(\boldsymbol{y}|\boldsymbol{x}) + P(\bar{\boldsymbol{y}}|\boldsymbol{x})]}_{1}\,P(\boldsymbol{x}) = P(\boldsymbol{x})\,. \end{aligned}$$

Thus, in this case,

$$P(\boldsymbol{x}) = P(\boldsymbol{x},\boldsymbol{y}) + P(\boldsymbol{x},\bar{\boldsymbol{y}})\,.$$

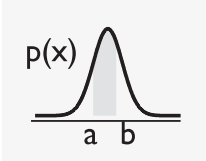
The argument is similar when $\boldsymbol{Y}$ has a countable number of elements. Marginalization can also be applied to elements of a vector $\boldsymbol{x} = \{x^{(1)},\cdots,x^{(N)}\}$. For example,

$$\sum_{x_i^{(1)}} P(\{x^{(1)},\cdots,x^{(N)}\}) = P(\{x^{(2)},\cdots,x^{(N)}\})\,, \tag{A.135}$$

where the sum is over all possible (countable) values of $x^{(1)}$. Any other component $x^{(i)}$ (or several) can equally well be chosen.

A.6.3 Probability with Continuous Event Sets

We can generalize the formalism of probability to handle a continuous set of outcomes, which corresponds to an uncountable event set X (or $\boldsymbol{X}$, if each event $\boldsymbol{x}$ is represented by an N-component vector). The probability to find a value x that is between a and b is given by an integral illustrated by the shaded value at right:

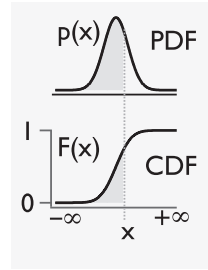


$$P(a \le x \le b) \equiv \int_a^b \mathrm{d}x\, p(x)\,, \tag{A.136}$$

where $p(x)$ is a *probability density function* (PDF). Since the probability to find x somewhere (that is, for the random variable x to take on a value between $-\infty$ and $+\infty$) should be one, we *normalize* $p(x)$ by requiring

$$\int_{-\infty}^{+\infty} \mathrm{d}x\, p(x) = 1\,. \tag{A.137}$$

Notice that whereas probabilities are numbers and have no dimensions, probability distribution functions have units that are inverse to the units of x. For example, if x is a length, $p(x)$ has units of length^{-1}. A second comment is that we will also be somewhat lazy in our language and refer to probability density functions as *continuous distributions*, or even as just *distributions*.



The *cumulative distribution function* (CDF) is defined to be $P(-\infty \le x)$, or

$$F(x) \equiv P(-\infty \le x) = \int_{-\infty}^{x} \mathrm{d}x'\, p(x')\,. \tag{A.138}$$

In words, $F(x)$ is the probability that an element drawn from $p(\cdot)$ has value $\le x$. Clearly, $F(-\infty) = 0$ and $F(\infty)$=1. Equivalently, $p(x) = \frac{\mathrm{d}}{\mathrm{d}x}F(x)$. See left for an illustration.

When $\boldsymbol{x}$ is N-dimensional, we define

$$P(\boldsymbol{x} \in \mathcal{V}) = \int_{\mathcal{V}} \mathrm{d}\boldsymbol{x}\, p(\boldsymbol{x})\,, \tag{A.139}$$

where $\mathcal{V}$ is a subvolume in the n-dimensional event space $\boldsymbol{X}$. Normalization implies $\int_{\boldsymbol{x}\in\boldsymbol{X}} \mathrm{d}\boldsymbol{x}\, p(\boldsymbol{x}) = 1$, which can be written more explicitly as

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathrm{d}x_1 \cdots \mathrm{d}x_n\, p(x_1, \cdots, x_n) = 1\,. \tag{A.140}$$

For simplicity, we now denote the components with a subscript, rather than the superscript we used in the discrete case, where the subscript was used to denote the realization, as in $x_j^{(i)}$, which refers to the ith component of $\boldsymbol{x}$ taking the jth value. For continuous probabilities, Bayes' theorem takes the same form as in Eq. (A.133). Marginalization does too, once we let sums go to the appropriate integral. For example, the *marginal distribution*

$$p(x_2, \cdots, x_n) = \int_{-\infty}^{\infty} \mathrm{d}x_1\, p(x_1, x_2, \cdots, x_n)\,. \tag{A.141}$$

Similarly,

$$p(\boldsymbol{x}) = \int_{\boldsymbol{y}\in\boldsymbol{Y}} \mathrm{d}\boldsymbol{y}\, p(\boldsymbol{x}, \boldsymbol{y})\,. \tag{A.142}$$

Marginalization is also referred to, in physics, as *integrating out* undesired variables.

We can unify the presentations of continuous and discrete event sets through use of the delta function. Thus, we can write the probability distribution function of a variable x that can take on values $\{x_1, x_2, \ldots, x_n\}$ with probabilities $\{p_1, p_2, \ldots, p_n\}$ as

$$p(x) = \sum_{i=1}^{n} p_i\, \delta\,(x - x_i)\,. \tag{A.143}$$

As an example of how this definition leads to the familiar rules for manipulating discrete probabilities, the normalization condition for Eq. (A.143) is

$$\int_{-\infty}^{\infty} \mathrm{d}x\, p(x) = \int_{-\infty}^{\infty} \mathrm{d}x \sum_{i=1}^{n} p_i\, \delta\,(x - x_i) = \sum_{i=1}^{n} p_i = 1\,, \tag{A.144}$$

which is just Eq. (A.131). The corresponding discrete distribution is $P(x_i) = p_i$.

We can extend these ideas to include mixed cases, neither discrete nor continuous. For example, $p(x) = \frac{1}{2}[\delta\,(x) + \mathrm{e}^{-x^2/2} \,/\, \sqrt{2\pi}]$ describes a variable that has a 50% chance of having a value of 0 and a 50% chance of being distributed as a Gaussian with mean 0 and variance 1. An even more general formulation of probability is based on *measure theory*.

A.6.4 Expected Value of Functions of Random Variables

We define the *expected value* of a function $\boldsymbol{f}(\boldsymbol{x})$ to be

$$\langle \boldsymbol{f}(\boldsymbol{x}) \rangle \equiv \underbrace{\sum_{\boldsymbol{x}\in\boldsymbol{X}} P(\boldsymbol{x})\, \boldsymbol{f}(\boldsymbol{x})}_{\text{discrete case}} \quad \text{or} \quad \underbrace{\int_{\boldsymbol{x}\in\boldsymbol{X}} \mathrm{d}\boldsymbol{x}\, p(\boldsymbol{x})\, \boldsymbol{f}(\boldsymbol{x})}_{\text{continuous case}}\,, \tag{A.145}$$

where $\boldsymbol{f}(\cdot)$ is an m-dimensional function defined on the n-dimensional random variable $\boldsymbol{x}$. As is traditional in quantum and statistical physics, the angle-bracket notation, $\langle\cdot\rangle$, does not explicitly state which distribution is being integrated over. In most cases, the context will make the distribution clear. If not, we will specify it explicitly. We distinguish the probabilistic notion of expected value from the statistical notion of *average*,

$$\overline{\boldsymbol{f}(\boldsymbol{x})} \equiv \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{f}(\boldsymbol{x}_i)\,, \qquad \boldsymbol{x}_i \sim p(\boldsymbol{x})\,. \tag{A.146}$$

The notation $\boldsymbol{x}_i \sim p(\boldsymbol{x})$ means that $\boldsymbol{x}_i$ is sampled from the distribution $p(\boldsymbol{x})$. The *law of large numbers* asserts that $\lim_{N\to\infty} \overline{\boldsymbol{f}(\boldsymbol{x})} = \langle \boldsymbol{f}(\boldsymbol{x}) \rangle$.

Below, we consider important special cases of functions $\boldsymbol{f}(\cdot)$. For brevity, we discuss only the continuous case. The discrete case has analogous formulas, substituting sums for integrals and distributions for densities.

Mean and Variance of a Probability Distribution

Although a full description of a continuous random variable requires knowing its distribution $p(\boldsymbol{x})$, reduced descriptions are very useful summaries. We begin by defining the *mean*, a measure of the "typical" value of $\boldsymbol{x}$ and the *variance*, a measure of the *width* of $p(\boldsymbol{x})$. We start with one variable and then generalize to n variables.

- The mean, $\mu \equiv \langle x \rangle$, of a distribution is defined as

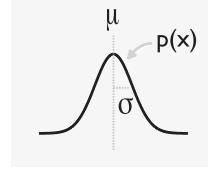
$$\mu \equiv \langle x \rangle \equiv \int_{-\infty}^{\infty} \mathrm{d}x\, p(x)\, x \tag{A.147}$$

and is a common measure of a "typical value" of a distribution. Others include the *mode* (global maximum of a probability distribution) and the *median* (which divides the lower and higher halves of the distribution). The angle brackets $\langle\cdot\rangle$ in Eq. (A.147) denote ensemble averages (over many trials). Ensemble averages can sometimes be calculated by taking a single, long time average (the *ergodic* property). For distributions that have a single, symmetric peak, the mean, mode, and median all coincide.

In our discussions of noise, we always assume a noise term of 0 mean because any constant could be absorbed into the deterministic part of the measurement.

- The *variance* is defined as a second *central moment*, which compensates for the mean:

$$\sigma^2 \equiv \langle\, \delta x^2 \rangle \equiv \mathrm{Var}\; x \equiv \left\langle (x-\mu)^2 \right\rangle = \langle x^2 \rangle - \mu^2 = \int_{-\infty}^{\infty} \mathrm{d}x\, p(x)\,(x-\mu)^2\,, \qquad \text{(A.148)}$$



and is a measure of the typical fluctuation of a variable x about its mean. The *standard deviation*, σ, is the square root of the variance. The mean and standard deviation are illustrated at left for a Gaussian distribution (discussed in Section A.7.3).

- With n-dimensional variables $\boldsymbol{x}$, the mean becomes a vector $\boldsymbol{\mu} = \int \mathrm{d}\boldsymbol{x}\, p(\boldsymbol{x})\, \boldsymbol{x}$, and the variance becomes the *covariance* matrix

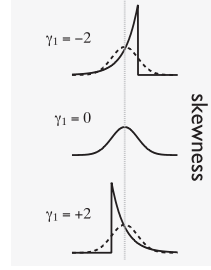
$$\boldsymbol{\Sigma} \equiv \langle\, \delta\boldsymbol{x}\, \delta\boldsymbol{x}^{\mathsf{T}} \rangle \equiv \mathrm{Cov}\; \boldsymbol{x} = \int \mathrm{d}\boldsymbol{x}\, p(\boldsymbol{x})\, (\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\,, \qquad \text{(A.149)}$$

where $\delta\boldsymbol{x} \equiv \boldsymbol{x} - \boldsymbol{\mu}$. The diagonal elements of $\boldsymbol{\Sigma}$ are the variances of the ith component of $\boldsymbol{x}$. The off-diagonal elements, $\langle x_i\, x_j \rangle$, give the tendency for fluctuations in x_i to correlate with those of x_j. Note that although the variances $\langle x_i^2 \rangle \geq 0$, the covariances may be positive or negative. A negative covariance means that a positive fluctuation in x_i is likely to be accompanied by a negative fluctuation in x_j. See Section A.7.5, below, for an example.

Higher-Order Moments

Higher-order moments such as the mth-order moment $\langle x^m \rangle$ and the mth-order central moment $\langle (x-\mu)^m \rangle$, which subtracts off the mean μ, are sometimes useful, too. The *skewness* γ_1 is the third central moment, normalized by the standard deviation σ,

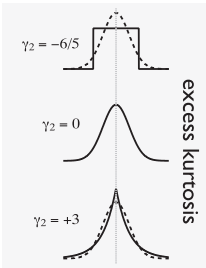
$$\gamma_1 \equiv \left\langle \left(\frac{x-\mu}{\sigma} \right)^3 \right\rangle . \qquad \text{(A.150)}$$



For a *symmetric* distribution, where $p(x) = p(-x)$, the skewness $\gamma_1 = 0$. At left are distributions with negative, zero, and positive skewness.

The *kurtosis* is a normalized fourth central moment, but it is more common to use the *excess kurtosis*, defined as

$$\gamma_2 \equiv \left\langle \left(\frac{x-\mu}{\sigma} \right)^4 \right\rangle - 3\,, \qquad \text{(A.151)}$$



where the factor of −3 is chosen so that a Gaussian distribution has $\gamma_2 = 0$ (Problem A.7.3). Distributions with $\gamma_2 < 0$ thus are "flat topped" relative to a Gaussian, while those with $\gamma_2 > 0$ are "peaky" and have fatter tails, as illustrated at right.

Conditional Moments

We can also define moments of conditional probability distributions, for example the *conditional mean*:[5]

$$\langle x \rangle_y \equiv \int \mathrm{d}x\, p(x|y)\, x\,. \tag{A.152}$$

If x and y are independent variables, the conditional mean reduces to the ordinary mean:

$$\langle x \rangle_y \equiv \int \mathrm{d}x\, p(x|y)\, x = \int \mathrm{d}x\, \frac{p(x,y)}{p(y)}\, x = \int \mathrm{d}x\, \frac{p(x)\,\cancel{p(y)}}{\cancel{p(y)}}\, x = \langle x \rangle\,. \tag{A.153}$$

Loosely, the conditional mean is the mean of the subset of those values of x that satisfy the condition y. Note that we have assumed $p(y) > 0$. If it equals zero, then y is impossible and, again, $\langle x \rangle_y = \langle x \rangle$.

In Section A.8.7, we show that the conditional mean $\langle x \rangle_y$ of a probability distribution $p(x|y)$ minimizes the mean-square deviation, $\langle (\langle x \rangle - x)^2 \rangle$, where the brackets are averaged over $p(x|y)$. It then serves as a "best" estimate of the quantity x (e.g., an unknown state) given the quantity y (e.g., a measurement).

Characteristic Function and Cumulants

The *characteristic function* is just the complex conjugate of the Fourier transform of a probability distribution.[6]

$$\varphi_x(k) = \left\langle \mathrm{e}^{\mathrm{i}kx} \right\rangle = \int_{-\infty}^{\infty} \mathrm{d}x\, p(x)\ \mathrm{e}^{\mathrm{i}kx} \tag{A.154}$$

We can use characteristic functions to generate the mth moment of a probability distribution function. From Eq. (A.154), we see that

$$\left\langle \mathrm{e}^{\mathrm{i}kx} \right\rangle = \sum_{m=0}^{\infty} \frac{(\mathrm{i}k)^m \langle x^m \rangle}{m!} \qquad \Longrightarrow \qquad \langle x^m \rangle = (-\mathrm{i})^m\, \frac{\mathrm{d}^m}{\mathrm{d}k^m} \varphi_x(k) \bigg|_{k=0}\,. \tag{A.155}$$

[5] An alternate notation, preferred in the statistics and control literature, uses $E(x)$ to define the expected value of x (the mean). Although I prefer the standard physics notation, $\langle x \rangle$, the statistics notation is perhaps clearer for conditional expectations, which are denoted $E(x|y)$, in closer analogy to the notation for conditional probability, $p(x|y)$. Unfortunately, $\langle x|y \rangle$, is used in physics for the inner product between two elements x and y. The notation $\langle x \rangle_y$ is a compromise.

[6] The physics literature often uses $\mathrm{e}^{\mathrm{i}kx}$ for the forward transform rather than the $\mathrm{e}^{-\mathrm{i}kx}$ used here. The characteristic function then equals the Fourier transform (no complex conjugate).

Likewise, the mth-order *cumulant* is defined in terms of the *cumulant generating function* $h_x(k) \equiv \ln \varphi_x(k)$ as

$$h_x(k) = \ln \left\langle e^{ikx} \right\rangle \equiv \sum_{m=0}^{\infty} \frac{(ik)^m \kappa_m(x)}{m!} \qquad \Longrightarrow \qquad \kappa_m(x) = (-i)^m \left. \frac{d^m}{dk^m} \ln \varphi_x(k) \right|_{k=0} . \quad \text{(A.156)}$$

It is easy to see that the first cumulant $\kappa_1(x) = \langle x \rangle = \mu$, the mean. Similarly, the second cumulant $\kappa_2(x) = \langle x^2 \rangle - \langle x \rangle^2 = \sigma^2$, the variance. Like central moments, cumulants are independent of μ. But, unlike moments and central moments, the cumulants of independent random variables add. For two independent random variables x and y,

$$h_{x+y}(k) = \ln \left\langle e^{ik(x+y)} \right\rangle = \ln \left(\left\langle e^{ikx} \right\rangle \left\langle e^{iky} \right\rangle \right) = h_x(k) + h_y(k) , \quad \text{(A.157)}$$

which implies that $\kappa_m(x + y) = \kappa_m(x) + \kappa_m(y)$ for all cumulant orders m. More generally, for N independent random variables, $\kappa_m(\sum_i x_i) = \sum_i \kappa_m(x_i)$.

Since we can inverse Fourier transform to find $p(x)$ from $\varphi_x(k)$, the characteristic function implies the moments and cumulants, which imply the distribution.[7] Thus, we can "expand" a distribution $p(x)$ in terms of its moments or, better, its cumulants. The mean, variance, skewness, and kurtosis are then related to the first four terms of that expansion. Often, we focus on the first two.

A.6.5 Functions of Random Variables

Sometimes, we have several random variables of known distribution, and we want to know the distribution of some function of these variables. We can transform from one probability distribution to another using the rules of probability and multivariable calculus.

As a preliminary step, we prove a delta-function identity. Let $g(x)$ be a function with r simple roots x_i, which satisfy $g(x_i) = 0$ and $g'(x_i) \neq 0$. Then,

$$\delta [g(x)] = \sum_{i=1}^{r} \frac{\delta (x - x_i)}{|g'(x_i)|} \quad \text{(A.158)}$$

To prove Eq. (A.158), assume, first, that $g(x)$ has a single root at $x = x_1$. Let $p(x)$ be an arbitrary test function and define $y = g(x)$. Realizing that the contribution to the delta function comes only when its argument is zero, we expand $g(x) \approx g'(x_1)(x - x_1)$ and write

$$\int_{-\infty}^{\infty} dx\, p(x)\, \delta [g(x)] = \int_{-\infty}^{\infty} dx\, p(x)\, \delta [g'(x_1)(x - x_1)] = \frac{p(x_1)}{|g'(x_1)|} . \quad \text{(A.159)}$$

The last step uses another delta-function identity, $\delta (ax) = \frac{1}{|a|} \delta (x)$, which is left as an exercise for the reader. If there are r simple roots, each contributes a similar term, giving Eq.(A.158). If $g(x)$ has no roots, then $\delta [g(x)] = 0$.

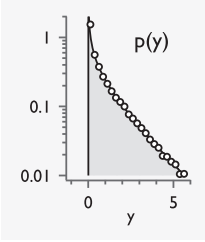
[7] Usually. There are exotic cases where two different distributions share the same moments.

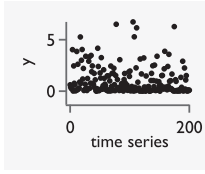
Functions of One Random Variable

Given a random variable x with probability density function $p(x)$, we calculate the density function $p(y)$ of the function $y = f(x)$ as follows:

$$\begin{aligned} p(y) &= \int \mathrm{d}x\, p(y,x) && \text{marginalization} \\ &= \int \mathrm{d}x\, p(y|x)\, p(x) && \text{conditional probability} \\ &= \int \mathrm{d}x\ \delta\left[y - f(x)\right] p(x) && \text{enforce } y = f(x) \\ &= \sum_{i=1}^{r} \frac{p(x_i)}{|f'(x_i)|} && \text{using Eq. (A.158)}\,. \end{aligned} \tag{A.160}$$

In Eq. (A.160), the x_i constitute the r inverses of y, satisfying $x = f^{-1}(y)$, since they are also roots of the function $g(x) = y - f(x) = 0$. If there is a unique inverse, the formula simplifies to $p(y) = p(x)/|f'(x)|$, which can also be justified using the rougher argument $p(y)\,\mathrm{d}y = p(x)\,\mathrm{d}x$ and noting that probability densities always have to be positive.





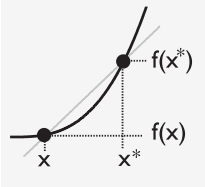
Example A.17 (Quadratic transformation) Given the Gaussian random variable $x \sim \mathcal{N}(0,1)$, what is the distribution of the random variable $y = x^2$?

Given y, there are two inverses: $x_1 = +\sqrt{y}$ and $x_2 = -\sqrt{y}$. Then $f'(x) = \pm 2\sqrt{y}$ and

$$p(y) = \frac{p(+\sqrt{y})}{|2(+\sqrt{y})|} + \frac{p(-\sqrt{y})}{|2(-\sqrt{y})|} = \frac{2}{2\sqrt{y}}\left(\frac{\mathrm{e}^{-\frac{y}{2}}}{\sqrt{2\pi}}\right) = \frac{1}{\sqrt{2\pi y}}\,\mathrm{e}^{-\frac{y}{2}}\,\theta(y)\,, \tag{A.161}$$

where the step function $\theta(y)$ enforces the condition that $y \geq 0$. (See right.) It is straightforward to verify that $\int_0^\infty \mathrm{d}y\, p(y) = 1$. The distribution $p(y)$ is also known as the χ^2 distribution for one degree of freedom. This example, reprised in Section 8.4.2, shows that a nonlinear transformation of a Gaussian-distributed variable is not, in general, Gaussian.

Jensen's Inequality



This useful relation shows up often in proofs, especially for information theory and stochastic thermodynamics. Let $f(x)$ be a *convex* $\smile$ function, as illustrated at right, and let $p(x)$ be a probability distribution. Then

$$f(\langle x \rangle) \leq \langle f(x) \rangle\,. \tag{A.162}$$

To prove this relation, notice that, for two points x and x^* on a convex function f, the slope of the chord is less than the local slope at the endpoint:

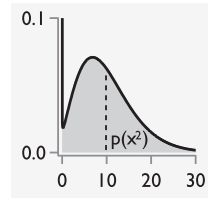
$$\frac{f(x^*) - f(x)}{x^* - x} \leq f'(x^*)\,, \tag{A.163}$$

Let $x^* = \langle x \rangle$. Then

$$f(\langle x \rangle) \leq f'(\langle x \rangle)\,(\langle x \rangle - x) + f(x)\,. \tag{A.164}$$

Taking the expectation of each side and noting that $\langle (\langle x \rangle - x) \rangle = 0$ proves the result. For us, important examples of convex functions are $\log 1/x$ and $x \log x$.

To understand the Jensen inequality more intuitively, apply the function $f(x) = x^2$ to transform the Gaussian distribution $x \sim \mathcal{N}(\mu, 1)$. We plot the distribution of x^2 at left for $\mu = 3$. Extending the results from Example A.17, we can easily verify that $\langle x^2 \rangle = 10$, which is greater than $\mu^2 = 3^2 = 9$. Intuitively, the function "stretches" higher values of x, which increases the mean of the transformed distribution. Note the peak near $x^2 = 0$, which results from "folding over" the two square roots.



Functions of More Than One Random Variable

We can also calculate functions of several random variables. We assume that $x \sim p(x)$ and $y \sim p(y)$ are independent random variables, and we wish to find $p(z)$ for $z = f(x, y)$. Then

$$\begin{aligned}
p(z) &= \iint \mathrm{d}x\,\mathrm{d}y\; p(z, x, y) && \text{marginalization} \\
&= \iint \mathrm{d}x\,\mathrm{d}y\; p(z|x, y)\; p(x, y) && \text{conditional probability} \\
&= \iint \mathrm{d}x\,\mathrm{d}y\; p(z|x, y)\; p(x)\; p(y) && x,\ y \text{ are independent} \\
&= \iint \mathrm{d}x\,\mathrm{d}y\; \delta\left[z - f(x, y)\right] p(x)\; p(y) && \text{enforce } z = f(x, y) \\
&= \sum_{i=1}^{r} \int \mathrm{d}x\; p(x)\, \frac{p(y_i)}{\left|\partial_y f(y_i)\right|} && \text{using Eq. (A.158)}
\end{aligned} \tag{A.165}$$

In Eq. (A.165), $\partial_y f(y_i)$ is short for $\partial f(x, y)/\partial y|_{y=y_i}$, where y_i is again the ith branch of the inverse of $z = f(x, y)$.

Example A.18 (Sum of two random variables) Here $z = f(x, y) = x + y$ and, thus, $|\partial_y f(x, y)| = 1$. The inverse is unique and given by $y = z - x$. Then,

$$p(z) = \int_{-\infty}^{\infty} \mathrm{d}x\; p(x)\; p(z - x) = p(x) * p(y)\,. \tag{A.166}$$

Thus, $p(z)$ is just the convolution of the distributions for x and y.

The characteristic function of the transformed probability distribution has a particularly simple form. Let $\boldsymbol{x}$ be a stochastic vector and $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ be a smooth function. Starting from

$$p(\boldsymbol{y}) = \int \mathrm{d}\boldsymbol{x}\ \delta\left[\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x})\right] p(\boldsymbol{x})\,, \tag{A.167}$$

we have

$$\varphi_{\boldsymbol{y}}(\boldsymbol{k}) = \int \mathrm{d}\boldsymbol{y}\, p(\boldsymbol{y})\ \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{y}} = \int \mathrm{d}\boldsymbol{x}\, \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{f}(\boldsymbol{x})}\ p(\boldsymbol{x}) = \left\langle \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{f}(\boldsymbol{x})} \right\rangle\,. \tag{A.168}$$

This formula works even when $\boldsymbol{x}$ and $\boldsymbol{y}$ have different numbers of components. For example, $y = x_1 + x_2$ implies $\varphi_y = \varphi_{x_1}\,\varphi_{x_2}$ and $p(y) = p(x_1) * p(x_2)$, as seen in Eq. (A.166). Transforming back from an explicit expression for a characteristic function often leads to contour integrals (Problem A.6.3).

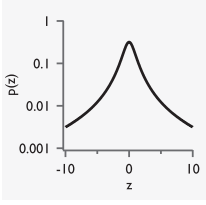
A.6.1 **Chain rule for probabilities**. Let $X^N = \{X_N, X_{N-1}, \ldots, X_1\}$. Show that

$$P(X^N) = \prod_{k=1}^{N} P(X_k | X^{k-1})\,.$$

A.6.2 **Moment exercises**. For the following probability density functions, find their mean μ, variance σ^2, skewness γ_1, and excess kurtosis γ_2.

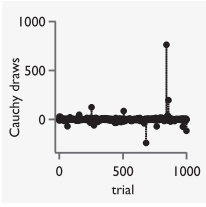
a *Uniform*: $p(x) = \frac{1}{2\sqrt{3}}$ for $-\sqrt{3} \le x \le \sqrt{3}$.
b *Exponential*: $p(x) = \mathrm{e}^{-x}$, for $x \ge 0$.
c *Laplace*: $p(x) = \frac{1}{\sqrt{2}}\,\mathrm{e}^{-\sqrt{2}|x|}$.

A.6.3 **Example A.17, again**. Find $p(y)$ for $y = x^2$ and $x \sim \mathcal{N}(0,1)$ via the characteristic function $\varphi_y(x)$. Transform back to $p(y)$ using contour integration and a branch cut.



A.6.4 **Ratio of normal variables is Cauchy.** Let x and y be independently distributed as $\mathcal{N}(0,1)$. Show that their ratio is distributed as a *Cauchy* (or *Lorentzian*) distribution. That is, if $z = y/x$, show that

$$p(z) = \left(\frac{1}{\pi}\right) \frac{1}{1+z^2}\,. \tag{A.169}$$



The Cauchy distribution, illustrated at right, is the "bad boy" of probability distributions. Its samples fluctuate so wildly (whenever x is near zero) that both mean and variance diverge and are undefined, although median and quartiles are. We see here that simple mechanisms involving standard random quantities can lead to such pathologies. The asymptotic power-law behavior $\sim z^{-2}$ is known as a *fat tail*.

A.7 Gaussian Integrals and Distributions

As we will often need to do integrals involving Gaussian functions, we review the required techniques here. Then we immediately apply those results to the discussion of Gaussian probability distributions, which play a key role in the analysis of the noisy linear systems discussed in Chapter 8.

A.7.1 Gaussian Integrals in One Variable

The basic Gaussian integral is

$$I = \int_{-\infty}^{\infty} \mathrm{d}x\, \mathrm{e}^{-\frac{1}{2}ax^2} \,. \tag{A.170}$$

To evaluate I, we write down two copies, in x and y,

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} \mathrm{d}x\, \mathrm{e}^{-\frac{1}{2}ax^2} \int_{-\infty}^{\infty} \mathrm{d}y\, \mathrm{e}^{-\frac{1}{2}ay^2} \\ &= \iint \mathrm{d}x\, \mathrm{d}y\, \mathrm{e}^{-\frac{1}{2}a(x^2+y^2)} \qquad\qquad \text{define } x = r\cos\theta,\ y = r\sin\theta \\ &= \int_0^{2\pi} \mathrm{d}\theta \int_0^{\infty} \mathrm{d}r\, r\ \mathrm{e}^{-\frac{1}{2}ar^2} = 2\pi \int_0^{\infty} \mathrm{d}u\, \mathrm{e}^{-au} = \frac{2\pi}{a} \,. \end{aligned} \tag{A.171}$$

Thus, $I = \sqrt{2\pi}/a$. Note that we could reduce this calculation to one where $a = 1$ by first changing variables to $x' = x/\sqrt{a}$.

We often need to calculate moments of the form $\int_{-\infty}^{\infty} \mathrm{d}x\, x^k\ \mathrm{e}^{-\frac{1}{2}ax^2}$. Since $\exp(-\frac{1}{2}ax^2)$ is even, the odd moments (k odd) vanish by symmetry. We can calculate the second moment with another trick:

$$\begin{aligned} I &= \int_{-\infty}^{\infty} \mathrm{d}x\, x^2\ \mathrm{e}^{-\frac{1}{2}ax^2} = \int_{-\infty}^{\infty} \mathrm{d}x \left(-2\frac{\partial}{\partial a}\right) \mathrm{e}^{-\frac{1}{2}ax^2} = -2\frac{\partial}{\partial a} \int_{-\infty}^{\infty} \mathrm{d}x\, \mathrm{e}^{-\frac{1}{2}ax^2} \\ &= -2\frac{\partial}{\partial a} \sqrt{\frac{2\pi}{a}} = \frac{\sqrt{2\pi}}{a^{3/2}} \,. \end{aligned} \tag{A.172}$$

Further derivatives with respect to a give higher even moments.

A.7.2 Gaussian Integrals in n Variables

To extend the results of Section A.7.1 to n variables, we look at the n-dimensional integral of the *multivariate Gaussian*

$$I = \int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{x}\, \mathrm{e}^{-\frac{1}{2}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x}} \,, \tag{A.173}$$

where we will assume the matrix $\boldsymbol{A}$ to be real, symmetric, and positive definite (see Section A.1.1). We substitute $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{z}$, where $\boldsymbol{U}$ is the unitary transformation that diagonalizes $\boldsymbol{A}$ and satisfies $\boldsymbol{U}\,\boldsymbol{U}^{\mathsf{T}} = \mathbb{I}$. That is $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\mathsf{T}}$, or $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{U} = \boldsymbol{D}$. Then,

$$\begin{aligned} I &= \int d\boldsymbol{z}\ \mathrm{e}^{-\frac{1}{2}\boldsymbol{z}^{\mathsf{T}}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{U}\boldsymbol{z}} = \int d\boldsymbol{z}\ \mathrm{e}^{-\frac{1}{2}\boldsymbol{z}^{\mathsf{T}}\boldsymbol{D}\boldsymbol{z}} = \prod_{k=1}^{n} \int_{-\infty}^{\infty} \mathrm{d}z\, \mathrm{e}^{-\frac{1}{2}\lambda_k z^2} = \prod_{k=1}^{n} \sqrt{\frac{2\pi}{\lambda_k}} \\ &= \frac{(2\pi)^{n/2}}{\sqrt{\det \boldsymbol{A}}} \,. \end{aligned} \tag{A.174}$$

Since the columns of $\boldsymbol{U}$ are orthonormal vectors, the Jacobian of the transformation $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{z}$ is one, and $d\boldsymbol{x} = d\boldsymbol{z}$.

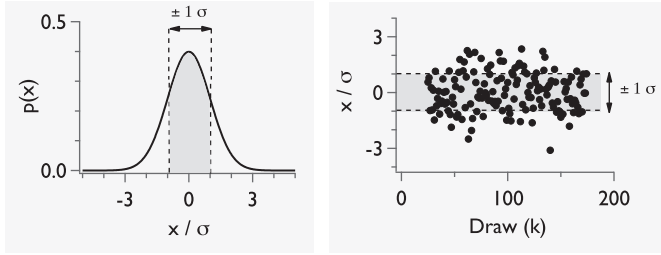


Fig. A.5 Gaussian distribution $\mathcal{N}(0, 1)$. Left: The probability for a point to fall within $\pm 1\,\sigma$ is $\approx \frac{2}{3}$. Right: 150 draws from the distribution on the left. Roughly $\frac{2}{3}$ of the draws fall within the shaded region.

The moments of a multivariate Gaussian function can be found by tricks similar to the one-dimensional case. For example, for a symmetric matrix $\boldsymbol{A}$,

$$\int \mathrm{d}\boldsymbol{x}\left(x_i x_j\right) \mathrm{e}^{-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}} = \lim_{\boldsymbol{b}\to 0} \frac{\partial^2}{\partial b_i\, \partial b_j} \int \mathrm{d}\boldsymbol{x}\, \mathrm{e}^{-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}\cdot\boldsymbol{x}} \,. \tag{A.175}$$

We evaluate the integral in Eq. (A.175) by completing the square:

$$\int \mathrm{d}\boldsymbol{x}\, \mathrm{e}^{-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}\cdot\boldsymbol{x}} = \int \mathrm{d}\boldsymbol{x}\, \mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{A}^{-1}\boldsymbol{b})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{A}^{-1}\boldsymbol{b})}\, \mathrm{e}^{\frac{1}{2}\boldsymbol{b}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{b}} = \left(\frac{(2\pi)^{n/2}}{\sqrt{\det \boldsymbol{A}}}\right) \mathrm{e}^{\frac{1}{2}\boldsymbol{b}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{b}} \tag{A.176}$$

Taking the derivative $\frac{\partial}{\partial b_j}$, recalling that $\boldsymbol{A}^{-1}$ is symmetric, and using the summation convention for repeated indices then gives

$$\frac{\partial}{\partial b_j} \exp\left[\tfrac{1}{2} b_k (\boldsymbol{A}^{-1})_{k\ell} b_\ell\right] = \tfrac{1}{2}[2 b_k (\boldsymbol{A}^{-1})_{kj}] \exp[\cdot] = (\boldsymbol{A}^{-1}\boldsymbol{b})_j] \exp[\cdot] \,. \tag{A.177}$$

Taking another derivative, with respect to b_i, gives

$$[(\boldsymbol{A}^{-1})_{ij} + (\boldsymbol{A}^{-1}\boldsymbol{b})_i\, (\boldsymbol{A}^{-1}\boldsymbol{b})_j] \exp\left[\tfrac{1}{2} b_i (\boldsymbol{A}^{-1})_{ij} b_j\right] \xrightarrow[\boldsymbol{b}\to 0]{} (\boldsymbol{A}^{-1})_{ij} \,. \tag{A.178}$$

Putting the pieces together leads to the final result,

$$\int \mathrm{d}\boldsymbol{x}\left(x_i x_j\right) \mathrm{e}^{-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}} = \frac{(2\pi)^{n/2}}{\sqrt{\det \boldsymbol{A}}}\, (\boldsymbol{A}^{-1})_{ij} \,. \tag{A.179}$$

A.7.3 Gaussian Distributions in One Variable

Gaussian probability distributions, the familiar "bell-shaped curve," play a key role in the analysis of stochastic linear systems. Let us define

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] , \tag{A.180}$$

Using the integrals derived in the previous section, we see that

$$\int_{-\infty}^{\infty} \mathrm{d}x\, p(x) = 1\,, \qquad \langle x\rangle \equiv \int_{-\infty}^{\infty} \mathrm{d}x\, p(x)\, x = \mu$$
$$\langle x^2\rangle \equiv \int_{-\infty}^{\infty} \mathrm{d}x\, p(x)\, x^2 = \sigma^2 + \mu^2 \qquad \Longrightarrow \qquad \mathrm{Var}\; x = \sigma^2\,. \tag{A.181}$$

We confirm these identities in Problem A.7.3.

Here are some more of the many useful features of a Gaussian:

- It has two (and only two) parameters, the mean μ and variance σ^2. The mean μ controls the location of the distribution, while the variance σ^2 controls its lateral scale.
- Because Gaussian distributions are so common, there is a special notation to denote that a random variable, say x, is distributed according to a Gaussian distribution of mean μ and variance σ^2.[8] We write that $x \sim \mathcal{N}(\mu, \sigma^2)$. The "$\sim$" denotes "distributed as." The $\mathcal{N}$ notation refers to an alternate name for a Gaussian, the *normal distribution*. Occasionally, we will need to be more explicit about the variable, in which case we extend the notation to read $\mathcal{N}(x; \mu, \sigma)$.
- Defining $z \equiv (x-\mu)/\sigma$, the universal Gaussian distribution $\mathcal{N}(0,1)$ is simply

$$p(z) = \frac{1}{\sqrt{2\pi}}\, \exp\left(-\tfrac{1}{2}z^2\right)\,. \tag{A.182}$$

  Notice that transforming dx to dz "absorbs" the σ in the prefactor.
- The *characteristic function* is

$$\varphi_x(k) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}x\, \mathrm{e}^{\mathrm{i}kx}\, \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \exp\left(\mathrm{i}k\mu - \tfrac{1}{2}\sigma^2k^2\right)\,, \tag{A.183}$$

  whose first and second moments are

$$\langle x\rangle = -\,\mathrm{i}\,\mathrm{d}_k \exp\left(\mathrm{i}k\mu - \tfrac{1}{2}\sigma^2k^2\right)|_{k=0} = \mu \tag{A.184}$$
$$\langle x^2\rangle = -\mathrm{d}_{kk} \exp\left(\mathrm{i}k\mu - \tfrac{1}{2}\sigma^2k^2\right)|_{k=0} = \mu^2 + \sigma^2\,. \tag{A.185}$$

- The *cumulant generating function*, $\ln\varphi_x(k)$, is

$$h_x(k) = \mathrm{i}k\mu - \tfrac{1}{2}\sigma^2k^2\,, \tag{A.186}$$

  which immediately gives $\kappa_1 = \mu$ and $\kappa_2 = \sigma^2$. The higher-order cumulants are zero.
- The *sum of two Gaussian variables is Gaussian*. Let x and y be two independent Gaussian random variables, with $x \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $y \sim \mathcal{N}(\mu_y, \sigma_y^2)$. Then Eq. (A.168) implies that

$$\varphi_z(k) = \varphi_x(k)\,\varphi_y(k) = \mathrm{e}^{\mathrm{i}k\mu_x - \frac{1}{2}\sigma_x^2k^2}\; \mathrm{e}^{\mathrm{i}k\mu_y - \frac{1}{2}\sigma_y^2k^2} = \mathrm{e}^{\mathrm{i}k(\mu_x+\mu_y) - \frac{1}{2}(\sigma_x^2+\sigma_y^2)k^2}\,. \tag{A.187}$$

  Taking the inverse Fourier transform of $\varphi_z(-k)$ gives the desired result.
- *Rescaling* gives $z = Ax \sim \mathcal{N}(A\mu_x, A^2\sigma_x^2)$.

[8] Some authors use $\mathcal{N}(\mu, \sigma)$ to indicate a Gaussian (normal) distribution of mean μ and standard deviation σ.

- *Central-limit theorem* (CLT): The average of n independent random variables tends to a Gaussian for large n. More precisely, for independent random variables with mean μ and variance σ^2, the distribution for nth approximation to the average, $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, converges to a Gaussian for $n \to \infty$:

$$p(\bar{x}_n) \to \mathcal{N}(\mu, \sigma^2/n)\,. \tag{A.188}$$

  Refinements to the basic theorem can relax the requirements for identically distributed variables. Density functions that lack a finite variance, such as the Cauchy distribution, do not obey the CLT. For such distributions, estimators of the average do *not* converge, no matter how many terms are used.

Problem A.7.1 Show that $\int_{-\infty}^{\infty} \mathrm{d}x \exp\left[-\frac{1}{2}(ax^2 + bx)\right] = \sqrt{\frac{2\pi}{a}}\, \mathrm{e}^{b^2/8a}$. (Complete the square.)

Problem A.7.2 **Characteristic function of a Gaussian.** Derive the characteristic function of a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. The result is quoted in Eq. (A.183).

Problem A.7.3 **Higher moments of a Gaussian.** For a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, use the characteristic function to calculate the first four moments, $\langle x^n \rangle$ ($n = 1, 2, 3, 4$). Verify that the skewness γ_1 and kurtosis γ_2 both vanish.

Problem A.7.4 **Central-limit theorem (CLT), via cumulants**. Consider N independent, identically distributed (i.i.d.) variables, each with mean zero and variance σ^2.

a. *Homogeneity*: Show that the cumulant $\kappa_m(\lambda x) = \lambda^m \kappa_m(x)$, where $\lambda > 0$.
b. Using additivity and homogeneity, find $\kappa_m(z_N)$ for $z_N \equiv \sum_{i=1}^{N}(x_i/\sqrt{N\sigma^2})$.
c. Argue that, for $N \to \infty$, the only nonzero cumulant is $m = 2$.
d. Conclude that $\lim_{N\to\infty} p(z_N) \sim \mathcal{N}(0, 1)$.
e. Define $\bar{x}_N = \frac{1}{N}\sum_i x_i$. Find $\lim_{N\to\infty} p(\bar{x}_N)$.

This is the essence of the CLT proof and can be generalized to the case where the x_i all have different distributions, each with its own mean μ_i and variance σ_i^2.

Problem A.7.5 **Multiple measurements lead to Gaussian states-of-knowledge**. The Central Limit Theorem also explains why the state-of-knowledge for a quantity x tends to be Gaussian after many independent measurements are made.

a. Use Bayes' theorem, a uniform prior, the relations between characteristic functions and repeated convolution and the CLT, and that a Fourier transform of a Gaussian is also Gaussian to argue this point. See Jacobs (2014), Section 1.2.2.
b. Explain why this claim is true but rather trivial when the individual measurements have Gaussian errors.
c. Explain why this claim is *not* true when the individual measurements have a uniform error distribution in the interval $[x - \frac{1}{2},\, x + \frac{1}{2}]$.

A.7.4 Gaussian Distributions in Two Variables

Bivariate Gaussian distributions describe two variables, x and y, that are separately Gaussian and are also linearly correlated with each other. We define their joint

Bivariate Gaussian distributions, for $\sigma_x = \sigma_y$ and various correlation coefficients ρ. Dots are drawn from $p(x, y)$; solid line is a 2σ contour enclosing $\approx 95\%$ of the points. **Fig. A.6**

distribution:

$$p(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{z}{2(1-\rho^2)}\right], \tag{A.189}$$

with

$$z \equiv \frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}. \tag{A.190}$$

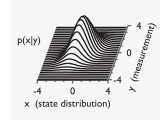
The bivariate Gaussian has some interesting properties (proved in problems):

1. Contours of equal probability are ellipses, with principal axes proportional to σ_x and σ_y, as illustrated in Fig. A.6.
2. The correlation coefficient ρ lies in the range $-1 \le \rho \le +1$ and measures how much fluctuations in x correlate with those in y. If the variables x and y are independent, then $\rho = 0$. Conversely, if $\rho = 0$, then x and y are independent. This converse relation is true for bivariate Gaussian distributions but not, in general, for other distributions.
3. Marginalizing, or "integrating out" y, leads to a Gaussian in x:

$$p(x) = \int_{-\infty}^{\infty} \mathrm{d}y\, p(x,y) = \mathcal{N}(\mu_x, \sigma_x^2). \tag{A.191}$$

   Likewise, marginalizing x leads to $p(y) = \mathcal{N}(\mu_y, \sigma_y^2)$.
4. If $p(x, y)$ is the bivariate distribution given in Eqs. (A.189) and (A.190), then the conditional distribution $p(x|y)$ is also Gaussian. We will see that

$$p(x|y) = \mathcal{N}(\mu_{x|y}, \sigma_{x|y}^2), \qquad \mu_{x|y} = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(y-\mu_y), \qquad \sigma_{x|y}^2 = \sigma_x^2(1-\rho^2). \tag{A.192}$$



We illustrate this statement at left, for $\rho = 0.5$. Each horizontal "cut" is a Gaussian-shaped curve. The varying size implies that the curve must be renormalized to form a proper distribution, $p(x|y)$. It is also plausible from the graph that the conditional mean $\mu_{x|y}$ increases linearly with y since, on average, when y is large, we expect x also to be large. Notice that $\mu_{x|y} \neq 0$ even if $\mu_x = 0$. Again, inspection of the graph at right should convince you of this. By contrast, the conditional variance is independent of y but depends on ρ. For $\rho = 0$, the x and y are independent, and we expect to see the full variance of the x measurement. For $\rho = 1$ and $\mu_x = \mu_y = 0$, we have $\mu_{x|y} = \frac{\sigma_x}{\sigma_y} y$, which shows that x is determined by y. The proportionality constant $\frac{\sigma_x}{\sigma_y}$ is a consequence of the different variances of the two

variables. Then, knowing y means that you know x perfectly, and the variance $\sigma^2_{x|y}$ must consequently be zero.

Problem A.7.6 Marginal and conditional distributions.

a. Show that if $p(x, y)$ is a bivariate Gaussian, then the marginal distributions are $p(x) = \mathcal{N}(\mu_x, \sigma_x^2)$ and $p(y) = \mathcal{N}(\mu_y, \sigma_y^2)$.
b. Show $p(x|y) = \mathcal{N}(\mu_{x|y}, \sigma^2_{x|y})$, with $\mu_{x|y} = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(y - \mu_y)$ and $\sigma^2_{x|y} = \sigma_x^2(1 - \rho^2)$.

Hint: Simplify by first defining $x' = (x - \mu_x)/\sigma_x$ and $y' = (y - \mu_y)/\sigma_y$.

A.7.5 Gaussian Distributions in n Variables

The results for a bivariate Gaussian distribution generalize to an n-dimensional multivariate distribution. To see how to generalize to n dimensions, we rewrite the bivariate distribution defined by Eqs. (A.189) and (A.190) in vector-matrix form, defining the vector $\boldsymbol{x} = \binom{x}{y}$. In terms of the mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, with determinant equal to $\det \boldsymbol{\Sigma}$, we have

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}. \tag{A.193}$$

and

$$p(\boldsymbol{x}) = \frac{1}{2\pi\sqrt{\det \boldsymbol{\Sigma}}} \exp\left[-(\boldsymbol{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right] \equiv \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \text{or} \quad \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{A.194}$$

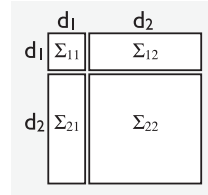
Equation (A.194) generalizes immediately to the case where $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^{\mathsf{T}}$ and where $\boldsymbol{\Sigma}$ is a symmetric $n \times n$, positive definite covariance matrix with $\boldsymbol{\Sigma}_{ij} = \langle x_i x_j \rangle$ (see Eq. A.179).

We list some of the useful properties of the multivariate Gaussian distribution, mostly without proof. (The proofs are analogous to the two-dimensional cases, but the algebra is more complicated.)

1. Contours of equal probability are ellipses, with principal axes equal to the square root of the eigenvalues of $\boldsymbol{\Sigma}$.
2. *A linear coordinate transformation* $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x}$ is Gaussian. If $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathsf{T}})$. We use this relation in our discussion of the Kalman filter in Chapter 8.
3. *Marginalization*. Marginalizing any subset of variables leaves a Gaussian in the remaining variables. In particular, if we partition $p(\boldsymbol{x})$ into two blocks of variables, $\boldsymbol{x}_1$ (d_1 dimensions) and $\boldsymbol{x}_2$ (d_2 dimensions, with $d_1 + d_2 = n$), then

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathcal{N}\left[\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\right], \tag{A.195}$$

where $\boldsymbol{\Sigma}$ has been partitioned into submatrices as depicted as left, so that



$$\underbrace{\boldsymbol{\Sigma}_{11}}_{d_1\times d_1} = \langle \boldsymbol{x}_1\, \boldsymbol{x}_1^{\mathsf{T}} \rangle, \qquad \boldsymbol{\Sigma}_{21}^{\mathsf{T}} = \underbrace{\boldsymbol{\Sigma}_{12}}_{d_1\times d_2} = \langle \boldsymbol{x}_1\, \boldsymbol{x}_2^{\mathsf{T}} \rangle, \qquad \underbrace{\boldsymbol{\Sigma}_{22}}_{d_2\times d_2} = \langle \boldsymbol{x}_2\, \boldsymbol{x}_2^{\mathsf{T}} \rangle\,.$$

The marginalization property is then

$$p(\boldsymbol{x}_1) = \int \mathrm{d}\boldsymbol{x}_2\, p(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})\,, \qquad p(\boldsymbol{x}_2) = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})\,. \tag{A.196}$$

4. *Conditional distributions.* For the same partitioning, $p(\boldsymbol{x}_1|\boldsymbol{x}_2) = \mathcal{N}(\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2})$, with

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)\,, \qquad \boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\,. \tag{A.197}$$

   As in the bivariate case, the conditional mean $\boldsymbol{\mu}_{1|2}$ depends linearly on the given variables $\boldsymbol{x}_2$ and is nonzero even when $\boldsymbol{\mu}_1 = \mathbf{0}$. Note that knowing $\boldsymbol{x}_2$ alters the estimate of $\boldsymbol{x}_1$ (i.e., the mean $\boldsymbol{\mu}_{1|2}$) in proportion to the covariance $\boldsymbol{\Sigma}_{1|2}$. The covariance also determines how much knowing $\boldsymbol{x}_2$ reduces the uncertainty in $\boldsymbol{x}_1$. For a nice proof based on completing the square and $p(\boldsymbol{x}_1|\boldsymbol{x}_2) = \frac{p(\boldsymbol{x}_1,\boldsymbol{x}_2)}{p(\boldsymbol{x}_2)}$, see Coolen et al. (2005).

5. *Characteristic function.* The multivariate generalization of Eq. (A.183) is

$$\varphi_{\boldsymbol{x}}(\boldsymbol{k}) = \left\langle \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \right\rangle = \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\mu} - \boldsymbol{k}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{k}}\,. \tag{A.198}$$

Problem A.7.7 Check that the expressions for mean and variance for the bivariate distribution (Eq. A.192) are compatible with their n-dimensional generalizations (Eq. A.197).

Problem A.7.8 **Propagating means and covariances.** Prove Property A.7.5 for multivariate Gaussians: if $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the linear coordinate transformation $z = \boldsymbol{A}\boldsymbol{x}$ is Gaussian with mean $\boldsymbol{A}\boldsymbol{\mu}$ and covariance $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathsf{T}}$. More succinctly, $z \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu}\,, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathsf{T}})$. Hint: You can use Eq. (A.179), but characteristic functions are simpler.

A.8 Statistics

Probability theory allows one to understand the consequences of randomness and indeterminacy in a "forward direction," as characterized by the probability distribution function (or its various moments). The goal of the field of *statistics* is to use data to infer, or *estimate* probabilities – to go "backwards" from data to inferences about the underlying probabilities. In our presentation, the main tool for carrying out this backwards inference will be Bayes' theorem.

A classic problem is to estimate, based on the data, quantities such as the mean and variance. For example, a common estimator for the mean is the *average*. Using

methods discussed below in Section A.8.2,[9] we will see that if there are N observations, x_i, then we can estimate the mean μ of a probability distribution as

$$\hat{\mu} \equiv \frac{1}{N} \sum_{i=1}^{N} x_i \tag{A.199}$$

We use the "hat" notation here and elsewhere to denote an estimator of a quantity.

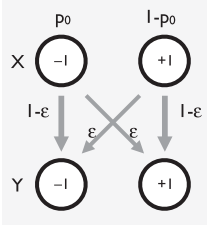
By similar means, we can also estimate the variance of a probability distribution as

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \hat{\mu})^2 \tag{A.200}$$

Note that estimators are not unique. Instead of the average, we could use just a single measurement, say $\hat{\mu}_1 = x_1$ and ignore the rest. Intuitively, the average, Eq. (A.199), should be better, as it uses all of the observations. We can quantify this by realizing that estimators are themselves random variables, or *statistics*. Thus, one can define the mean and variance for the estimator itself. For the estimators of the mean, $\hat{\mu}$ and $\hat{\mu}_1$, the variances are σ/N and σ, respectively. We then prefer $\hat{\mu}$ to $\hat{\mu}_1$ because it has lower variance for $N > 1$.

Another desirable property of an estimator is that its mean should equal the value of the quantity it is estimating. If it does, the estimator is *unbiased*. Both estimators of the mean, $\hat{\mu}$ and $\hat{\mu}_1$ are unbiased. The estimator $\hat{\sigma}^2$ is also unbiased. If the mysterious $N - 1$ factor in the denominator were replaced with the more obvious N, the new estimator [called the *population variance*, in contrast to the *sample variance* of Eq. (A.200)] would be biased for any finite value of N. Of course, when $N \gg 1$, the bias is small and vanishes in the limit $N \to \infty$. We then say that the population variance is *asymptotically unbiased*.

A.8.1 Simple Inference



To begin to understand how observations lead to inference on underlying events, let us consider the case where the underlying system that is being observed can be in one of two states. In Chapter 12, we will denote such states as +1 and −1, but any two symbols ({0, 1}, {L,R}), and so on, will do. The system is in state +1 with probability p_0. We measure the system in a way that also gives two outcomes, but the measurement is noisy: the observation is incorrect with probability ε (for both cases; see right). Then, if the result of an observation is $y = 1$, what is the probability that the underlying state is the same?

To translate the words and images into mathematics, we define $p(X = +1) = p_0$, so that $p(X = -1) = 1 - p_0$. For the measurement process, $p(Y = +1|X = +1) = p(Y =$

[9] Using N independent observations to make a single inference is equivalent to fitting noisy data to the function $f(x) = \mu$, with μ a single, free parameter.

$-1|X = -1) = 1 - \varepsilon$ and $p(Y = +1|X = -1) = p(Y = -1|X = +1) = \varepsilon$. From Bayes' theorem,

$$\begin{aligned} p(X = +1|Y = +1) &= \frac{p(Y = +1|X = +1)p(X = +1)}{p(Y = +1)} \\ &= \frac{p(Y = +1|X = +1)p(X = +1)}{p(Y = +1|X = +1)p(X = +1) + p(Y = +1|X = -1)p(X = -1)} \\ &= \frac{(1 - \varepsilon)p_0}{(1 - \varepsilon)p_0 + \varepsilon(1 - p_0)} . \end{aligned} \tag{A.201}$$

Notice that $\varepsilon \to 0$ implies that $p(X = +1|Y = +1) \to 1$: With no noise, the inference has no uncertainty. In the other extreme case, $\varepsilon \to 0.5$ and $p(X = +1|Y = +1) \to p_0$. If the observation conveys no information (returns ± 1 with equal probabilities, no matter what the underlying state), then we fall back on our prior, which is our estimate of the probabilities before we made the measurement Y. In the intermediate case, the probability is between p_0 and 1.

In Chapter 12, we extend this kind of interference to the case of a sequence of observations of a hidden Markov model (HMM). In that case, the underlying state probabilities remain p_0 and $1-p_0$, and rates of transition from one state to another are given. The problem then is to estimate the current state probabilities given a sequence of observations going back indefinitely into the past.

Example A.19 (Batch or sequential processing?) Is there a fundamental difference between analyzing a *batch* of data $y^k = \{y_1, y_2, \ldots, y_k\}$, as opposed to doing the analysis *sequentially*? A simple Bayesian analysis confirms our hunch that the two are equivalent. Let us discuss this for two observations, y_1 and y_2, with the goal of making an inference about a state x. Then, the batch analysis tells us that

$$p(x|y_1, y_2) \propto p(y_1, y_2|x)p(x) \propto p(y_1|x)\, p(y_2|x)\, p(x)\,, \tag{A.202}$$

where the first step uses Bayes' theorem and the second the independence of the two observations. Note that we leave out the normalization constant. If needed, it can be computed at the last step.

Now let us analyze the problem sequentially. In the absence of measurements, all we know is encapsulated in $p(x)$. After the first measurement, Bayes' theorem implies that $p(x|y_1) \propto p(y_1|x)\, p(x)$. After the second measurement, we have

$$p(x|y_2, y_1) \propto p(y_2|x, y_1)\, p(x|y_1) \propto p(y_2|x)\, p(x|y_1) \propto p(y_2|x)\, p(y_1|x)\, p(x)\,, \tag{A.203}$$

which is the same as the batch algorithm. The first step uses Bayes' theorem, and the second uses the independence of observation y_2 from the previous observation y_1.

Thus, it makes no difference whether we analyze our data all at once or one by one. Bayes' theorem is a statement about *logical* relationships, not *temporal* ones. We use this result several times in the book. An elementary example from Chapter 8 is that we can estimate a sample mean equivalently by either a batch or a recursive algorithm.

Problem A.8.1 Two Gaussian measurements. Often, we need to combine information from independent measurements with different precision. Assume that there are two independent measurements x_1 and x_2, distributed as $x_1 \sim \mathcal{N}(\mu, \sigma_1^2)$ and $x_2 \sim \mathcal{N}(\mu, \sigma_2^2)$. The variances σ_1 and σ_2 are known, and we wish to infer the mean, μ. Assume a uniform prior for μ.

a. Using Bayes' theorem, show that you can estimate μ as $\hat{\mu} \pm \sigma_\mu$, with

$$\hat{\mu} = \sigma_\mu^2 \left(\frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2} \right), \qquad \frac{1}{\sigma_\mu^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}.$$

b. Show that $p(\mu | x_1, x_2, \sigma_1^2, \sigma_2^2)$ is in fact a Gaussian, with the mean and variance given above. If you have a computer algebra program, do this in general. If you are doing the problem by hand, show the claim assuming that $\sigma_1^2 = \sigma_2^2 = 1$.

A.8.2 Estimating Parameters

Another estimation problem is to infer the values of parameters entering into a relation between two variables. Let us assume that we have a data set of N points (x_i, y_i), $i = 1 \cdots N$, that is generated by the model $y = f(x, \boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^*$ is a vector of K parameters. That is,

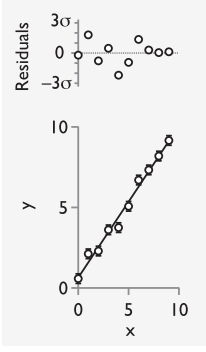
$$y_i = f(x_i, \boldsymbol{\theta}^*) + \xi_i \,, \tag{A.204}$$

where $\xi_i \sim \mathcal{N}(0, \sigma^2)$ is Gaussian noise of mean 0 and variance σ^2. We also require different ξ_i to be independent: if we collect the N noise terms into a vector $\boldsymbol{\xi}$, we have $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I})$. The usual least-squares procedure is then to minimize the sum χ^2,

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left(\frac{1}{\sigma^2} \right) [y_i - f(x_i, \boldsymbol{\theta})]^2 \,, \tag{A.205}$$

where the model function is

$$y = f(x, \boldsymbol{\theta}) \,, \quad \boldsymbol{\theta} = \begin{pmatrix} \theta_1 & \dots & \theta_K \end{pmatrix}^{\mathsf{T}} . \tag{A.206}$$



More generally, the noise could have a different variance at each point i (then $\sigma^2 \to \sigma_i^2$), and there can even be correlations between different noise components.

The χ^2 statistic is minimized by taking derivatives with respect to the θ_ℓ. For example, a linear model (straight line) would have the form $y = \theta_0 + \theta_1 x$, as illustrated at right. Note that we also plot the normalized residuals, $\varepsilon_i \equiv \frac{1}{\sigma}[y_i - f(x_i, \boldsymbol{\theta})]$. As we discuss below, the residuals should be approximately distributed as $\mathcal{N}(0, 1)$; however, the K fit parameters lead to correlations between the residuals, so that they are not independent. For example, in a linear fit, $\sum_i \varepsilon_i = 0$; see below.

To understand where this "recipe" comes from, let us return to Bayesian ideas of parameter estimation. To estimate the parameters $\boldsymbol{\theta}$ given a set of observations $Y \equiv \{y_i\}$ taken at $X \equiv \{x_i\}$, we can use Bayes' theorem to evaluate

$$p(\boldsymbol{\theta} | Y, X) = \frac{1}{Z} p(Y | X, \boldsymbol{\theta}) p(\boldsymbol{\theta}, \boldsymbol{X}) \,, \tag{A.207}$$

where Z normalizes the posterior $p(\boldsymbol{\theta}|Y,X)$. Note that the background information in all of the probability distributions includes the "knowledge" that the observations Y are described by the function $f(x,\boldsymbol{\theta})$. To keep the notation simple, we omit an explicit reference to this hypothesis in the conditioning of probabilities. If the a priori distribution of parameters and data points $p(\boldsymbol{\theta},\boldsymbol{X})$ is taken to be uniform, then

$$p(\boldsymbol{\theta}|Y,X) \propto p(Y|X,\boldsymbol{\theta}) \equiv \mathcal{L}\,. \tag{A.208}$$

where $\mathcal{L}$ is known as the *likelihood*. If the noise affecting each observation is independent of the others, then

$$\mathcal{L} = p(Y|X,\boldsymbol{\theta}) = \prod_{i=1}^{N} p(y_i|x_i,\boldsymbol{\theta})\,, \tag{A.209}$$

Now, we specialize to the case of Gaussian measurement errors, which implies that

$$\begin{aligned}\mathcal{L} &= \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{[y_i - f(x_i,\boldsymbol{\theta})]^2}{2\sigma^2}\right] \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left\{-\sum_{i=1}^{N}\left[\frac{[y_i - f(x_i,\boldsymbol{\theta})]^2}{2\sigma^2}\right]\right\} \\ &\equiv \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \mathrm{e}^{-\frac{1}{2}\chi^2}\,,\end{aligned} \tag{A.210}$$

where χ^2 is defined above in Eq. (A.205). To summarize, the probability distribution of the parameters $\boldsymbol{\theta}$ given observations Y and Gaussian measurement noise is

$$p(\boldsymbol{\theta}|Y,X) = \tfrac{1}{Z}\ \mathrm{e}^{-\frac{1}{2}\chi^2}\,, \tag{A.211}$$

where we include in the normalization Z all terms that do not depend on the parameters $\boldsymbol{\theta}$.

Maximum-likelihood (ML) estimation is useful even in cases where the measurement noise is not Gaussian. Then (for independent measurements), we revert to Eq. (A.209). The ML estimator has many desirable features (Krishnamurthy, 2016), notably

- *Strong Consistency*: $\lim_{N\to\infty} \hat{\theta}_{\mathrm{ML}}(N) \to \theta^*$, where θ^* is the "true" value of the parameter (in a frequentist perspective).
- *Asymptotic Normality*: $\lim_{N\to\infty} \hat{\theta}_{\mathrm{ML}}(N) - \theta^* \sim \frac{1}{\sqrt{N}}\mathcal{N}(\mathbf{0},\mathcal{I})$, where the covariance matrix $\mathcal{I} = \langle(\nabla_{\boldsymbol{\theta}} \log p(y|x,\boldsymbol{\theta}))^2\rangle$ is known as the Fisher information matrix.

Finally, we can put to rest one worry that often comes up: a truly uniform prior on an infinite range is not normalizable. But all that we require in our use of Bayes' theorem in Eq. (A.207) is that the prior be broader than the likelihood function, so that a large but finite range for each parameter is acceptable. In Bayesian terminology, we require that the likelihood dominate over the prior in determining the posterior. More intuitively, the method of maximum likelihood is relevant when the data are more important than our prior beliefs in estimating parameter values. We usually

want to be in this limit. The case of state estimation is an exception because our prior is based on many previous observations, and we ask how to update based on a single additional observation. The power of the Kalman filter and related ideas explored in Chapter 8 is precisely that a well-chosen prior can lead to a much better estimate of the state vector than is given by a single observation.

A.8.3 Nonlinear Fits

In general, the model $f(x, \boldsymbol{\theta})$ will depend nonlinearly on the parameters $\boldsymbol{\theta}$. For example,a fit to an exponential plus background would have $f(x) = \theta_1 + \theta_2 \exp(-\theta_3 x)$, which depends linearly on θ_1 and θ_2 but nonlinearly on θ_3. Minimizing $\chi^2(\boldsymbol{\theta})$ is then done numerically using the same ideas as other optimization problems (cf. Appendix A.5 and Chapter 7). The *Levenberg-Marquardt algorithm* is a standard way to find a local solution given initial conditions for $\boldsymbol{\theta}_0$ that are "close enough" to the true values $\boldsymbol{\theta}^*$. You have to supply the $\boldsymbol{\theta}_0$. More complicated algorithms such as the *genetic algorithm* attempt to find a global minimum for χ^2 by allowing different initial conditions to compete, applying biologically inspired strategies that mimic evolution. Standard curvefit software almost always offers the former and usually the latter – or other possibilities, such as *simulated annealing*.

A.8.4 Linear Fits

If $f(x, \boldsymbol{\theta})$ is linear in the parameters $\boldsymbol{\theta}$, then $\chi^2(\boldsymbol{\theta})$ is quadratic in $\boldsymbol{\theta}$ and hence convex, with a unique minimum. Indeed, taking a derivative leads to K linear equations, which have a single solution for the K parameters $\boldsymbol{\theta}$. In more detail, we define the linear model

$$f(x, \boldsymbol{\theta}) = \sum_{k=1}^{K} \theta_k \, e_k(x) \,, \tag{A.212}$$

with K *basis functions* $e_k(x)$. Note that $f(x, \boldsymbol{\theta})$ is linear in the parameters $\boldsymbol{\theta}$ but not necessarily in the variable x.

In elementary discussions of curve fitting, we often consider $e_k = 1, x, x^2, \ldots$, but many other choices are possible. If f and g are real and the scalar product defined to be $f \cdot g \equiv \int_a^b \mathrm{d}x \, w(x) \, f(x) \, g(x)$, then an orthonormal basis set will satisfy

$$\int_a^b \mathrm{d}x \, w(x) \, e_j(x) \, e_k(x) = \delta_{jk} \,. \tag{A.213}$$

Here, a and b define a domain that can be finite or infinite in extent, and $w(x)$ is a weight function that in some cases equals 1. Although forming a complete basis (all integrable functions can be expanded over them), the monomials are not orthogonal. Bases that are orthogonal include Chebyshev polynomials, sines and cosines, and Bessel functions.

When $f(x, \boldsymbol{\theta})$ is linear in the parameters $\boldsymbol{\theta}$, the likelihood function $\mathcal{L}$ is a Gaussian in the parameters $\boldsymbol{\theta}$ that is peaked about a set of values $\boldsymbol{\theta}^*$ that both maximizes $\mathcal{L}$ and

minimizes χ^2. If the error distribution widths σ_i are all similar and approximately equal to σ, it is easy to see that the density P(theta | Y,X) for each parameter has a Gaussian form, with variance $\approx \sigma^2/N$. As N gets large, the distribution narrows, and there is little uncertainty in each parameter θ_m. Even when the function f is not linear in its parameters $\boldsymbol{\theta}$, a second-order Taylor expansion gives a Gaussian approximation to the likelihood function.

Since Gaussian distributions are fully characterized by their mean and covariance matrix, we conclude that when the observation noise is Gaussian, we can infer parameters $\boldsymbol{\theta}^*$ by minimizing the χ^2 statistic. More generally, if the observation noise is not Gaussian, we can follow the same procedure to the point of maximizing the log likelihood function $\ln \mathcal{L}$ and refer, in that case, to the method of *maximum likelihood*. (Recall that the parameters that maximize $\ln \mathcal{L}$ also maximize $\mathcal{L}$.) We could generalize even further – as we do in the state estimation problem or in the recursive inference of parameters – by allowing for a more informative prior $P(\boldsymbol{\theta})$ than the minimal uniform distribution assumed here.

We can solve directly for the optimal parameter estimates when the model function is linear in parameters. We want to minimize

$$\chi^2 = \frac{1}{\sigma^2}\sum_{i=1}^{N}\left[y_i - \sum_{k=1}^{K}\theta_k\, e_k(x_i)\right]^2 \equiv \frac{1}{\sigma^2}\sum_{i=1}^{N}\left[y_i - \sum_{k=1}^{K}\Phi_{ik}\,\theta_k\right]^2$$

$$\equiv \frac{1}{\sigma^2}\,(\boldsymbol{y}-\boldsymbol{\Phi\theta})^{\mathsf{T}}\,(\boldsymbol{y}-\boldsymbol{\Phi\theta}) = \frac{1}{\sigma^2}\,\|\,\boldsymbol{y}-\boldsymbol{\Phi\theta}\,\|^2\,, \tag{A.214}$$

where the *design matrix* $\boldsymbol{\Phi}(\boldsymbol{x})$ has elements $\Phi_{ik} \equiv e_k(x_i)$. For example, with $f(x,\boldsymbol{\theta}) = \theta_1 x + \theta_2 x^2$ ($K = 2$ parameters) and with $N = 3$ data points, the data are generated from

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} x_1 & x_1^2 \\ x_2 & x_2^2 \\ x_3 & x_3^2 \end{pmatrix}\begin{pmatrix} \theta_1^* \\ \theta_2^* \end{pmatrix} + \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix}, \quad \text{or} \quad \boldsymbol{y} = \boldsymbol{\Phi\theta}^* + \boldsymbol{\xi}\,. \tag{A.215}$$

where θ_k^* are the "true" parameter values of the process. To find the estimate of $\boldsymbol{\theta}$ that minimizes $\chi^2(\boldsymbol{\theta})$, we differentiate with respect to θ_ℓ:

$$\frac{\partial\chi^2}{\partial\theta_\ell} = \frac{2}{\sigma^2}\sum_{i=1}^{N}\left[y_i - \sum_{k=1}^{K}\theta_k\Phi_{ik}\right][-\Phi_{i\ell}] = 0\,, \tag{A.216}$$

which implies, in component and then vector notation, the *normal equations*:

$$\sum_{i=1}^{N} y_i\Phi_{i\ell} = \sum_{i,k}\Phi_{ik}\Phi_{i\ell}\theta_k$$

$$\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{y} = \boldsymbol{\Phi}^{\mathsf{T}}\,\boldsymbol{\Phi}\,\boldsymbol{\theta}\,, \tag{A.217}$$

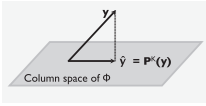
In this notation, $\boldsymbol{y}$ is an N-component vector of data, $\boldsymbol{\theta}$ is still a K-component vector of parameters, and $\boldsymbol{\Phi}$ is an $N \times K$ matrix. If the $K \times K$ square

matrix $\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}$ is invertible,[10] we can solve for $\hat{\boldsymbol{\theta}}$, the parameter estimate that minimizes χ^2:

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{y} \equiv \boldsymbol{\Phi}^{+}\,\boldsymbol{y}\,, \tag{A.218}$$

where $\boldsymbol{\Phi}^+$ is the *Moore-Penrose pseudoinverse* of $\boldsymbol{\Phi}$.[11] In Problem A.8.2, we show that if the measurement noise $\boldsymbol{\xi}$ is white, then $\hat{\boldsymbol{\theta}} \sim \mathcal{N}[\hat{\boldsymbol{\theta}}^*, \sigma^2(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi})^{-1}]$. In other words, the estimate $\hat{\boldsymbol{\theta}}$ is unbiased. The best estimate for the data $\boldsymbol{y}$ is

$$\hat{\boldsymbol{y}} = \boldsymbol{\Phi}\,\hat{\boldsymbol{\theta}} \equiv \boldsymbol{P}^{(K)}\boldsymbol{y}\,, \qquad \boldsymbol{P}^{(K)} = \boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\mathsf{T}\,. \tag{A.219}$$



In Problem A.8.3, we will see that $\boldsymbol{P}^{(K)}$ projects the N-dimensional vector of observations $\boldsymbol{y}$ onto the K-dimensional subspace spanned by the columns of $\boldsymbol{\Phi}$.[12] Intuitively, as shown at right, the projection of a vector gives the closest distance between the original vector and the plane, which represents the space of all possible model vectors. Because $\boldsymbol{P}^{(K)}$ is a projector matrix, it must be *idempotent*, meaning $\boldsymbol{P}^{(K)2} = \boldsymbol{P}^{(K)}$. We can show this directly:

$$\boldsymbol{P}^{(K)2} = \underbrace{\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\mathsf{T}}_{\boldsymbol{P}^{(K)}}\,\underbrace{\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\mathsf{T}}_{\boldsymbol{P}^{(K)}} = \boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\mathsf{T} = \boldsymbol{P}^{(K)}\,. \tag{A.220}$$

$\boldsymbol{P}^{(K)}$ is also symmetric, hence diagonalizable: $\boldsymbol{P}^{(K)} = \boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^{-1}$. The diagonal matrix of eigenvalues then satisfies $\boldsymbol{D}^2 = \boldsymbol{D}$. Correspondingly, each eigenvalue λ satisfies $\lambda^2 = \lambda$, and is thus 0 or 1. The rank of $\boldsymbol{P}^{(K)}$ is then the number of eigenvalues that = 1, which is just K. Note that $\boldsymbol{P}^{(K)}$ itself is $N \times N$ dimensional.

A.8.5 Assessing a Curve Fit

If observational noise is Gaussian, if each point has independent errors, if the model function is linear in the unknown parameters, and if we have little prior knowledge of the parameter values, then we are in the limit where the posterior distribution is Gaussian in the parameters. To assess whether all these hypotheses hold, we can use the value of the χ^2 statistic itself. Its mean value is

$$\begin{aligned}\langle\chi^2\rangle &= \frac{1}{\sigma^2}\sum_{i=1}^{N}\left\langle\left[y_i - \sum_{k=1}^{K}\theta_k\Phi_{ik}\right]^2\right\rangle = \frac{1}{\sigma^2}\left\langle\|\boldsymbol{y}-\hat{\boldsymbol{y}}\|^2\right\rangle = \frac{1}{\sigma^2}\left\langle\left\|(\boldsymbol{I}-\boldsymbol{P}^{(\boldsymbol{K})})\boldsymbol{y}\right\|^2\right\rangle \\ &= \frac{1}{\sigma^2}\mathrm{Tr}\left(\mathbb{I}-\boldsymbol{P}^{(K)}\right)^2\langle y^2\rangle = \frac{\sigma^2}{\sigma^2}\mathrm{Tr}\left(\mathbb{I}-\boldsymbol{P}^{(K)}\right)^2 = \mathrm{Tr}\left(\mathbb{I}-\boldsymbol{P}^{(K)}\right) = N-K \equiv \nu\,.\end{aligned} \tag{A.221}$$

In the last step, we recall that the trace of a matrix equals the sum of its eigenvalues. The quantity ν is the number of *degrees of freedom* of the fit. Intuitively, each

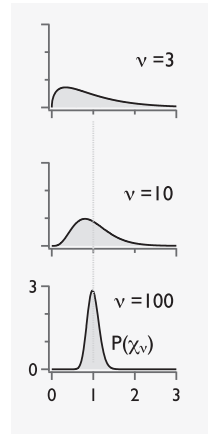
[10] The invertibility of $\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}$ can be a delicate question. In Chapters 6 and 10, we use least-squares fits for system identification. There, the $\boldsymbol{\Phi}$ matrix is formed from both inputs u and outputs y of a dynamical system. The input must satisfy conditions of *persistent excitation* to ensure that $\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}$ is invertible.

[11] Notice that if $\boldsymbol{\Phi}$ were invertible, then $\boldsymbol{\Phi}^+ = \boldsymbol{\Phi}^{-1}$. Here, $\boldsymbol{\Phi}$ is almost never invertible because there would then be exactly as many data points as parameters ($N = K$). Nonetheless, we can write $\boldsymbol{\Phi}^+\boldsymbol{\Phi} = \mathbb{I}$.

[12] The statistics literature uses the notation $\boldsymbol{H}$, for *hat matrix*, because it "puts the hat onto $\boldsymbol{y}$." While this makes for a cute in-joke, I prefer to emphasize the geometrical nature of the transformation.

fit parameter "absorbs" one degree of freedom and reduces the expected χ^2 accordingly. We note that this argument also shows that the residuals $\boldsymbol{y} - \hat{\boldsymbol{y}} = \left(\mathbb{I} - \boldsymbol{P}^{(K)}\right)\boldsymbol{y}$ are correlated Gaussian variables.

A more detailed argument shows that the variance of $P(\chi^2)$ is given by 2ν and, moreover, that for $\nu \gg 1$, the χ^2 distribution is asymptotically normal: $\chi^2 \sim \mathcal{N}(\nu, 2\nu)$. We illustrate this convergence at left, where we normalize χ^2 by ν, so that the average should be unity. Values of χ^2 much outside the range $\nu \pm \sqrt{2\nu}$ should be treated with suspicion. High values of χ^2 may result from *underfitting* – the model is too simple to describe the data and there are systematic errors – or from using error estimates σ_i that are too small. (Recall that $\chi^2 \sim 1/\sigma^2$.) Conversely, low values of χ^2 may result from *overfitting* – the model has too many parameters and is fitting the noise – or from values of σ that are too large.

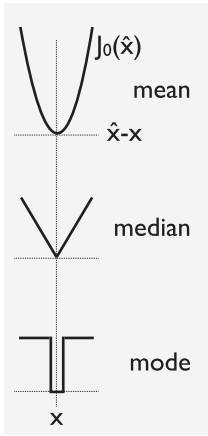


A.8.6 Model Structure Assessment

The χ^2 statistic is not the whole story of model assessment because we must also choose the model structure, the $f(x, \boldsymbol{\theta})$ used in the fit. We should not be impressed by models that can "fit anything," even if they lead to better fits than does a simple model, as they may simply be fitting to the noise. Although it would be logical to develop this idea here, its novelty to many physicists has led me to include it as part of the main text in the discussion of System Identification in Section 6.4. Briefly, we describe two different ways to assess a model structure. The first, *minimum description length* (MDL), looks for the "simplest" model, where simplicity is given a technical definition that amounts to counting fit parameters. It is most appropriate for situations where the goal is to find and describe the underlying "true" model describing the data. The second way of assessing a curve fit uses *cross-validation*, a model's ability to predict new data. It is often appropriate for control problems, where the performance of a control system is limited by the ability to predict the next observation. Curiously, the two principles can lead to somewhat different model-selection criteria.

A.8.7 From Posterior Distribution to Best Estimate

The result of a Bayesian analysis is a posterior distribution. For example, in Section 8.3.1, we looked at a very simple example, the problem of inferring a constant x from a noisy observation $y = x + \xi$ when the prior $p(x)$ and likelihood function $p(y|x)$ are known Gaussian distributions. The resulting posterior distribution, Eq. (8.66), for $p(x|y)$ also was Gaussian. Since a Gaussian is completely characterized by its mean and variance, we can equally well give those values instead of the entire formula for a Gaussian. This is what we do implicitly when we "make a measurement" and report $x \pm \Delta x$. More generally, we would like to summarize a posterior distribution by a "typical" value and an "error," even if the distribution is not a Gaussian. Obviously,



such an approach only makes sense when the distribution is at least approximately a "bell-shaped curve."

But how to go from a probability distribution to a couple of numbers? Here, there is no unique answer. A sensible approach, *decision theory*, draws on ideas from optimal control (Chapter 7). The strategy is to define a *cost function* $J(\hat{x}) \equiv \langle J_0(\hat{x}) \rangle_y$, which quantifies the average "cost" of choosing $\hat{x}$ as the "best" estimator of x, given the posterior distribution for x that is conditioned on the observation y. Here $J_0(\hat{x})$ is the cost function for the deterministic problem.

As the sketch at right shows, there are several reasonable candidates for $J_0(\hat{x})$. From top to bottom, we minimize the square of deviations, their absolute value, or we add a constant cost for all deviations outside a range $\pm\Delta$ from the true value. In the first case, for example, we look for the value $\hat{x}$ that minimizes the square of the deviations from the unknown "true" value x, which is drawn from $p(x|y)$. Then, with $J_0(\hat{x}) = (\hat{x} - x)^2$, we have

$$J(\hat{x}) = \left\langle (\hat{x} - x)^2 \right\rangle_y = \int \mathrm{d}x \, p(x|y) \, (\hat{x} - x)^2 \, . \tag{A.222}$$

Differentiating with respect to $\hat{x}$ to find the optimum gives

$$\left. \frac{dJ}{d\hat{x}} \right|_{\hat{x}=\hat{x}^*} = 2 \int \mathrm{d}x \, p(x|y) \, (\hat{x}^* - x) = 2 \left(\hat{x}^* - \langle x \rangle_y \right) = 0 \, , \tag{A.223}$$

using $\int \mathrm{d}x \, p(x|y) = 1$. Thus,

$$\hat{x}^*(y) = \langle x \rangle_y \, . \tag{A.224}$$

Since $J(\hat{x})$ is convex and $J''(\hat{x}^*) = 2 > 0$, we conclude that $\hat{x}^* = \langle x \rangle_y$ minimizes J. Thus, with a mean-square-deviation cost function, the "best" value $\hat{x}^*$ is the (conditional) mean. We can use the same idea in many contexts, for example to estimate the "best" parameters in a least-squares fit. Often, we will drop the * and not indicate the explicit y dependence, using simply $\hat{x}$.

In Problems A.8.4 and A.8.5, you will explore the other two cost functions, showing that minimizing the absolute value of deviations leads to the choice of the *median* value of $p(x|y)$ and assigning a constant cost to deviations outside a small range $\pm\Delta$ leads to the *mode*. In the context of Bayesian parameter estimation, the choice of mode is termed the *maximum a posteriori* (MAP) estimate.

A.8.8 Regularization

As discussed in Section A.8.5, when the number of parameters in a model becomes large, there can be a danger of overfitting. The χ^2 value is then anomalously low.

Ridge regression provides a way to *regularize* the χ^2 statistic to[13]

$$\chi^2(\boldsymbol{\theta}) = \frac{1}{\sigma^2} \| \boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta} \|^2 + \beta \|\boldsymbol{\theta}\|^2 \ . \tag{A.225}$$

In Eq. (A.225), the β parameter is equivalent to putting a prior distribution on the values of the unknown parameters $\boldsymbol{\theta}$. Indeed, if the prior for each θ_i is Gaussian, with standard deviation σ_θ, we can identify $\beta = 1/\sigma_\theta^2$. More informally, Eq. (A.225) expresses a compromise between choosing the parameters to fit the data (first term) while keeping parameter values small (second term).

Minimizing the augmented χ^2 statistic in Eq. (A.225) leads to a straightforward generalization of Eq. (A.218):

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi} + \beta \mathbb{I} \right)^{-1} \boldsymbol{\Phi}^\mathsf{T} \boldsymbol{y} \, , \tag{A.226}$$

where $\mathbb{I}$ is the $K \times K$ identity matrix. We see explicitly that for $\beta \to 0$, the estimate $\hat{\boldsymbol{\theta}}$ reduces to that of Eq. (A.218), while for $\beta \to \infty$, we have $\hat{\boldsymbol{\theta}} \to \boldsymbol{0}$.

Despite its connection to notions of a Bayesian prior, regularization can seem ad hoc. Its popularity stems from two key points: (1) overfitting tends to produce parameter fluctuations that are much larger than plausible parameter values, and (2) the inversion in Eq. (A.226) is numerically much more stable as β is increased from zero. Regularization is thus often a practical step for parameter estimation.

A.8.9 Monte Carlo Methods

Monte Carlo methods constitute an alternate way of approximating integrals over a probability distribution density $p(\boldsymbol{x})$ that is based on samples drawn from the distribution itself. In evaluating the integral in Eq. (8.103), we sampled uniformly from the n-dimensional domain bounding the set of possible states $\boldsymbol{x}$. We can ask, in a typical case, what fraction of the normalized weights P_i are appreciably nonzero? The question is equivalent to asking, "Over what fraction of the space does $p(\boldsymbol{x})$ have a non-negligible amplitude?" Imagine, for example, that in each dimension, $p(\cdot)$ is nonzero for $\frac{1}{10}$ of the total domain coordinate. Then in two dimensions, p would be nonzero on only 1% of the area. In three dimensions, it would be 0.1% of the volume, and in n-dimensions, it would be a fraction 10^{-n}. Thus, another consequence of the *curse of dimensionality* is that only a tiny fraction of the weights P_i differ appreciably from 0 in high dimensions. Evaluating p on all the grid points is then terribly inefficient.

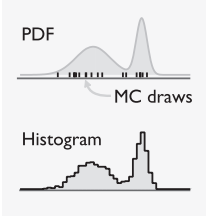
Let us then try the other extreme and draw a single point, $\boldsymbol{x}^1$ from $p(\boldsymbol{x})$ and try $\varphi(\boldsymbol{x}^1)$ as our estimator. Observe that the expectation value of $\varphi(\boldsymbol{x}^1)$ is simply given by the

[13] There are many ways to regularize the χ^2 statistic. A slight variant, *Tikhonov regularization*, defines $\chi^2(\boldsymbol{\theta}) = \frac{1}{\sigma^2} \| \boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta} \|^2 + \|\boldsymbol{\Gamma}\boldsymbol{\theta}\|^2$, with $\boldsymbol{\Gamma}$ a matrix. In this more general kind of regularization, we can impose more subtle constraints. For example, $\boldsymbol{\Gamma}$ can approximate a spatial-derivative operator that corresponds to differentiating $\boldsymbol{\theta}$ twice in a discrete representation. If the parameters correspond to the spatial degrees of freedom of a field, we would be favoring smoother solutions that had reduced "roughness." Thus, Tikhonov regularization can be appropriate for *inverse problems* where the goal is to find a non-parametric estimate of a function, while ridge regression can be appropriate for parameter estimation problems in curve fitting.

integral $\int d\boldsymbol{x}^1\, \varphi(\boldsymbol{x}^1) p(\boldsymbol{x}^1) \equiv \langle\varphi\rangle$. This single-point estimator is unbiased! Although you will immediately object that the variance, equal to Var φ, is large – if you happen to choose a point where $\varphi(\boldsymbol{x})$ has a very atypical value, your estimate will have a huge error – we can remedy that by picking a reasonable number N of *independently drawn* points and averaging over them. Thus, we arrive at the Monte Carlo estimate for $\langle\varphi(\boldsymbol{x})\rangle$:

$$\hat{p}(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} \delta\left(\boldsymbol{x} - \boldsymbol{x}^i\right), \qquad \hat{\varphi} = \frac{1}{N}\sum_{i=1}^{N} \varphi_i\,, \qquad \operatorname{Var}\hat{\varphi} = \frac{1}{N}\operatorname{Var}\varphi\,. \tag{A.227}$$

Equation (A.227) is of the form of Eq. (8.103), except that the weights w_i have become uniform ($w_i = \frac{1}{N}$) and the points $\boldsymbol{x}^i$ now are drawn from $p(\boldsymbol{x})$ and not evaluated on a grid. Note that we should, but do not, use a different symbol for the Monte Carlo estimator $\hat{\varphi}$; the context will be enough to distinguish which estimator is referred to. At right, Var $\hat{\varphi}$ states that the variance of the mean of the average N independent estimates is just $\frac{1}{N}$ the variance of one estimate (standard error of the mean). In Problem A.8.6, you will review this statement. Notice that the variance of $\hat{\varphi}$ depends only on the variance of φ itself and the number of samples N drawn from p. It is independent of the dimension n of the space!



To picture Monte Carlo sampling, consider the figure at right, which shows a continuous probability distribution at top (sum of two Gaussians). The ticks at bottom mark the positions of 20 draws, with their even weight represented by the even height of the lines. The bottom graph is a histogram, based on 1000 draws, which demonstrates that we can use the Monte Carlo approximation to derive a more conventional grid approximation of the distribution. Of course, to make this scheme work, we need to know how to draw numbers from odd-shaped probability distributions.[14]

The above argument that sampling from a PDF allows us to approximate the continuous density via a histogram is intuitively reasonable but not rigorous. To justify $p(\boldsymbol{x}) \approx \frac{1}{N}\sum_i \delta\left(\boldsymbol{x} - \boldsymbol{x}^i\right)$ more carefully, we look at the corresponding approximation to the cumulative density function (CDF). For simplicity, we restrict ourselves to a one-dimensional case, where x^i is distributed as $p(x)$. Then

$$\mathrm{CDF}(a) = \int_{-\infty}^{a} \mathrm{d}x\, p(x) \approx \int_{-\infty}^{a} \mathrm{d}x\, \frac{1}{N}\sum_{i=1}^{N} \delta\left(x - x^i\right) = \sum_{i=1}^{N} \frac{1}{N} 1_{(-\infty,a)}(x^i) \tag{A.228}$$

where $1_{(a,b)}(x)$ is the *indicator function*, which is 1 when $a \le x \le b$ and 0 otherwise. Equation (A.228) implies that our Monte-Carlo approximation to $p(x)$ is equivalent to an approximation to the CDF where the probability that $x \le a$ is by the fraction of elements $x^i \le a$, based on N trials. Since the x^i are drawn from $p(x)$, the law of large numbers implies that the fraction $\le a$ converges to the probability $x \le a$ and, hence, $\hat{p}(x) \to p(x)$.

Although Eq. (A.227) would seem to solve our problem, we still need to know how to draw points from a high-dimensional, irregular distribution. Above, we argued that

[14] For the example here, we sample using the technique of *rejection sampling*, which we do not cover because of its limited applicability to state-estimation problems. See MacKay (2003).

$p(\boldsymbol{x}) \approx 0$ for all but a vanishingly small subset of points, which, in general, are hard to find. In other words, at best, we have reduced one hard problem to another. Still, that is progress, and we can now draw on a wide array of techniques for drawing points from complicated distributions. We focus on two, *importance sampling* and *bootstrap resampling*, that have proven useful in particle-filter applications of MC techniques.

Importance Sampling

We have seen that if we can draw numbers from a probability distribution, we can approximate nicely all the quantities we need in state estimation. Yet we have argued that drawing numbers from a high-dimensional distribution is hard, as the density function p vanishes almost everywhere. Assume, then, that we have another density function $q(\boldsymbol{x})$ that is simple enough that we know how to draw numbers from it. Then, we can use q to help us evaluate $\langle\varphi\rangle_p$. Notice that we now are careful to specify whether we average over numbers drawn from p or from q. When $\boldsymbol{x}_i$ is distributed as $q(\boldsymbol{x})$, with $q(\boldsymbol{x})$ strictly positive, we have,

$$\hat{\varphi} = \frac{\sum_i w_i \varphi_i}{\sum_i w_i}, \qquad w_i \equiv \frac{p^*(\boldsymbol{x}^i)}{q(\boldsymbol{x}^i)}, \tag{A.229}$$

where $\varphi_i \equiv \varphi(\boldsymbol{x}_i)$. In Eq. (A.229), P may not be normalized, so that $p = p^*/Z_p$, where Z_p is the normalizing factors (partition functions). The weights w_i now are adjusted to have different *importance* depending on whether the distribution p^* or q is larger at a particular point $\boldsymbol{x}^i$.

Let us first show that $\hat{\varphi}$ is an unbiased estimator for $\langle\varphi\rangle$ when $N \to \infty$.[15] The easiest approach is to look separately at the numerator and the denominator. For the latter, averaging the point $\boldsymbol{x}^i$ over $q(\boldsymbol{x}^i)$, we have

$$\left\langle \sum_i w_i \right\rangle_q = \sum_i \int \mathrm{d}\boldsymbol{x}^i \, \frac{p^*(\boldsymbol{x}^i)}{q(\boldsymbol{x}^i)} q(\boldsymbol{x}^i) = N \int \mathrm{d}\boldsymbol{x}^i \, \frac{p^*(\boldsymbol{x})}{q(\boldsymbol{x})} q(\boldsymbol{x}) = N Z_p \,. \tag{A.230}$$

For the numerator, we have, similarly,

$$\left\langle \sum_i w_i \varphi_i \right\rangle_q = N Z_p \langle\varphi\rangle_p \,. \tag{A.231}$$

Note that although the numerator and denominator are each unbiased for any N, their ratio is unbiased only for $N \to \infty$. See Problem A.8.6.

Also in Problem A.8.6, we investigate the variance of $\hat{\varphi}$ in Eq. (A.229). The upshot will be that, assuming $\varphi(\boldsymbol{x}) > 0$, we can choose $q(\boldsymbol{x}) = p(\boldsymbol{x})\,\varphi(\boldsymbol{x})$ and make the variance vanish! Before celebrating, we note that the desired $q(\boldsymbol{x})$ must be normalized. And if we can normalize $Z_q = \int \mathrm{d}\boldsymbol{x}\, p(\boldsymbol{x})\,\varphi(\boldsymbol{x})$, then we can do our original integral and have no business trying to solve it by Monte Carlo methods. Still, the general implication is that by making $q(\boldsymbol{x})$ resemble $p(\boldsymbol{x})\,\varphi(\boldsymbol{x})$ as much as possible, we can reduce the variance of the importance-sampling estimator.

[15] Such estimators are termed *asymptotically unbiased*.

Bootstrap Resampling

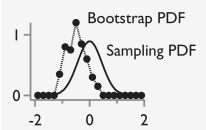
Bootstrap resampling is a simple idea that seems almost magical when you first hear of it. Imagine that we have drawn N "data points" (samples) $\boldsymbol{x}^i$ from a data set that follows a distribution $p(\boldsymbol{x})$. We denote this set of data points by $\{\boldsymbol{x}\}$. The twist is that we do not know the distribution explicitly. For example, the samples may come from an experiment that we do not understand well, or, in the case we are interested in this book, from a high-dimensional state vector whose distribution is too complicated to write down easily.

To get a feel for the bootstrap method, consider a statistic $\bar{\varphi}$ that is defined as a function of the data set. We distinguish $\bar{\varphi}$, which is evaluated over $\{\boldsymbol{x}\}$, from $\langle\varphi\rangle_p$, which is an integral over $p(\boldsymbol{x})$. To estimate the distribution of $\bar{\varphi}$, we start from the original data set $\{\boldsymbol{x}\}$ and define a new set $\{\boldsymbol{x}^*\}$ of N elements by *resampling with replacement*. That is, we draw N values $\boldsymbol{x}^i$ from a "bag" containing the original data set, each time replacing the value we picked before the next draw. The claim then is that we can approximate the sampling distribution of $\bar{\varphi}$ by evaluating it on the set of resampled elements:

$$\hat{p}(\bar{\varphi}) = \frac{1}{N_B^*} \sum_{i=1}^{N_B^*} p[\bar{\varphi}(\{\boldsymbol{x}^*\})] \,, \tag{A.232}$$

where $\{\boldsymbol{x}^*\}$ is a new set of N samples drawn, *with replacement*, from $\{\boldsymbol{x}\}$. The distribution in Eq. (A.232), known as the *exact bootstrap* estimate, requires evaluating all possible resamplings (with replacement), N_B^*. Since N_B^* is impractically large for even modest N,[16] we use a Monte Carlo estimate for Eq. (A.232) that is based on $N_B \ll N_B^*$ bootstrap resamplings. Again, each resampling involves drawing N elements from the original data set with replacement.

Example A.20 (Bootstrap estimate of the variance of the sample mean.) The most common application of bootstrap resampling is to estimate Var φ rather than the whole sampling PDF. We illustrate the Monte Carlo bootstrap method at right. We draw 5 samples from a Gaussian distribution, $\mathcal{N}(0,1)$, and are interested in the average, $\bar{\varphi} = \frac{1}{5}\sum_i x_i$, which approximates the mean $\varphi = \int \mathrm{d}x\, x\, p(x)$. The smooth curve shows the exact sampling distribution for $\bar{\varphi}$, which is simply $\mathcal{N}(0, \frac{1}{5})$. The histogram estimate of the PDF at right is based on 100 resamplings with replacement. Note that its form is close to the Gaussian of the true sampling distribution. For example, its standard



[16] The number of times n that an element $\boldsymbol{x}^i$ is included in a single sample with replacement is given by the multinomial distribution of N elements that each have probability $1/N$, which implies that the number of resamplings is $N_B^* = (2N-1)!/(N-1)!$. For $N = 15$, the number of possible resamplings is already $O(10^8)$.

deviation, 0.35, is close to $\frac{1}{\sqrt{5}} = 0.45$. But although the mean of the bootstrap distribution, −0.50, is close to the mean of the 5-element data set, −0.46, it is not (and cannot) be centered on 0, since all that we know about the true mean is derived from the original data set. But remarkably, we have found a good estimate of the variance. Indeed, we seem to have gotten something for nothing: with just one experiment or simulation, we have estimated the variability in a quantity, something that would have normally required many (costly) repetitions or a theoretical calculation. We know how to do the latter for an estimate of the mean, of course, but not for more complicated statistics.

A variation of resampling that is of direct interest for state estimation is to use importance sampling to estimate properties of a *different* distribution. The calculations reprise those of standard importance sampling. Assume that we have samples $\boldsymbol{x}^i \sim q(\boldsymbol{x})$ and wish to transform them into samples from the distribution $p(\boldsymbol{x}) = \frac{1}{Z_p} p^*(\boldsymbol{x})$, with $Z_p = \int \mathrm{d}\boldsymbol{x}\, p^*(\boldsymbol{x})$. To understand how to do this, we return to the estimate of the CDF in Eq. (A.228), with $q(x) \approx \sum_i \frac{1}{N} \delta(x - x^i)$. Then,

$$\begin{aligned}\mathrm{CDF}_p(a) = \int_{\infty}^{a} \mathrm{d}x\, p(x) &= \frac{\int_{-\infty}^{a} \mathrm{d}x\, \frac{p^*}{q}\, q(x)}{\int_{-\infty}^{-\infty} \mathrm{d}x\, \frac{p^*}{q}\, q(x)} \approx \frac{\frac{1}{N}\int_{-\infty}^{a} \mathrm{d}x \sum_i \frac{p^*(x)}{q(x)} \delta(x - x^i)}{\frac{1}{N}\sum_i \frac{p^*(x)}{q(x)} \delta(x - x^i)} \\ &= \frac{\frac{1}{N}\sum_i \frac{p^*(x^i)}{q(x^i)} 1_{(-\infty,a)}(x^i)}{\frac{1}{N}\sum_i \frac{p^*(x^i)}{q(x^i)}} = \frac{\frac{1}{N}\sum_i w_i 1_{(-\infty,a)}(x^i)}{\frac{1}{N}\sum_i w_i} = \sum_{i=1}^{N} \tilde{w}_i 1_{(-\infty,a)}(x^i) \qquad \text{(A.233)}\end{aligned}$$

where we have defined weights $w_i \equiv \frac{p^*(x^i)}{q(x^i)}$ and $\tilde{w}_i \equiv \frac{w_i}{\sum_i w_i}$. Thus, if we weight the event x^i by $\tilde{w}_i$ in all sums involving expectation values, the result will be equivalent to having drawn a value from $p(x)$, rather than $q(x)$. To put this result in the context of Bayes' theorem, if $p(H|D) = \frac{1}{Z} p(D|H)\, p(H)$ and if we can sample from the prior $x \sim p(H)$, then we can generate samples from the posterior using Eq. (A.233), with $p^* = p(D|H)\, p(H)$.

One issue with transforming the sampling distribution is that when $p(x)$ is very different from $q(x)$, the resulting wide range of weights $\tilde{w}_i$ reduces the effective number of points sampled. Points with very little weight do not contribute much to the sum in Eq. (A.233). As a consequence, more samples are needed to make the effective number of samples of p equal to that from the original distribution q. When the new distribution is not too different from the old, the method can work well.

For further information on Monte Carlo methods, see Press et al. (2007) and MacKay (2003).

Problem A.8.2 Show that if $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I})$, then the best estimate $\hat{\boldsymbol{\theta}}$ of a linear fit is Gaussian distributed about the true values $\hat{\boldsymbol{\theta}}^*$, with variance $\sigma^2(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi})^{-1}$. Notice how the variance of the parameter estimates is proportional to the variance of the original data. Qualitatively, what happens if the measurement noise is colored, so that $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}_\xi)$, with a covariance matrix that has off-diagonal elements?

Problem A.8.3 Show that for an orthonormal basis in function space, $e_j \cdot e_k = \delta_{jk}$ that

$$\hat{\theta}_k = \boldsymbol{y} \cdot e_k \,, \qquad \text{or } \boldsymbol{y} = \sum_{k=1}^{K} (\boldsymbol{y} \cdot e_k) e_k \,, \tag{A.234}$$

justifying the interpretation of $\boldsymbol{P}^{(K)}$ as a projector matrix in Eq. (A.219).

Problem A.8.4 For the cost function $J(\hat{x}) = \langle |x - \hat{x}| \rangle = \int \mathrm{d}x \, |\hat{x} - x| \, p(x|y)$, show that minimizing J implies that $\hat{x}^*$ is the median of $p(x|y)$. Investigate $J''(\hat{x}^*)$, too.

Problem A.8.5 For the bottom "box-shaped" cost function with a small width Δ, show that minimizing $J(\hat{x})$ implies that $\hat{x}^*$ is the mode of $p(x|y)$.

Problem A.8.6 **Important details about importance sampling.** We fill in some gaps in our discussion of importance sampling. Assume a scalar variable x.

a. As a warmup, use Eq. (A.227) to show that, for N independent draws x^i from $p(x)$, if $\hat{\varphi} = \frac{1}{N} \sum_{i=1}^{N} \varphi_i$, then $\mathrm{Var}\, \hat{\varphi} = \frac{1}{N} \mathrm{Var}\, \varphi$.

b. Importance sampling estimates the average of $\varphi(x)$ over the distribution $p(x)$ using a second, proposal distribution $q(x)$. In general, p and q need not be normalized, but here we assume they are. Then, $\langle \varphi \rangle_p \approx \hat{\varphi} = \sum_{i=1}^{N} w_i \, \varphi(x^i)$, with weights $w_i = \frac{p(x^i)/q(x^i)}{\sum_i p(x^i)/q(x^i)}$. The x^i are N independent draws $x^i \sim q(x)$.

   Show that $\langle \hat{\varphi} \rangle = \langle \varphi \rangle_p$ and $\mathrm{Var}_q[\hat{\varphi}] = \frac{1}{N} (\langle \frac{p\varphi^2}{q} \rangle_p - \langle \varphi \rangle_p^2)$. Show that the variance vanishes when we pick $q(x) = p(x)\,\varphi(x)$, a choice that is valid only for $\varphi(x) > 0$.

c. *Bias of the ratio.* $\hat{\varphi}$ is biased for finite N. To see how this can arise, consider a case of N random variables n_i and d_i (numerator and denominator, in this case). Assume that they are correlated, as this is true for the importance-sampling case. (Why?) Let $\langle n \rangle$ and $\langle d \rangle$ be the mean values of the random variables (e.g., $\langle n \rangle = \int dn \, n \, p(n)$). Let bars denote arithmetic averages (e.g., $\overline{n} = \frac{1}{N} \sum_{i=1}^{N} n_i$). Then show $\left\langle \frac{\bar{n}}{\bar{d}} \right\rangle \neq \frac{\langle n \rangle}{\langle d \rangle}$ by writing $n_i = \langle n \rangle + \delta n_i$ and giving the lowest-order corrections.

A.9 Stochastic Processes

We can extend the notion of a random variable x to a sequence of random variables x_k, for example a *time series* of values measured at regular times, $x_k \equiv x(kT_s)$. Examples of such times series include the noise "sources" v and w from our discussion of Kalman filters (Chapter 8). The time series are Gaussian *white noise processes*, with zero mean and specified variance. That is, each random variable $x_k \equiv x \sim \mathcal{N}(\mu, \sigma^2)$ is drawn from a Gaussian distribution,

$$p(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left[-\frac{x^2}{2\sigma^2} \right] . \tag{A.235}$$

We are often interested in *noise terms* v_k and ξ_k, which are *stochastic processes*, or *time series*. For example, in Chapter 8, we consider dynamical systems with noise

added to the inputs (the *process noise* v_k) and to the outputs (the *measurement noise* ξ_k). At each time kT_s (or k, for short), one draws a noise term from a probability distribution (Gaussian in many cases). Every element in an ensemble would have a different actual number for the kth draw, and every repetition of a sequence of draws would also be different. We characterize the stochastic processes by characterizing their *statistics*. For example,

$$\begin{aligned} \langle v_k \rangle &= \langle \xi_k \rangle = 0 \\ \langle v_k \, v_{k'} \rangle &= \nu^2 \, \delta_{kk'} \,, \quad \langle \xi_k \, \xi_{k'} \rangle = \xi^2 \, \delta_{kk'} \,, \end{aligned} \tag{A.236}$$

where the Kronecker delta symbol

$$\delta_{kk'} = \begin{cases} 1 & \text{if } k = k' \,, \\ 0 & \text{otherwise} \,. \end{cases} \tag{A.237}$$

and where the angle brackets $\langle \cdots \rangle$ denote ensemble averages over the respective Gaussian probability distributions, $\mathcal{N}(0, \nu^2)$ and $\mathcal{N}(0, \xi^2)$.

Note that in Eq. (A.236), the second moment is the variance, since the mean is zero. Otherwise, subtract off its square. Note, too, that the variances ν^2 and ξ^2 are independent of time k. Time series whose moments (mean, variance, etc.) are constant are *stationary*. To denote a nonstationary time series, we would use ν_k^2 and ξ_k^2 (and would need to carefully distinguish those quantities from the square of the random variable at time k).

We can extend these definitions to continuous-time stochastic processes. For example, a one-dimensional stochastic process $v(t)$ has, for each time t, a probability distribution. We will often be interested in the case of Gaussian noise, where that probability distribution, $p(v, t)$ is Gaussian, as in Eq. (A.235), above. In such a case, all we need do is specify the moments. A common idealized case is *delta-correlated noise*, where the probability distributions from two different times are independent of each other. Formally, we can specify such a case using the notation

$$\langle v(t) \rangle = 0 \,, \qquad \langle v(t) \, v(t') \rangle = \delta \, (t - t') \,. \tag{A.238}$$

For a multivariable case $\boldsymbol{v}(t)$ is an n-dimensional vector, each of whose elements is a stochastic process, and we write

$$\langle \boldsymbol{v}(t) \rangle = \boldsymbol{0} \,, \qquad \langle \boldsymbol{v}(t) \, \boldsymbol{v}^{\mathsf{T}}(t') \rangle = \boldsymbol{Q}_v \, \delta \, (t - t') \,, \tag{A.239}$$

where the $n \times n$ matrix $\boldsymbol{Q}_v$ is diagonal when each element of $\boldsymbol{v}$ is an independent stochastic process. Off-diagonal elements give the covariance of element i with element j. In the mathematical literature, an alternate notation is usually used:

$$\boldsymbol{v}(t) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_v) \,, \tag{A.240}$$

with a mention that the $\boldsymbol{v}(t)$ are i.i.d. (independent, identically distributed) variables.

A.9.1 Independent Processes

The simplest kind of stochastic process is one in which the probability distributions governing each element are independent. For a discrete-time process X with realizations x_k, independence implies

$$p(x_1, x_2, \ldots, x_N) = p(x_1)p(x_2)\cdots p(x_N)\,. \tag{A.241}$$

If the distribution for $p(x_k)$ is independent of time k, the random variables are *i.i.d.*: independent and identically distributed and constitute a *purely random process*. Using our more compact notation, we can also write $p(x^N) = \prod_k p(x_k)$, with k ranging from 1 to N.

A.9.2 Markov Processes

Markov processes are an important special case of a stochastic process, where the present state alone determines the future. Here, a series of states, x_k, is a Markov process if

$$p(x_{k+1}|x^k) = p(x_{k+1}|x_k)\,. \tag{A.242}$$

In Eq. (A.242), the notation x^k is shorthand for the set $\{x_k, x_{k-1}, \ldots, x_1\}$. In other words, the probability distribution function for the random variable x_{k+1} depends only on the most recent state x_k and not on the previous states x^{k-1}.

Although the Markov assumption would seem to be very restrictive, many processes can be put in that form. A process at time k that depends on both $k-1$ and $k-2$ can be rewritten as a Markov process by enlarging the state vector to be the pair x_k and x_{k+1}. Thus, increasing the dimension of the state vector can turn a history-dependent process into a Markov one. In that sense, the question to ask is whether a finite-dimensional state vector suffices and, if so, whether the required dimension is small enough to allow useful calculations.

The *Chapman–Kolmogorov equation* is an important relation satisfied by Markov processes. Using the marginalization identity of Eq. (A.134) and the definition of conditional probability, Eq. (A.132), we can always write

$$p(x_2) = \int \mathrm{d}x_1\, p(x_2, x_1) = \int \mathrm{d}x_1\, p(x_2|x_1)p(x_1)$$

Similarly, we write

$$\begin{aligned} p(x_3, x_1) = p(x_3|x_1)p(x_1) &= \int \mathrm{d}x_2\, p(x_3, x_2, x_1) \\ &= \int \mathrm{d}x_2\, p(x_3|x_2, x_1)p(x_2, x_1) \\ &= \int \mathrm{d}x_2\, p(x_3|x_2)p(x_2|x_1)p(x_1)\,, \end{aligned} \tag{A.243}$$

where we use the Markov property that $p(x_3|x_2, x_1) = p(x_3|x_2)$. Dividing by $p(x_1)$ gives the Chapman–Kolmogorov equation for discrete-time processes,

$$p(x_3|x_1) = \int \mathrm{d}x_2\, p(x_3|x_2)p(x_2|x_1)\,. \tag{A.244}$$

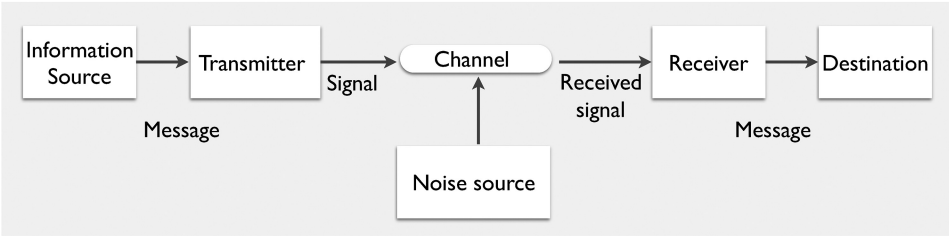
For the case where x_k can take only discrete values (e.g., Chapters 12 and 13), the integrals over x become sums.

A.10 Information Theory

We give a very quick review. Information theory was "born as an adult" in a remarkably clear and complete two-part paper from 1948 by Claude Shannon that is still well worth reading. For later developments, the standard text in the field is Cover and Thomas (2006), while Gibson (2014) is compact, accessible, and has an especially nice treatment of rate distortion theory. Papoulis and Pillai (2002) discusses many connections between information theory and signal processing. MacKay (2003) discusses connections to Bayesian inference and statistical physics. Bialek (2012) discusses biological applications.

Working at Bell Laboratories during World War II, Shannon's goal was to develop a general way to understand communication systems. His 1948 article is entitled "A Mathematical Theory of Communication,"[17] with no mention of "information." Shannon's abstract schematic diagram of a communication system is redrawn as Fig. A.7. The schematic has five elements:

1. *Information source*. Produces the message(s) to be communicated. For example, a set of English words. The words are then encoded into a standard alphabet. For example, each letter in the English alphabet is converted to a sequence of 0s and 1s.
2. *Transmitter*. The code is converted into a form suitable for transmission (e.g., voltage levels for 0 and 1).
3. *Channel*. The medium used to transmit the signal. For example, the message could be communicated electrically over wires, by radio waves, by photons over an optical



Abstract elements of a communication system. Adapted from Shannon (1948). Fig. A.7

[17] By the time the articles were reprinted a year later in book form (Shannon and Weaver, 1949), the title had changed to "The Mathematical Theory of Communication."

fiber, and so on. The channel is usually noisy: the output can differ, stochastically, from the input.
4. *Receiver*. Inverts the coding operation. The equipment could be an antenna for radio waves, a photodiode for an optical fiber, and so on.
5. *Destination*. The person (or thing) for whom the message is intended.

"Communication" is thus a very general concept. In this book, our most important example of it is measurement. That is, we will view the physical quantity we are trying to measure as the information source and the sensor as the noisy channel and receiver.

Here, we begin by defining and discussing briefly some basic quantities.

A.10.1 Entropy

Shannon viewed the transmission of a message across a noisy communication channel as a process of selecting from a set of possible messages. In the discrete case, the message is a random variable X that takes values ("outcomes") x that belong to a finite (or countably infinite) set $\mathcal{X} = \{x_i\}$, which is also called the *alphabet* and is similar to the event space of probability theory (Section A.6). For example, a binary signal may transmit a 0 or 1, a two-letter alphabet $\mathcal{X} = \{0, 1\}$. The notion of information then quantifies how much we learn – how "surprised" we are – when we receive a symbol. If, for example, a symbol $x = 0$ or 1 with equal probability – i.e., $P(x = 0) = P(x = 1) = \frac{1}{2}$ – then receiving a 0 or 1 allows us to answer one "yes or no" question. We define this to be a *bit* (= binary digit) of information. With n bits of information, we can answer n "yes or no" questions, selecting from among 2^n possibilities. Inspired by an earlier, informal effort by Hartley (1928), Shannon called the logarithm of the number of possibilities the *entropy* of the message. Here, $\log_2(2^n) = n$ and measures the *uncertainty* associated with a random variable. Conversely, it tells how much information is gained after measuring the quantity (neglecting measurement noise). Note that information, as defined here, has nothing to do with meaning or *semantics*.

Let $N = |\mathcal{X}|$ be the number of elements in the alphabet. Equally likely possibilities have probability $P = 1/N$ and entropy $\log N = \log(1/P)$. When events occur with different probabilities $P_X(x = x_i) \equiv P(x_i)$, the average entropy of the random variable X is

$$H(X) = \left\langle \log \frac{1}{P(x)} \right\rangle = -\sum_{x_i \in \mathcal{X}} P(x_i) \log P(x_i) \,. \tag{A.245}$$

Notice that the entropy is a function of the probability distribution $P(x)$. The logarithm base in Eq. (A.245) determines the units as conventionally base 2 (*bits*) or e (*nats*). The term "entropy" reflects a formal similarity to the entropy defined by Boltzmann and Gibbs in statistical physics, the difference being one of units, as reflected in the dimensional prefactor k_B (Boltzmann's constant). Below, we will see (Eq. A.280) that for any assignment of probabilities $P(x)$, the entropy is in the range $0 \le H(X) \le \log N$.

Shannon actually derived the form of Eq. (A.245) from a small number of axioms:

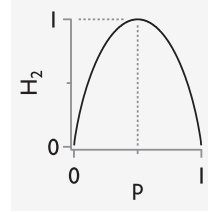
1. $H(X)$ is a continuous function of the probabilities $P(x_i)$.
2. If $P(x_i) = 1/N$, then $H(X)$ is a monotonically increasing function of N.
3. If a choice is broken down into two different series of sequential steps, the value of H should be consistent with either partition.

The full statement of the last axiom and the proof that they lead to a unique function is nicely discussed in Shannon's original 1948 paper. Pressé et al. (2013) discuss the derivation in a broad context. Here, we focus on the properties of entropy.

One of the powerful features of information theory is that it leads to proofs that hold for arbitrary probability distributions. Here, we mostly forego such general results and consider, instead, two archetypical examples: a discrete binary sequence and a Gaussian random variable, where the set of possible states is the real line (and hence a continuous variable that is uncountable). We begin by defining them and evaluating their entropy.

Example A.21 (Discrete Binary Sequence) Let the signal be a discrete binary sequence x_i, representing, say, two states of a system. We can label them $x = 0$ and $x = 1$ for convenience, but any two "letters" will work equally well. Then, if the probability of observing 0 is $P(x = 0) = p$ and of observing 1 is $P(x = 1) = 1 - p$, the entropy is

$$H_2(p) = -p \log p - (1 - p) \log(1 - p) . \tag{A.246}$$



The graph at left shows H_2 in units of bits. Notice that, as discussed above, the maximum 1 bit of uncertainty occurs at $p = \frac{1}{2}$, and the minimum value, 0, occurs at $p = 0$ or 1.

Differential entropy. When there are a continuous number of states, the alphabet $\mathcal{X}$ is a continuous set (for example, the real numbers). Let us define the "obvious" generalization to the continuous case, the *differential entropy*:

$$H(X) = -\int_{x \in \mathcal{X}} \mathrm{d}x \, p(x) \, \log p(x) . \tag{A.247}$$

To see why this definition is not quite the same thing as the "continuous version" of the discrete sum in Eq. (A.245), let us define a *coarse-grained* entropy by integrating over an interval Δx, with

$$P_i = \int_{i\Delta x}^{(i+1)\Delta x} \mathrm{d}x \, p(x) \approx P(x_i) \, \Delta x . \tag{A.248}$$

Then, at the discretization scale Δx, the coarse-grained entropy is

$$\begin{aligned} H_{\Delta x}(X) &= -\sum_i P_i \log P_i = -\sum_i P(x_i) \, \Delta x \log[P(x_i) \, \Delta x] \\ &= -\sum_i \Delta x \, P(x_i) \log P(x_i) - (1) \log \Delta x \\ &\to -\int \mathrm{d}x \, p(x) \log p(x) - \log \Delta x . \end{aligned} \tag{A.249}$$

Thus, the differential entropy differs from the coarse-grained approximation by a constant, $-\log \Delta x$, that diverges as $\Delta x \to 0$.

Example A.22 (Uniform distribution) If $p(x)$ is a uniform distribution between a and b, the differential entropy is

$$H(X) = + \int_a^b \mathrm{d}x \left(\frac{1}{b-a} \right) \log(b-a) = \log(b-a) \,. \tag{A.250}$$

Notice that the differential entropy is infinite if the range of x is infinite.

We can also now interpret the differential entropy more precisely: it is the difference between the entropy of the distribution $P(x)$ and that of a distribution that is uniform over $\mathcal{X}$. That is, the entropy for a continuous distribution is infinite – it takes an infinite number of bits or digits to specify an element completely and thus be certain of its identity – but the difference in entropy between two distributions may be finite.

A Gaussian distribution also has finite differential entropy.

Example A.23 (Gaussian distribution) Let the random variable X have values $x \sim \mathcal{N}(0, \sigma^2)$ that is a Gaussian random variable. The differential entropy of X is

$$H(X) = -\int_{-\infty}^{\infty} \mathrm{d}x \, p(x) \, \log p(x) = \left\langle \log\left(\sigma \sqrt{2\pi}\right) + \frac{x^2}{2\sigma^2} \log \mathrm{e} \right\rangle = \tfrac{1}{2} \log\left(2\pi \, \mathrm{e} \, \sigma^2\right) . \tag{A.251}$$

Up to constants, $H \sim \log \sigma$. It is easy to show that the entropy is the same if $\langle x \rangle = \mu$. Inverting Eq. (A.251) gives the variance in terms of the entropy:

$$\sigma^2 = \frac{1}{2\pi \, \mathrm{e}} \, \mathrm{e}^{2H(X)} \,. \tag{A.252}$$

Joint entropy. The definitions of discrete and continuous entropy functions generalize immediately to multiple random variables X_i with outcomes x_i, with an important case being the stochastic processes discussed in Section A.9. For the discrete case, we denote the set of random variables $\{X_1, X_2, \ldots, X_N\}$ by X^N, as in Section 15.2. Then,

$$H(X^N) = -\sum_{x^N} P(x^N) \, \log P(x^N) \,. \tag{A.253}$$

The N-fold sum is (implicitly) over the entire alphabet $\mathcal{X}$ for each random variable X_i. Likewise, the differential entropy is an N-dimensional integral:

$$H(X^N) = -\int_{x^N} \mathrm{d}x^N \, p(x^N) \, \log p(x^N) \,. \tag{A.254}$$

For example, if $X^N \equiv \boldsymbol{X}$ is a multivariate Gaussian with realizations $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then you will show in Problem A.10.2 that

$$H(\boldsymbol{X}) = \tfrac{1}{2} \log \det\left(2\pi \, \mathrm{e} \, \boldsymbol{\Sigma}\right) \,. \tag{A.255}$$

Conditional entropy. Consider the reduction in uncertainty, or gain of information, on measuring y, given that x is known. Let us revert to an explicit notation where the random variable X has outcomes $x \in \mathcal{X}$. Likewise, Y has outcomes $y \in \mathcal{Y}$, and sums or integrals of x and y are taken over their respective alphabets. Then,

$$H(Y|x) \equiv H(Y|X = x) = -\sum_y P(y|x) \log P(y|x)\,, \tag{A.256}$$

which is just the reduction in uncertainty, or gain of information, on measuring Y, given that another random variable X is known and has a particular value x.

The average reduction in uncertainty on measuring Y, averaging over X, is

$$\begin{aligned} H(Y|X) \equiv \langle H(Y|x)\rangle_x &= -\sum_{x,y} P(x)\, P(y|x) \log P(y|x) \\ &= -\sum_{x,y} P(x,y) \log P(y|x)\,. \end{aligned} \tag{A.257}$$

From Eq. (A.132), we also have

$$H(Y|X) = -\sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)} = H(X,Y) - H(X)\,. \tag{A.258}$$

Chain rule for entropy. Rearranging the terms in Eq. (A.258) gives

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)\,. \tag{A.259}$$

The latter relation comes from decomposing the conditional probability the "other way," $P(x,y) = P(x|y)P(y)$. Similarly, for three variables, the chain rule is

$$\begin{aligned} H(X,Y,Z) &= H(X) + H(Y,Z|X) \\ &= H(X) + H(Y|X) + H(Z|X,Y)\,. \end{aligned} \tag{A.260}$$

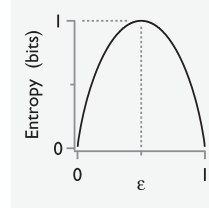
For N variables, it is

$$H(X^N) = \sum_{k=1}^{N} H(X_k|X^{k-1})\,. \tag{A.261}$$

Compare this relation with the chain rule for probabilities (Problem A.6.1).

Example A.24 (Two-state Markov chain) Let X be a random variable with alphabet $\mathcal{X} = \{0,1\}$, with transition probabilities $P(0|0) = P(1|1) = 1-\varepsilon$ and $P(0|1) = P(1|0) = \varepsilon$, as illustrated in the *state diagram* in Figure A.8. We can calculate the conditional entropy directly:

$$H(X_2|X_1) = -\sum_{x_1,x_2} P(x_1,x_2) \log P(x_2|x_1) = H_2(\varepsilon) \tag{A.262}$$



again illustrated at left. The sum for each x_i is over $\mathcal{X} = \{0,1\}$. The limits make sense: When $\varepsilon = 0$ or 1, the sequence is completely determined by its initial condition, and further observations do not reduce the uncertainty (i.e., they are certain). Thus,

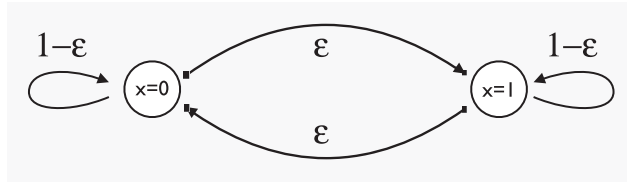


Fig. A.8 Two-state symmetric Markov chain.

$H(X_2|X_1) = 0$. On the other hand, when $\varepsilon = \frac{1}{2}$, there is a 50-50 chance that the next symbol will be 0 or 1, implying a $\log 2$ reduction in uncertainty per observation.

Changing Variables

If the entropy of X is $H(X)$, what is the entropy $H(Y)$ if $y = f(x)$? This question will have very different answers, depending on whether the instances x of the random variable X are drawn from a discrete or a continuous set.

Discrete case. Let us first assume that the function $f(x)$ is one-to-one, with a unique inverse. Then, every element x_i has a unique element y_i. Then $P(x_i) = P(y_i)$ and, hence, $H(Y) = H(X)$.

If $f(x)$ is not invertible, then at least two elements x_i and x_j map to the same new element y_k, implying that $H(Y) \leq H(X)$. Intuitively, there are "fewer possibilities" for the variable Y than for X and its uncertainty is therefore lower.

Continuous case. This case is completely different, for reasons that trace back to the definition of differential entropy itself, which depends on the choice of coordinates. For example, if we rescale $y = ax$, then $p(y) = p(x)\,|\,\mathrm{d}x/\mathrm{d}y\,| = p(x)/|a|$, and

$$H(Y) = -\int_y \mathrm{d}y\, p(y)\, \log p(y) = H(X) + \log|a|\,. \tag{A.263}$$

Compare this result to the discrete case, where rescaling is an invertible transformation and $H(Y) = H(X)$. Since $|a|$ can be less than 1, $\log|a|$ and hence $H(Y)$ can be negative. The absolute value in Eq. (A.263) comes from the transformation of probabilities, which must be ≥ 0. Lack of invariance also traces back to the fact that the probability density function $p(x)$ has units, whereas probabilities $P(x)$ do not.

As an example, consider the transformation $y = 2x$. Let a real number x have the binary expansion $x = 0.a_1a_2a_3\ldots$, where $a_i \in \{0, 1\}$. Then $y = 2x$ just shifts the binary digits one place over to the left:

$$x = 0.a_1a_2a_3\ldots \quad \Longrightarrow \quad y = a_1.a_2a_3a_4\ldots\,. \tag{A.264}$$

For $P(a_i = 0) = P(a_i = 1) = \frac{1}{2}$, at a given discretization scale, the binary expansion of y has one more bit (binary digit) than does x. Equivalently, the uncertainty in y is one bit greater than that of x, which is what the $\log|a| = 1$ term indicates.

For multiple variables $X^N = \{X_1, X_2, \ldots, X_N\}$, the differential entropy transforms as

$$H(Y^N) = H(X^N) + \langle \log |J| \rangle \,, \tag{A.265}$$

where J is the determinant of the *Jacobian* matrix of the coordinate transformation between the vectors X^N and Y^N, assuming the transformation to be invertible. In the special case $Y^N = \boldsymbol{A} X^N$, we have

$$H(Y^N) = H(X^N) + \log |\det \boldsymbol{A}| \,. \tag{A.266}$$

As in the discrete case, if the transformation is not invertible, then $H(Y^N) < H(X^N) + \langle \log |J| \rangle$.

A.10.2 Entropy Rate and Stochastic Processes

The entropy of a stochastic process is a special case of the notion of joint entropy, defined in Eq. (A.253). Since we expect the entropy to be proportional to the number of variables N (*extensive*), we define the *entropy rate* (or *density*)

$$\mathcal{H}(X) = \lim_{N\to\infty} \tfrac{1}{N} H(X^N) \,. \tag{A.267}$$

Clearly, $\mathcal{H}(X)$ is a kind of average entropy per variable and is most interesting when those variables are correlated; however, to build up our intuition, let us first consider N independent variables X_i. Then $P(x^N) = P(x_1)\, P(x_2) \ldots P(x_N)$ and

$$H(X^N) = -\sum_{x_i \in X} P(x_1) \ldots P(x_N) \sum_{i=1}^{N} \log P(x_i) = N\, H(X) \,. \tag{A.268}$$

The entropy of N independent, identically distributed variables is just N times the entropy of a single variable. In Eq. (A.268), there are N terms, each with N sums. In each sum, $N-1$ terms "go away," since $\sum_{x_i} P(x_i) = 1$. The entropy rate is then

$$\mathcal{H}(X) = \lim_{N\to\infty} \tfrac{1}{N} [N\, H(X^N)] = H(\boldsymbol{X}) \,. \tag{A.269}$$

As expected, the entropy rate of a sequence of uncorrelated, identically distributed random variables is just the entropy of a single random variable.

Example A.25 For a more interesting case, consider a stationary Markov process (Sec. A.9.2), where $P(x_{k+1}|x_k, x_{k-1}, \ldots, x_1) \equiv P(x_{k+1}|x^k) = P(x_{k+1}|x_k)$. Then

$$\begin{aligned} H(X^N) &= -\sum_{x_i \in X} P(x_N|x_{N-1}) \ldots P(x_2|x_1) P(x_1) \log \left[P(x_N|x_{N-1}) \ldots P(x_2|x_1) P(x_1) \right] \\ &= H(X_N|X_{N-1}) + \ldots + H(X_2|X_1) + H(X_1) \\ &= (N-1) H(X_2|X_1) + H(X_1) \\ &= (N-1) H(X_{+1}|X) + H(X) \,. \end{aligned} \tag{A.270}$$

In the last step, we introduce X_{+1} as the random variable one time step after X. Because the time series is stationary, $H(X_{k+1}|X_k) = H(X_{+1}|X)$ for any k. Taking the limit $N \to \infty$, we see that the entropy rate is simply $\mathcal{H}(X) = H(X_2|X_1)$. In other words, the average increase in entropy per measurement is the conditional entropy. Because of correlations among measurements, the conditional entropy is less than the entropy of a single measurement: $H(X_1|X_0) \leq H(X_1)$. Thus, $H(X^N)/N \to \mathcal{H}(X)$ from above, with corrections $O(N^{-1})$. Finally, see Section A.10.4 for a proof that conditioning reduces entropy.

The entropy rate for a general stationary Markov process $P(X_{k+1}|X_k)$ follows the arguments given in Example A.25, which leads to

$$\mathcal{H} = -\sum_{(x,x_{+1})\in\mathcal{X}} P(x)P(x_{+1}|x)\log P(x_{+1}|x)\,, \tag{A.271}$$

where $P(x_{k+1}|x_k) = P(x_{+1}|x)$ for any k and where the steady-state distribution $P(x)$ satisfies

$$P(x) = \sum_{x\in\mathcal{X}} P(x_{+1}|x)\,P(x)\,. \tag{A.272}$$

For an application, see Problem A.10.7.

Example A.26 Consider the transformation

$$y_k = \sum_{n=0}^{k} a_n x_{k-n}\,. \tag{A.273}$$

Since the random variables $y_0, \ldots, y_k$ are causally related to $x_0, \ldots, x_k$, the transformation matrix is lower triangular:

$$\boldsymbol{A} = \begin{pmatrix} a_0 & 0 & \cdots & 0 \\ a_1 & a_0 & \cdots & 0 \\ \vdots & & & \\ a_k & a_{k-1} & \cdots & a_0 \end{pmatrix}. \tag{A.274}$$

The determinant of $\boldsymbol{A}$ is easy to calculate: $\det \boldsymbol{A} = (a_0)^{k+1}$. The differential entropy is then $H(Y) = H(X) + (k+1)\log|a_0|$. Dividing by $k+1$ gives the entropy rate:

$$\mathcal{H}(Y) = \mathcal{H}(X) + \log|a_0|\,. \tag{A.275}$$

In words, the uncertainty typically increases or decreases by $\log|a_0|$ each observation, depending on whether $|a_0| > 1$ or < 1.

A.10.3 Relative Entropy

The *relative entropy*, or *Kullback–Leibler divergence*, is a useful measure of the "distance" between a probability distribution $Q(x)$ and a reference distribution $P(x)$:

$$D(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} . \tag{A.276}$$

Obviously, $D(P\|P) = 0$. In general, $D(P\|Q) \geq 0$ (Gibbs' inequality). To see this, we consider the convex function $f(u) = -\log u$ and apply Jensen's inequality, $\langle f(u) \rangle \geq f(\langle u \rangle)$, for $P \equiv P(x)$ and $Q \equiv Q(x)$:

$$D = -\sum_x P \log\left(\frac{Q}{P}\right) \geq -\log \sum_x P\left(\frac{Q}{P}\right) = -\log \sum_x Q = -\log 1 = 0 . \tag{A.277}$$

The quantity D is not quite a metric, as it is asymmetric between the two distributions – $D(P\|Q) \neq D(Q\|P)$, in general – and does not satisfy the triangle inequality.

The continuous version of the relative entropy is

$$D(p\|q) = \int_{\mathcal{X}} \mathrm{d}x\, p(x) \log \frac{p(x)}{q(x)} . \tag{A.278}$$

Unlike differential entropy, the relative entropy between continuous distributions is invariant under a coordinate transformation $y = f(x)$. The Jacobians cancel in the ratio of p/q, and we transform probabilities by $p(x)\, \mathrm{d}x = p(y)\, dy$. See Section A.6.5.

As an application, $H(X) \leq \log N$, where $N = |\mathcal{X}|$ is the number of elements in the alphabet $\mathcal{X}$. To see this, let $Q(x) = 1/N$ and apply the Gibbs' inequality:

$$D(P\|Q) = \sum_x P(x) \log \frac{P(x)}{1/N} = -H(X) + \log N \geq 0 . \tag{A.279}$$

We can strengthen this conclusion. Complete certainty means knowing that a measurement will lead to a particular outcome, for example, x_j. Thus, $P(x_i) = \delta_{ij}$, which implies that $H(X) = 0$. Further, for discrete probabilities, $0 < P(x_i) < 1$, since probabilities are positive and must sum to one: $\sum_i P(x_i) = 1$. Then $\log 1/P(x_i) > 0$ and $H = \langle \log 1/P \rangle$ must also be positive. Thus,

$$0 \leq H(X) \leq \log N . \tag{A.280}$$

The relative entropy can also be interpreted as the amount of information lost when the "wrong" probability distribution (Q) is used to interpret events that actually come from P.

Example A.27 (Counting and probabilities) Let p_i be the probability that event i occurs. For N trials and K possible outcomes, you observe n_i counts of event i, with $\sum_{i=1}^{K} n_i = N$. We claim that $\hat{p}_i \equiv n_i/N$ is the maximum likelihood estimate of p_i. The slick proof is based on relative entropy. We use the notation $P(n_i; q_i)$ to denote the likelihood of observing n_i counts given the estimator q_i of the probability p_i.

We need to show that $P(n_i;\hat{p}_i) > P(n_i;q_i)$ for all numbers $q_i \neq \hat{p}_i$, or, equivalently, that $\log P(n_i;\hat{p}_i)/P(n_i;q_i) > 0$. Independent events of probability p_i occur n_i times with probability $\propto p_i^{n_i}$. Then, with $n = \{n_i\}$, $\hat{p} = \{\hat{p}_i\}$, and $q = \{q_i\}$, we have

$$\log \frac{P(n;\hat{p})}{P(n;q)} = \log \frac{\prod_i \hat{p}_i^{n_i}}{\prod_i q_i^{n_i}} = \sum_i n_i \log \frac{\hat{p}_i}{q_i} = N \sum_i \hat{p}_i \log \frac{\hat{p}_i}{q_i} = ND(\hat{p},q) \geq 0\,, \qquad \text{(A.281)}$$

with equality if and only if $q_i = \hat{p}_i$. In words, in the absence of any prior information, the most likely estimate of p_i is just its empirical frequency of occurrence, n_i/N. This example is from Durbin et al. (1998).

Example A.28 (Relative entropy and maximum likelihood) The Kullback–Leibler divergence also leads to an interesting interpretation of the use of maximum likelihood in curve fitting. To see this in the one-parameter case, let $P_\theta(x)$ be the probability distribution of a discrete variable x, parameterized by θ. Then, in nats,

$$\begin{aligned} D(P_\theta||P_{\theta+\Delta\theta}) &= \sum_x P_\theta(x) \ln \frac{P_\theta(x)}{P_{\theta+\Delta\theta}(x)} \\ &= \sum_x P_\theta(x) \ln P_\theta(x) - \sum_x P_\theta(x) \ln P_{\theta+\Delta\theta}(x)\,. \end{aligned} \qquad \text{(A.282)}$$

We can expand the latter term in a Taylor series,

$$\ln P_{\theta+\Delta\theta}(x) = \ln P_\theta(x) + \frac{\partial \ln P_\theta}{\partial\theta}\Delta\theta + \frac{1}{2}\frac{\partial^2 \ln P}{\partial\theta^2}(\Delta\theta)^2 + \cdots\,. \qquad \text{(A.283)}$$

The $\Delta\theta$ term vanishes when averaged. To see this, we write

$$\begin{aligned} \sum_x P_\theta(x)\frac{\partial \ln P_\theta}{\partial\theta}\Delta\theta &= \sum_x P_\theta(x)\frac{1}{P_\theta(x)}\frac{\partial P_\theta}{\partial\theta}\Delta\theta \\ &= \sum_x \frac{\partial P_\theta}{\partial\theta}\Delta\theta \\ &= \frac{\partial}{\partial\theta}\left(\sum_x P_\theta(x)\right)\Delta\theta = \left[\frac{\partial}{\partial\theta}(1)\right]\Delta\theta = 0\,. \end{aligned} \qquad \text{(A.284)}$$

We can then evaluate Eq. (A.282):

$$\begin{aligned} D(P_\theta||P_{\theta+\Delta\theta}) &\approx \sum_x P_\theta(x) \ln P_\theta(x) - \sum_x P_\theta(x) \ln P_{\theta+\Delta\theta}(x) \\ &= \sum_x P_\theta(x) \ln P_\theta(x) - \sum_x P_\theta(x) \ln P_\theta(x) + 0 - \frac{1}{2}\sum_x P_\theta(x)\frac{\partial^2 \ln P_\theta}{\partial\theta^2}(\Delta\theta)^2 \\ &= -\frac{1}{2}\left\langle \frac{\partial^2 \ln P_\theta}{\partial\theta^2}\right\rangle (\Delta\theta)^2\,. \end{aligned} \qquad \text{(A.285)}$$

The quantity $-\langle\partial_{\theta\theta} \ln P_\theta\rangle = +\langle(\partial_\theta \ln P_\theta)^2\rangle \geq 0$ is known as the *Fisher information*. (The identity is proved by applying the chain rule to the second derivative.) It quantifies how measurements of x reduce the uncertainty in a parameter θ. Combining this result with

extensions of Example A.27 leads to the conclusion that maximizing the likelihood in a curve fit minimizes the relative entropy between the probability distributions based on the inferred parameter ($\theta + \Delta\theta$) and one based on the "true" parameter value θ (*Sanov's theorem*).

A.10.4 Mutual Information

The *mutual information*, $I(X;Y)$ indicates how much, on average, measuring Y reduces the uncertainty in X. More precisely,

$$I(X;Y) \equiv H(X) - H(X|Y)\,. \tag{A.286}$$

One of the motivations for this definition is to account for measurement noise. For a noiseless measurement, the physical quantity X is deterministically related to the measurement Y, implying that $H(X|Y) = 0$: there is no uncertainty in X after the measurement. In that limit, $I(X;Y) = H(X)$. In the other limit where measurement noise dominates, $H(X|Y) \approx H(X)$: the measurement is so bad that it does not reduce the uncertainty in X at all. In this case $I(X;Y) = 0$, and we say that, on average, the measurement Y does not give any information about the physical quantity X. In between these extremes, $I(X;Y)$ tells us the amount we learn about X after making a noisy measurement Y.

Let us establish a few properties of the mutual information:

- $I(X;Y) = I(Y;X)$. Mutual information is symmetric between input and output. What X tells about Y is the same as what Y tells about X. To see this,

$$\begin{aligned} I(X;Y) &= H(X) + \sum_{x,y} P(x,y)\log P(x|y) = H(X) + \sum_{x,y} P(x,y)\log\frac{P(x,y)}{P(y)} \\ &= H(X) + H(Y) - H(X,Y)\,, \end{aligned} \tag{A.287}$$

  which is manifestly symmetric in X and Y. Expanding all the entropy terms gives

$$I(X;Y) = \sum_{x,y} P(x,y)\log\frac{P(x,y)}{P(x)\,P(y)}\,. \tag{A.288}$$

- $I(X;Y) = H(Y) - H(Y|X)$. This is a simple consequence of symmetry, with $X \leftrightarrow Y$.
- $I(X;X) = H(X)$. Entropy can be viewed as "self-information."
- $I(X;Y) = D[P(x,y)\|P(x)P(y)]$. This follows from interpreting Eq. (A.288) as an expression for relative entropy.
- $I(X;Y) \geq 0$, with $I = 0$ only when X and Y are independent random variables. This follows from the Gibbs inequality, $D \geq 0$. Two useful corollaries follow immediately:
  - $H(X|Y) \leq H(X)$. This follows from $I(X;Y) = H(X) - H(X|Y)$ and $I(X;Y) \geq 0$. On average, conditioning can only decrease uncertainty. Curiously, an individual measurement *can* increase the uncertainty about X. See Problem A.10.6.
  - $H(X^N) \leq \sum_i H(X_i)$. This follows from the previous result and the chain rule for entropy.

- $I(X;Y) \leq \min\,[H(X), H(Y)]$. This follows from the definition of $I(X;Y)$ and the positivity of entropies. Thus, a measurement Y cannot tell us more about X than its a priori uncertainty. Likewise, we cannot learn more about X than the entropy contained in the measurement (as might be deduced from looking at a long series of measurements). Putting the last two results together, we have

$$0 \leq I(X;Y) \leq \min[H(X), H(Y)]\,. \tag{A.289}$$

- *Data-processing inequality*. Consider the random variables $X \to Y \to Z$, by which we imply a Markov chain where X affects Y and Y affects Z. But X does not directly affect Z. In probabilities, this last statement implies that $P(z|x,y) = P(z|y)$, independent of x. The data-processing inequality then states that $I(X;Z) \leq \min[I(X;Y), I(Y;Z)]$:

$$\begin{aligned} I(X;Z) = H(Z) - H(Z|X) &\leq H(Z) - H(Z|X,Y) \\ &= H(Z) - H(Z|Y) = I(Y;Z) \end{aligned} \tag{A.290}$$

  A similar argument shows $I(X;Z) \leq I(X;Z)$ (it also requires showing $X \to Y \to Z \implies Z \to Y \to X$). Thus, any kind of data manipulation at best preserves information; it can never add information. Extending the argument to a chain $X_1 \to X_2 \to \cdots \to X_N$ shows that the mutual information is bounded by the weakest link: $I(X_1;X_N) \leq \min_i I(X_i;X_{i+1})$, which can also be viewed as a *bottleneck*.

- *Continuous case*. The definition of mutual information is the obvious extension:

$$I(X;Y) = \int_{x\in\mathcal{X},\, y\in\mathcal{Y}} \mathrm{d}x\,\mathrm{d}y\; p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}\,. \tag{A.291}$$

  Unlike the differential entropy, the continuous version of mutual information is coordinate invariant. That is, if we transform $x \to x'$ and $y \to y'$, the Jacobian factors cancel out in the log term, since numerator and denominator transform in the same way. Coordinate invariance gives the mutual information absolute meaning: it tells us what we learn from a noisy measurement of a physical quantity. An alternate form for $I(X;Y)$ that is sometimes useful is

$$I(X;Y) = \int_{x\in\mathcal{X}} \mathrm{d}x\; p(x) \int_{y\in\mathcal{Y}} \mathrm{d}y\; p(y|x)\, \log \frac{p(y|x)}{p(y)}\,. \tag{A.292}$$

- *Gaussian signals and noise*. Consider a signal with additive Gaussian noise. If $y = x + \xi$, with signal $x \sim \mathcal{N}(0, \sigma_x^2)$ and noise $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$, then Eq. (A.291) implies,

$$I(X;Y) = \tfrac{1}{2} \log\left(1 + \mathrm{SNR}^2\right) \equiv \tfrac{1}{2} \log\left(1 + \frac{\sigma_x^2}{\sigma_\xi^2}\right). \tag{A.293}$$

See Problem A.10.11. We can define a signal-to-noise ratio (SNR) = σ_x/σ_ξ.[18] In the limit $\sigma_x \ll \sigma_\xi$ (or SNR $\ll 1$), the measurement tells us nothing about the signal, and the mutual information $I \approx 0$. When $\sigma_x \gg \sigma_\xi$ (or SNR $\gg 1$), the mutual information $I \approx \log\left(\sigma_x/\sigma_\xi\right) \equiv \log n$, where n is the number of possible measurements, given the noise level, that we can make of the continuous signal. Then $I \approx \log n$, as expected for a "discrete" system of n equally likely states.

Equation (A.293) is a celebrated result, so much so that it is worth emphasizing that it is a particular case – a linear measurement with additive Gaussian noise. The mutual information for nonlinear measurements and for other types of noise is different.

- *Correlated time series, part I.* Consider N noisy measurements $Y^N = \{y_1, y_2, \ldots, y_N\}$ of N variables $X^N = \{x_1, x_2, \ldots, x_N\}$, taken at intervals T_s, with $T = NT_s$, and related by $y_k = \sum_n G_{kn} x_n + \xi_k$. The random variables X^N are multivariate Gaussian, with zero mean and covariance $(\boldsymbol{S}_{xx})_{ij} = \langle x_i\, x_j \rangle$. Similarly, measurement noise is correlated, with $\xi^N \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Xi})$. If the ξ^N are uncorrelated with the states X^N, then Eq. (A.293) generalizes to (Problem A.10.12),

$$I(X^N; Y^N) = \frac{1}{2} \mathrm{Tr} \log \left[\mathbb{I} + \boldsymbol{\Xi}^{-1} (\boldsymbol{G}\, \boldsymbol{S}_{xx}\, \boldsymbol{G}^{\mathsf{T}}) \right] . \tag{A.294}$$

Equation (A.294) simplifies in the frequency domain, where each frequency component is independent of the others. After discrete Fourier transformation, each frequency ω_k acts as an independent Gaussian channel, with a signal-to-noise ratio $\mathrm{SNR}^2(\omega) = |G|^2(\omega)\, S_{xx}(\omega)/\Xi(\omega)$. The mutual information sums these independent contributions:

$$\begin{aligned} I(X^N; Y^N) &= \frac{1}{2} \sum_\omega \log\left(1 + \frac{|G|^2(\omega) S_{xx}(\omega)}{\Xi(\omega)}\right) \to \frac{T}{2} \int \frac{\mathrm{d}\omega}{2\pi} \log\left(1 + \frac{|G|^2(\omega) S_{xx}(\omega)}{\Xi(\omega)}\right) \\ &= \frac{T}{2} \int \frac{\mathrm{d}\omega}{2\pi} \log\left[1 + \mathrm{SNR}^2(\omega)\right] , \end{aligned} \tag{A.295}$$

for $T \to \infty$. The frequency limits are given by the Nyquist frequency, $\omega_N = \pm \frac{\pi}{T}$, which go to $\pm\infty$ for a continuously sampled signal. The $\mathrm{SNR}^2(\omega) \equiv |G|^2(\omega) S_{xx}(\omega) \,/\, \Xi(\omega)$, the signal-to-noise ratio at frequency ω. Equation (A.295) generalizes the result in Eq. (A.293) for $I(X;Y)$, since each Fourier component is statistically independent of the others. The time series X and noise terms must have stationary time series, and the time-domain response function $G(t)$ must be time-translation invariant.

- *Mutual information rate.* We can define the rate at which information is communicated between X and Y by analogy with the definition of entropy rate:

$$\mathcal{I}(X;Y) \equiv \lim_{N\to\infty} \frac{1}{N} I(X^N;\, Y^N) . \tag{A.296}$$

[18] As always, we define SNR as an *amplitude* ratio. Some authors define it to be the *power* ratio σ_x^2/σ_ξ^2.

By substituting the definitions of individual mutual-information functions in Eq. (A.296) in terms of entropy functions, we can easily see that $\mathcal{I}(X;Y) = \mathcal{H}(X) - \mathcal{H}(X|Y) = \mathcal{H}(Y) - \mathcal{H}(Y|X)$. In other words, an information-theory identity in terms of entropies and mutual information terms can be transformed into a corresponding rate identity (e.g., by setting $H \to \mathcal{H}$). Dividing by N in Eq. (A.296) gives the rate per observation, while dividing by $T = NT_s$ would give the rate per time. We will use both conventions.

In the Gaussian example above, the information acquired per unit time is,

$$\mathcal{I}(X;Y) = \lim_{T\to\infty} \frac{1}{T} I(X;Y) = \frac{1}{2} \int \frac{\mathrm{d}\omega}{2\pi} \log\left[1 + \mathrm{SNR}^2(\omega)\right] . \tag{A.297}$$

If Y^N is a deterministic and invertible function of X^N, then $H(Y^N|X^N) = 0$. That is, if we know the X^N, then we know the Y^N, as well. As a consequence, $\mathcal{I}(X;Y) = \mathcal{H}(Y)$.

In Example A.26, $y_k = \sum_{n=0}^{k} a_n x_{k-n}$, implies that $\mathcal{I}(X;Y) = \mathcal{H}(X) + \log|a_0|$.

- *Correlations*. Mutual information generalizes the notion of linear correlation. For example, if the random variables $\{X, Y\}$ are stationary, correlated processes, distributed as a bivariate Gaussian with correlation coefficient $\rho = \sigma_{xy}/(\sigma_x \sigma_y)$ (Section A.7.4), then

$$I(X;Y) = \log\left(\frac{1}{\sqrt{1-\rho^2}}\right) = -\frac{1}{2}\log\left(1-\rho^2\right). \tag{A.298}$$

See Problem A.10.8. For $\rho = 0$ (independent variables) $I = 0$, as expected. In the deterministic limit $|\rho| \to 1$, the mutual information diverges ($I \to \infty$) because a measurement of y determines the value of x, and a real number takes an infinite amount of information to describe completely. (Its decimal expansion has an infinite number of digits.)

The conventional correlation coefficient ρ does not do a good job of quantifying the dependence of variables that are nonlinearly related. For example, if $y = x^2$ and $x \sim \mathcal{N}(0,1)$, then $\rho = 0$ but $I(X;Y) > 0$. That is, the linear correlation coefficient wrongly suggests that the two variables are independent, while the mutual information correctly quantifies the relation. (See Problem A.10.10.)

- *Correlated time series, part II*. Equation (A.298) gives $I(X;Y)$ for two Gaussian random variables. Following the logic that led from Eq. (A.293) to Eq. (A.297), we can calculate $\mathcal{I}(X;Y)$ between two correlated, stationary Gaussian processes by integrating over frequency:[19]

$$\mathcal{I}(X;Y) = -\frac{1}{2} \int \frac{\mathrm{d}\omega}{2\pi} \log\left[1 - \rho^2(\omega)\right] , \qquad \rho^2(\omega) \equiv \frac{|S_{xy}|^2}{S_{xx}(\omega)\, S_{yy}(\omega)} , \tag{A.299}$$

[19] For example, see Munakata and Kamiyabu (2006) or, for more rigor, Liptser and Shiryaev (2000).

where the *squared coherence function* $\rho^2(\omega)$ is equivalent to the bivariate correlation coefficient ρ^2 for each frequency component, with $S_{xx} = \langle |x|^2(\omega) \rangle$ the power spectrum of the input signal $x(t)$ and $S_{yy} = \langle |y|^2(\omega) \rangle$ the power spectrum of the output signal $y(t)$. The *cross spectral density* $S_{xy}(\omega) \equiv \langle x(\omega)\, y^*(\omega) \rangle$ is also the Fourier transform of the covariance function $\langle x(t)\, y(t) \rangle$. As with the correlation coefficient ρ^2, the quantity $\rho^2(\omega) \in [0, 1]$ for each frequency. We see that

$$\begin{aligned} &x(t) \text{ and } y(t) \text{ are independent signals} \quad &\Longrightarrow \quad &\rho^2 = 0 \text{ and } \mathit{I}(X;Y) = 0. \\ &x(t) \text{ completely determines } y(t) \quad &\Longrightarrow \quad &\rho^2 = 1 \text{ and } \mathit{I}(X;Y) \to \infty. \end{aligned}$$

Although Eqs. (A.297) and (A.299) for $\mathit{I}(X;Y)$ look rather different, they are equivalent when signal and noise are uncorrelated at equal times.

- *Chain rule for mutual information*. For the time series X^N and Y^N,

$$I(X^N; Y^N) = \sum_{k=1}^{N} I(X_k; Y^N | X^{k-1}) \,. \tag{A.300}$$

This follows from the chain rule for entropy, Eq. (A.261). We write

$$\begin{aligned} I(X^N; Y^N) &= H(X^N) - H(X^N|Y^N) \\ &= \sum_{k=1}^{N} H(X_k|X^{k-1}) - \sum_{k=1}^{N} H(X_k|X^{k-1}, Y^N) = \sum_{k=1}^{N} I(X_k\,; Y^N|X^{k-1}) \,. \end{aligned} \tag{A.301}$$

Note that Y may have a different number of elements than X. Also, the symmetry of I implies a similar expansion with $X \leftrightarrow Y$.

A.10.5 Some Fundamental Theorems

Much of the interest historically in information theorem is due to Shannon's three fundamental theorems on *source coding* (a long sequence of N independent, identically distributed random variables with entropy $H(X)$ cannot be compressed into fewer than $NH(X)$ bits without loss of information), *channel coding* (information can be sent without loss up to a rate known as the channel capacity), and *rate distortion*, which sets bounds on the amount of loss (distortion) that occurs if information is sent at a rate higher than the channel capacity. These topics are mostly outside our purview, and you should consult standard references such as Cover and Thomas (2006) or MacKay (2003). We do, however, make use of the concept of channel coding, which we introduce below.

Channel Coding

We first define the notion of a communication channel, following the careful treatment of Massey (1990). A *discrete* channel has a source (input) X with finite alphabet $\mathcal{X}$. When used N times, the input distribution is $P(x^N)$. Similarly, the receiver (output

Y with finite output alphabet $\mathcal{Y}$) has output distribution $P(y^N)$. The channel is *memoryless* if, for each k (with $1 \le k \le N$), we have

$$P(y_k|x^k, y^{k-1}) = P(y_k|x_k)\,. \tag{A.302}$$

The channel is used *without feedback* if

$$P(x_k|x^{k-1}, y^{k-1}) = P(x_k|x^{k-1})\,. \tag{A.303}$$

Note that $P(x_k)$ commonly does depend on previous values, x^{k-1}. For example, if an English-language text transmits "b,u,z," the next letter is more likely to be "z" than "e", even though the overall probability of receiving an "e" is higher than "z."

If a channel is memoryless *and* used without feedback, then (Problem A.10.13)

$$P(y^N|x^N) = \prod_{k=1}^{N} P(y_k|x_k)\,. \tag{A.304}$$

The output y_k then depends *only* on the input x_k and no other value of x or y. Notice that our standard measurement relation for a linear system, $y_k = \boldsymbol{C}\boldsymbol{x}_k + \xi_k$, defines a DMC. We can think of measurement as a communication between a system and its measuring device.

We now define the *capacity* C of a DMC as

$$C \equiv \max_{P(x)} I(X;Y)\,. \tag{A.305}$$

In other words, we evaluate $I(X;Y)$ for each possible input distribution $P(x)$ of a DMC. The capacity is given by choosing $P(x)$ to maximize the mutual information between source and receiver. Shannon's theorem then states that it is possible to transmit information, with arbitrarily small probability of error at a rate up to the channel capacity. The theorem is remarkable, as we would naively expect a crossover between regimes of small and large transmission errors. But no: for long messages, at rates less than C, there is no error; at rates greater than C, there will be errors. For a DMC, it turns out that using feedback does not increase the capacity.

Problem A.10.1 **Entropy of a function of a stochastic process**. For any deterministic function $f(\cdot)$ and discrete random variable X, show that $H[f(X)] \le H(X)$. Hint: Apply the chain rule to $H[X, f(X)]$.

Problem A.10.2 **Entropy of a multivariate Gaussian distribution**. The N-dimensional vector $\boldsymbol{X}$ has realizations $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Show that $H(\boldsymbol{X}) = \frac{1}{2}\log\det(2\pi\,\mathrm{e}\,\boldsymbol{\Sigma})$.

Problem A.10.3 **Chain rule for relative entropy**. Show that $D[p(x,y) \parallel q(x,y)] = D[p(x) \parallel q(x)] + D[p(y|x) \parallel q(y|x)]$ – i.e., *coarse graining reduces relative entropy*.

Problem A.10.4 **Relative entropy for Gaussians**.

a. If $x_1 \sim \mathcal{N}(\mu_1, \sigma^2)$ and $x_2 \sim \mathcal{N}(\mu_2, \sigma^2)$, show $D(p_{x_1}\|p_{x_2}) = (\mu_1 - \mu_2)^2/2\sigma^2$.

b. If $x_1 \sim \mathcal{N}(0, \sigma_1^2)$ and $x_2 \sim \mathcal{N}(0, \sigma_2^2)$, show $D(p_{x_1}\|p_{x_2}) = (\sigma_1^2 - \sigma_2^2)/(2\sigma_2^2) + \log\frac{\sigma_2}{\sigma_1}$. Verify that $D(p_{x_1}\|p_{x_2})$ is nonnegative.

Problem A.10.5 Relative entropy and Fisher information. Consider a one-parameter family of probability distributions $P_\theta(x)$ over a discrete variable x. Show that the relative entropy between $P_\theta(x)$ and $P_{\theta+\Delta\theta}(x)$ is given, to lowest order in $\Delta\theta$, by

$$D(P_\theta \| P_{\theta+\Delta\theta}) \approx \frac{(\Delta\theta)^2}{2} \left\langle \left(\frac{d \log P_\theta}{\mathrm{d}\theta} \right)^2 \right\rangle \equiv \frac{(\Delta\theta)^2}{2} F \,,$$

where F is the *Fisher information*.

Problem A.10.6 Numerical example. Consider two variables X and Y with two-letter alphabets $\{1, 2\}$. The joint probabilities $P(x, y)$ are given in the table at left. Calculate

$P(x,y)$	$x=1$	$x=2$
$y=1$	0.3	0.3
$y=2$	0.4	0

a. the marginal distributions $P(x)$ and $P(y)$;
b. the entropies $H(X)$, $H(Y)$, and $H(X, Y)$;
c. conditional entropies $H(Y|x = 1)$ and $H(Y|x = 2)$. Show that $H(Y|x = 1) > H(Y)$. What does this mean?
d. the average conditional entropy $H(Y|X)$. Show that $H(Y|X) < H(Y)$, and reconcile this result with the previous part.
e. the mutual information $I(X; Y)$.

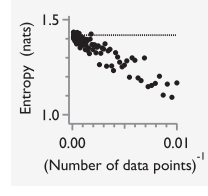
Problem A.10.7 Entropy rate of a Markov chain. Generalize the result for the entropy rate of a symmetric 2-state Markov chain (Eq. A.262) to the asymmetric case, with transition probabilities α and β, using Eq. (A.271).

a. Show that $\mathcal{H}(X) = \left(\frac{\beta}{\alpha+\beta}\right) H_2(\alpha) + \left(\frac{\alpha}{\alpha+\beta}\right) H_2(\beta)$.
b. Which values of (α, β) maximize and which minimize $\mathcal{H}(X)$?
c. What is wrong with the following argument: For $\alpha = \beta = \varepsilon$, the single-symbol entropy is 1 bit. The time spent in each state before jumping to the other is typically ε^{-1}. Therefore, the entropy rate should be $\mathcal{H} \approx 1$ bit$/(\varepsilon^{-1}) = \varepsilon$ bits/trial.

Problem A.10.8 Mutual information of a bivariate Gaussian. Derive Eq. (A.298).

Problem A.10.9 Estimating entropy from limited experimental data is tricky.

a. Simulate N draws $x \sim \mathcal{N}(0, 1)$ of the random variable X, and histogram the results. Let N_i be the number of data points in bin i and let $f_i = N_i/N$ be the corresponding frequency estimates. If p_i are the exact frequencies, then $H = -\sum p_i \ln p_i$ and a "naive" estimator is $\hat{H} = -\sum f_i \ln f_i$, in nats. Confirm the plot of $\hat{H}$ versus N^{-1} at right, where the dashed line is the expected result, $\frac{1}{2} \ln 2\pi \mathrm{e} \approx 1.42$. Use 100 bins, over $x \in [-5, 5]$. The linearity of the plot implies an $O(1/N)$ bias. The bias is large: more than 20% for 100 points.



b. Use Jensen's inequality to show that the bias is general: $\langle \hat{H} \rangle \leq H$.
c. Taylor expand $\hat{H}$ about the exact H. Assuming that N_i is Poisson distributed about the expected value, Np_i, show that $\langle \hat{H} \rangle = H(X) - (1/2)(N_{\mathrm{bin}}/N) + O(N^{-2})$, which has the $O(N^{-1})$ dependence noted in part (a). An obvious way to deal with finite-N bias is to extrapolate N^{-1} to zero, as shown in the figure. Appendix A.8 of Bialek (2012) discusses more sophisticated estimators with smaller bias.

Problem A.10.10 **Mutual information for a nonlinear relationship**. Let $y = x^2$ and let $x \sim \mathcal{N}(0, 1)$. Show the following:

a. The linear correlation coefficient $\rho = 0$.

b. $I(X;Y) \approx 1.13$ bits. Why is $I < H(X) \approx 2.05$ bits? Hint: Show that $H(Y|X) = 0$.

Problem A.10.11 **Noisy measurements and Gaussian channels**. Derive Eq. (A.293).

Problem A.10.12 **Measurements of noisy, correlated variables**. Derive Eq. (A.294). Hint: Follow Problem A.10.11, and use Tr log = log det (see Problem A.1.21).

Problem A.10.13 **Discrete memoryless channel (DMC)**. Prove Eq. (A.304).

Problem A.1.14 Show that if $\boldsymbol{A}$ is symmetric, then so is $\boldsymbol{A}^{-1}$.

Solution.

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{A}^{-1} &= \mathbb{I} \\
\left(\boldsymbol{A}^{-1}\right)^{\mathsf{T}} \boldsymbol{A}^{\mathsf{T}} &= \mathbb{I} && \text{take transpose} \\
\left(\boldsymbol{A}^{-1}\right)^{\mathsf{T}} \boldsymbol{A} &= \mathbb{I} && \boldsymbol{A} \text{ is symmetric} \\
\left(\boldsymbol{A}^{-1}\right)^{\mathsf{T}} &= \boldsymbol{A}^{-1}. && \text{take inverse on right}
\end{aligned}$$

Problem A.1.15 Verify the Sherman–Morrison formula, Eq. (A.15).

Solution.

$$\begin{aligned}
\left(\boldsymbol{A}+\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}\right)^{-1}\left(\boldsymbol{A}+\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}\right) &= \left(\boldsymbol{A}^{-1}-\frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}}{1+\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{u}}\right)\left(\boldsymbol{A}+\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}\right) \\
&= \mathbb{I}+\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}-\frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}+\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}}{1+\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{u}} \\
&= \mathbb{I}+\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}-\frac{\boldsymbol{A}^{-1}\boldsymbol{u}\cancel{\left(1+\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{u}\right)}\boldsymbol{v}^{\mathsf{T}}}{\cancel{1+\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{u}}} \\
&= \mathbb{I}+\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}-\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}} \\
&= \mathbb{I}.
\end{aligned}$$

The above is a verification, not a proof, as we assumed we knew the formula already!

Problem A.1.16 Show $\frac{\partial^2}{\partial \boldsymbol{x}^{\mathsf{T}}\partial \boldsymbol{x}}\left(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x}\right)=\boldsymbol{A}+\boldsymbol{A}^{\mathsf{T}}$; $\frac{\partial}{\partial \boldsymbol{x}}\mathrm{Tr}\left(\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\right)=2\boldsymbol{x}^{\mathsf{T}}$, and $\frac{\partial}{\partial \boldsymbol{x}}\left(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{y}\right)=\boldsymbol{y}^{\mathsf{T}}$.

Solution.

a. The first identity follows from Eq. (A.39):

$$\frac{\partial^2}{\partial \boldsymbol{x}^{\mathsf{T}}\partial \boldsymbol{x}}\left(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x}\right)=\frac{\partial}{\partial \boldsymbol{x}^{\mathsf{T}}}\left[\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{x}}\left(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x}\right)\right]=\frac{\partial}{\partial \boldsymbol{x}^{\mathsf{T}}}\left[\boldsymbol{x}^{\mathsf{T}}\left(\boldsymbol{A}+\boldsymbol{A}^{\mathsf{T}}\right)\right]=\boldsymbol{A}+\boldsymbol{A}^{\mathsf{T}}.$$

b. For the second identity,

$$\frac{\partial}{\partial \boldsymbol{x}} \mathrm{Tr}\left(\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\right) \to \frac{\partial}{\partial x_j} \sum_i \left(x_i^2\right) = 2x_j \to 2\boldsymbol{x}^{\mathsf{T}} .$$

c. For the third identity,

$$\frac{\partial}{\partial \boldsymbol{x}} \left(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{y}\right) \to \frac{\partial}{\partial x_j} \sum_i x_i y_i = y_j \to y^{\mathsf{T}} .$$

Problem A.1.17 Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be $n \times m$ matrices. Show $\partial_{\boldsymbol{A}} \left(\mathrm{Tr}\, \boldsymbol{A}\boldsymbol{B}^{\mathsf{T}}\right) = \partial_{\boldsymbol{A}} \left(\mathrm{Tr}\, \boldsymbol{B}\boldsymbol{A}^{\mathsf{T}}\right) = \boldsymbol{B}^{\mathsf{T}}$. Hint: Make sure your definition of derivative with respect to a matrix is consistent with the previously defined limiting case $m = 1$ for a vector.

Solution.

Let

$$f \equiv \mathrm{Tr}\, \boldsymbol{A}\boldsymbol{B}^{\mathsf{T}} = \sum_{ij} A_{ij} B^{\mathsf{T}}_{ji} = \sum_{ij} A_{ij} B_{ij} .$$

Then

$$\frac{\partial f}{\partial A_{k\ell}} = \frac{\partial}{\partial A_{k\ell}} \sum_{ij} A_{ij} B_{ij} = B_{k\ell} .$$

To know whether the derivative is $\boldsymbol{B}$ or $\boldsymbol{B}^{\mathsf{T}}$ in our convention, let us look at the case $m = 1$. Then $\boldsymbol{A}$ and $\boldsymbol{B}$ are vectors, and we have already established that the derivative of a scalar with respect to a column vector is a row vector. Thus,

$$\frac{\partial}{\partial \boldsymbol{A}} \left(\mathrm{Tr}\, \boldsymbol{A}\boldsymbol{B}^{\mathsf{T}}\right) = \boldsymbol{B}^{\mathsf{T}} .$$

For the other identity,

$$f \equiv \mathrm{Tr}\, \boldsymbol{B}\boldsymbol{A}^{\mathsf{T}} = \sum_{ij} B_{ij} A^{\mathsf{T}}_{ji} = \sum_{ij} B_{ij} A_{ij} .$$

This is just the same as before and gives the same answer.

Problem A.1.18 Let $\boldsymbol{A}$ be an $n \times m$ matrix and let $\boldsymbol{B}$ be an $m \times m$ matrix. Show that $\partial_{\boldsymbol{A}} \left(\mathrm{Tr}\, \boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^{\mathsf{T}}\right) = \left(\boldsymbol{B} + \boldsymbol{B}^{\mathsf{T}}\right) \boldsymbol{A}^{\mathsf{T}}$.

Solution.

Let

$$f \equiv \mathrm{Tr}\, \boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^{\mathsf{T}} = \sum_{ijk} A_{ij} B_{jk} A^{\mathsf{T}}_{ki} = \sum_{ijk} A_{ij} B_{jk} A_{ik} .$$

Then

$$\frac{\partial f}{\partial A_{ab}} = B_{bk} A_{ak} + A_{aj} B_{jb} = B_{bj} A_{aj} + B_{jb} A_{aj} = \left(B_{bj} + B^{\mathsf{T}}_{bj}\right) A_{aj}$$

Let us look at the case $m = 1$. Then $\boldsymbol{A}$ is a vector and $\boldsymbol{B}$ is a scalar. Since the answer must be a row vector, it must involve $\boldsymbol{A}^{\mathsf{T}}$. Thus,

$$\frac{\partial}{\partial \boldsymbol{A}} \left(\mathrm{Tr}\, \boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^{\mathsf{T}} \right) = \left(\boldsymbol{B} + \boldsymbol{B}^{\mathsf{T}} \right) \boldsymbol{A}^{\mathsf{T}} .$$

Problem A.1.19 Show that if $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{A}$, then $\mathrm{e}^{\boldsymbol{A}+\boldsymbol{B}} = \mathrm{e}^{\boldsymbol{A}}\, \mathrm{e}^{\boldsymbol{B}}$. If you are clever, no calculations are required! This identity does *not* hold when $\boldsymbol{A}$ and $\boldsymbol{B}$ do not commute.

Solution.

The straightforward way to prove this would be show that

$$\left(\mathbb{I} + \frac{\boldsymbol{A}+\boldsymbol{B}}{1!} + \cdots + \frac{(\boldsymbol{A}+\boldsymbol{B})^n}{n!} + \cdots \right) = \left(\mathbb{I} + \frac{\boldsymbol{A}}{1!} + \cdots + \frac{\boldsymbol{A}^n}{n!} + \cdots \right) \times \left(\mathbb{I} + \frac{\boldsymbol{B}}{1!} + \cdots + \frac{\boldsymbol{B}^n}{n!} + \cdots \right),$$

by expanding out the terms and matching coefficients. Alternatively, we can argue for the same by noting that if the matrices commute, then the calculation is exactly the same as for real numbers. But we know the identity holds in that case, and hence it does here, too.

For reference, if the matrices do not commute, then the *Baker-Campbell-Hausdorff* formula states that

$$\log \left(\mathrm{e}^{\boldsymbol{A}}\, \mathrm{e}^{\boldsymbol{B}} \right) = \boldsymbol{A} + \boldsymbol{B} + \frac{1}{2}[\boldsymbol{A}, \boldsymbol{B}] + \cdots ,$$

where the higher-order terms involve repeated commutators of $\boldsymbol{A}$ and $\boldsymbol{B}$. There are deep connections to the theory of Lie Algebras.

Problem A.1.20 In analogy with the matrix exponential, we can define a matrix logarithm via the identity $\ln(\mathbb{I} + \boldsymbol{A}) = \boldsymbol{A} - \frac{1}{2}\boldsymbol{A}^2 + \frac{1}{3}\boldsymbol{A}^3 - \cdots$. Use the previous problem to show that if $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{A}$, then $\log \boldsymbol{A}\boldsymbol{B} = \ln \boldsymbol{A} + \log \boldsymbol{B}$.

Solution.

The basic idea is again the same: Since $\boldsymbol{A}$ and $\boldsymbol{B}$ commute, they obey all the ordinary identities of algebra, and thus all the ordinary identities, such as $\mathrm{e}^{a+b} = \mathrm{e}^{a}\, \mathrm{e}^{b}$, for scalars a and b.

Problem A.1.21 Show $\ln \det \boldsymbol{A} = \mathrm{Tr}\, \ln \boldsymbol{A}$, for symmetric, positive-definite matrices $\boldsymbol{A}$. The identity holds for more general $\boldsymbol{A}$ using the complex logarithm.

Solution.

From the identities for trace and det, both are invariant under coordinate transformation. Thus, choose coordinates so that $\boldsymbol{A} = \boldsymbol{D}$ the diagonal matrix of eigenvalues λ_i. This is always possible if $\boldsymbol{A}$ is symmetric and positive definite. In this coordinate system,

$$\ln \det \boldsymbol{A} = \ln \prod_i \lambda_i = \sum_i \ln \lambda_i = \mathrm{Tr}\ \ln \boldsymbol{A} .$$

Problem A.4.1 Let us calculate some simple Fourier series.

a. *Square wave*. Verify the coefficients given in Eq. (A.60) of the Fourier series for the square wave ⊓, defining it to be an even function about $t = 0$.
b. *Square wave with variable duty cycle*. Find the coefficients of an even, asymmetric square wave that equals 1 for a quarter period and 0 for the rest, ⊓.
c. An even function satisfies $f(t) = -f(-t)$. Show that if the function also satisfies $f(t) = -f(t + T/2)$, the even cosine terms will vanish in the Fourier series.
d. *Triangle wave*. Find the coefficients of the Fourier series for the even triangle wave ⋀. Show, in particular, that $a_n \sim O(\frac{1}{n^2})$.

Solution.

a. When the square wave is defined as an even function, the sine terms (b_n in Eq. (A.58)) are zero. The DC term is $\frac{1}{T}\int_{-T/2}^{T/2} \mathrm{d}t\, f(t) = \frac{1}{2}$. The finite-order terms are then, with $\omega = 2\pi/T$,

$$\begin{aligned} a_n &= \frac{2}{T}\int_{-T/2}^{T/2} \mathrm{d}t\, f(t)\cos n\omega t \\ &= \frac{2}{T}\int_{-T/4}^{T/4} \mathrm{d}t\,(1)\,\cos n\omega t \\ &= \frac{4}{T}\int_{0}^{T/4} \mathrm{d}t\cos n\omega t \\ &= \frac{4}{n\omega T}\sin n\omega t\Big|_0^{T/4} \\ &= \left(\tfrac{2}{\pi}\right)\left(\tfrac{1}{n}\right)\sin\tfrac{1}{2}\pi n\,. \end{aligned}$$

The sine term evaluates to +1, 0, −1, 0 +1, …, and thus,

$$\mathrm{sq}(t) = \tfrac{1}{2} + \tfrac{2}{\pi}\left(\cos\omega t - \tfrac{1}{3}\cos 3\omega t + \tfrac{1}{5}\cos 5\omega t - \ldots\right),$$

which is just Eq. (A.60).

b. We can again repeat most of the square-wave derivation in (a). Picking up where things begin to be different,

$$a_n = \frac{4}{T}\int_0^{T/8} \mathrm{d}t\cos n\omega t = \frac{4}{n\omega T}\sin n\omega t\Big|_0^{T/8} = \left(\tfrac{2}{\pi}\right)\left(\tfrac{1}{n}\right)\sin\tfrac{1}{4}\pi n\,.$$

The pattern of the sine term is now $\frac{1}{\sqrt{2}}, 1, \frac{1}{\sqrt{2}}, 0, -\frac{1}{\sqrt{2}}, -1, -\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}, 0\ldots$, and the DC term is $\frac{1}{4}$, which gives

$$\mathrm{sq}_{1/4}(t) = \tfrac{1}{4} + \tfrac{2}{\pi}\left(\tfrac{1}{\sqrt{2}}\cos\omega t + \tfrac{1}{2}\cos 2\omega t + \tfrac{1}{3\sqrt{2}}\cos 3\omega t - \tfrac{1}{5\sqrt{2}}\cos 5\omega t - \cdots\right).$$

Notice the "missing orders" $\cos 4\omega t$, $\cos 8\omega t$, etc. These are analogous to the case in optics where missing orders in a diffraction grating occur when there is a rational ratio between wavelength and slit spacing.

c. As we have seen, an even function has only cosine terms. Now let us impose the additional symmetry $f(t) = -f(t + T/2)$. In words, the function is invariant if you slide it half a period and invert. Notice that the standard square wave satisfies this symmetry, but a square wave with asymmetric duty cycle does not. Let us compute the Fourier series of the shifted function, noting again that there are only cosine terms. With the substitution $s = t + T/2$ and using the fact that we can integrate over one period starting and ending anywhere we like, we have

$$a_n = \frac{2}{T}\int_{-T/2}^{T/2} \mathrm{d}t\, f(t + T/2)\cos n\omega t = \frac{2}{T}\int_{-T/2}^{T/2} \mathrm{d}s\, f(s)\cos n\omega(s - T/2)\,.$$

Let us use the trig identity

$$\cos n\omega(s - T/2) = \cos n\omega s\, \cos n\pi - \sin \omega s\, \sin n\pi\,.$$

The second term is zero ($\sin n\pi = 0$ for integer n). If we look at even orders, $n = 2m$, then $\cos 2m\pi = 1$ for integer m. Thus, for our shifted function,

$$a_{2m} = \frac{2}{T}\int_{-T/2}^{T/2} \mathrm{d}s\, f(s)\cos 2m\omega s\,.$$

But this is also what we would find for the same coefficient of the expansion of $f(t)$. Thus, the only way for a function to satisfy the symmetry $f(t) = -f(t + T/2)$ is to have the even orders $a_{2m} = 0$.

d. The sine terms are again 0, and $a_0 = \frac{1}{2}$. The Fourier cosine terms have amplitude

$$\begin{aligned} a_n &= \frac{4}{T}\int_0^{T/2} \mathrm{d}t\left(\frac{2t}{T}\right)\cos n\omega t \\ &= \frac{8t}{T^2}\,\frac{\sin n\omega t}{n\omega}\bigg|_0^{T/2} - \frac{8}{n\omega T^2}\int_0^{T/2} \mathrm{d}t\,\sin n\omega t \\ &= -\left(\frac{8}{n\omega T^2}\right)\left(\frac{2}{n\omega}\right)\frac{1}{2}\,(1 - \cos \pi n) \\ &= -\left(\frac{4}{\pi^2 n^2}\right)\frac{1}{2}\,(1 - \cos \pi n)\ . \end{aligned}$$

where we integrate by parts and note that the boundary term is zero. The cosine terms give a pattern $1, 0, 1, 0\ldots$ for $n = 1, 2, 3, 4\ldots$ and, thus, for general T,

$$\mathrm{tr}(t) = \tfrac{1}{2} - \left(\tfrac{2}{\pi}\right)^2\left(\cos \omega t + \tfrac{1}{9}\cos 3\omega t + \tfrac{1}{25}\cos 5\omega t + \cdots\right)\,.$$

Problem A.4.2 **Poisson summation formula**. Prove the following version of the Poisson summation formula, which relates Fourier coefficients to Fourier transforms for a periodic function $f(t) = f(t + T)$ built out of non-periodic "basis" functions $g(t)$. Show that $f(t) = \sum_{k=-\infty}^{\infty} g(t + kT) = \frac{1}{T}\sum_{n=-\infty}^{\infty} G(n\omega)\,\mathrm{e}^{\mathrm{i}n\omega t}$, where $\omega = \frac{2\pi}{T}$ and the Fourier transform $G(\omega) = \int_{-\infty}^{\infty} \mathrm{d}t\, g(t)\ \mathrm{e}^{-\mathrm{i}\omega t}$.

Solution. Since $f(t)$ is periodic with period T, it has a Fourier series representation

$$f(t) = \sum_{n=-\infty}^{\infty} c_n \, \mathrm{e}^{\mathrm{i}n\omega t}$$

with coefficients given by Eq. (A.63):

$$\begin{aligned}
c_n &= \frac{1}{T}\int_0^T \mathrm{d}t\, f(t)\ \mathrm{e}^{-\mathrm{i}n\omega t} \\
&= \frac{1}{T}\int_0^T \mathrm{d}t \sum_{k=-\infty}^{\infty} g(t+kT)\ \mathrm{e}^{-\mathrm{i}n\omega t} \\
&= \frac{1}{T}\sum_{k=-\infty}^{\infty}\int_0^T \mathrm{d}t\, g(t+kT)\ \mathrm{e}^{-\mathrm{i}n\omega t} \\
&= \frac{1}{T}\sum_{k=-\infty}^{\infty}\int_{kT}^{(k+1)T} \mathrm{d}t'\, g(t')\ \mathrm{e}^{-\mathrm{i}n\omega(t'-kT)} \\
&= \frac{1}{T}\sum_{k=-\infty}^{\infty}\int_{kT}^{(k+1)T} \mathrm{d}t'\, g(t')\ \mathrm{e}^{-\mathrm{i}n\omega t'}\ \underbrace{\mathrm{e}^{\mathrm{i}nk\omega T}}_{=1} \\
&= \frac{1}{T}\int_{-\infty}^{\infty} \mathrm{d}t'\, g(t')\ \mathrm{e}^{-\mathrm{i}n\omega t'} \\
&= \frac{1}{T} G(n\omega)\,.
\end{aligned}$$

Problem A.4.3 **Parseval's theorem**. Show that $\int_{-\infty}^{\infty} \mathrm{d}t\,[f(t)]^2 = \int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi}\left|F_f(\omega)\right|^2$. Check the relation explicitly for $f(t) = \theta(t)\ \mathrm{e}^{-t}$, with $\theta(t)$ the Heaviside step function.

Solution.

$$\begin{aligned}
\int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} F_f(\omega) F_f(\omega)^* &= \int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}t\, f(t)\ \mathrm{e}^{-\mathrm{i}\omega t} \int_{-\infty}^{\infty} \mathrm{d}t'\, f(t')\ \mathrm{e}^{+\mathrm{i}\omega t'} \\
&= \int_{-\infty}^{\infty} \mathrm{d}t\, \mathrm{d}t' f(t)\, f(t') \int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi}\ \mathrm{e}^{\mathrm{i}\omega(t'-t)} \\
&= \int_{-\infty}^{\infty} \mathrm{d}t\, \mathrm{d}t' f(t)\, f(t')\, \delta(t'-t) \\
&= \int_{-\infty}^{\infty} \mathrm{d}t\,[f(t)]^2\ .
\end{aligned}$$

using Eq. (A.77) for the delta function.

For $f(t) = \theta(t)\ \mathrm{e}^{-t}$,

$$\int_{-\infty}^{\infty} \mathrm{d}t\,[f(t)]^2 = \int_0^{\infty} \mathrm{d}t\, \mathrm{e}^{-2t} = \frac{1}{2}\,.$$

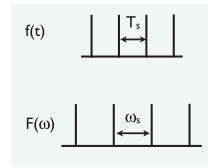
On the other hand,

$$F_f(\omega) = \int_{-\infty}^{\infty} \mathrm{d}t\, f(t)\ \mathrm{e}^{-\mathrm{i}\omega t} = \int_0^{\infty} \mathrm{d}t\, \mathrm{e}^{-(1+\mathrm{i}\omega)t} = \frac{1}{1+\mathrm{i}\omega}\,,$$

so that

$$\int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} \left|F_f(\omega)\right|^2 = \int_{-\infty}^{\infty} \frac{\mathrm{d}\omega}{2\pi} \frac{1}{1+\omega^2} = \frac{1}{2\pi}\,(2\pi\,\mathrm{i})\,\frac{1}{2\,\mathrm{i}} = \frac{1}{2}\,,$$

where we use the Residue Theorem to evaluate the contour integral.



Problem A.4.4 **Fourier transform of a comb function**. By applying the Poisson summation formula to the delta function, $g(t) = \delta\,(t)$, show that the Fourier transform of the time-domain comb function, $f(t) = \sum_k \delta\,(t - kT_\mathrm{s})$, is the frequency-domain comb function $F(\omega) = \frac{2\pi}{T_\mathrm{s}} \sum_n \delta\,(\omega - n\omega_\mathrm{s})$.

Solution. In the Poisson summation formula from Problem A.4.2, we set $g(t) = \delta\,(t)$, which implies

$$G(\omega) = 1\,.$$

Thus, the Fourier series representation is

$$f(t) = \sum_{k=-\infty}^{\infty} \delta\,(t - kT_\mathrm{s}) = \frac{1}{T_\mathrm{s}} \sum_{n=-\infty}^{\infty} \mathrm{e}^{\mathrm{i}n\omega_\mathrm{s}t}\,.$$

Next, we take the Fourier transform:

$$\begin{aligned} F(\omega) &= \int_{-\infty}^{\infty} \mathrm{d}t\, f(t)\, \mathrm{e}^{-\mathrm{i}\omega t} \\ &= \sum_{n=-\infty}^{\infty} \frac{1}{T_\mathrm{s}} \int_{-\infty}^{\infty} \mathrm{d}t\,(1)\, \mathrm{e}^{\mathrm{i}n\omega_\mathrm{s}t}\, \mathrm{e}^{-\mathrm{i}\omega t} \\ &= \frac{1}{T_\mathrm{s}} \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} \mathrm{d}t\, \mathrm{e}^{-\mathrm{i}(\omega - n\omega_\mathrm{s})t} \\ &= \frac{2\pi}{T_\mathrm{s}} \sum_{n=-\infty}^{\infty} \delta\,(\omega - n\omega_\mathrm{s})\,. \end{aligned}$$

Here, we use Eq. (A.76) in the form

$$2\pi\,\delta\,(\omega) = \int_{-\infty}^{\infty} \mathrm{d}t\, \mathrm{e}^{-\mathrm{i}\omega t}\,.$$

Problem A.4.5 **Laplace transform of integral**. Show that $\mathcal{L}[\int_0^t \mathrm{d}t'\, f(t')] = \frac{1}{s} F(s)$.

Solution. The Laplace transform of $\int_0^t \mathrm{d}t'\, f(t')]$ is

$$\begin{aligned} \mathcal{L}\left[\int_0^t \mathrm{d}t'\, f(t')\right] &= \int_0^\infty \mathrm{d}t \underbrace{\left[\int_0^t \mathrm{d}t'\, f(t')\right]}_{u} \underbrace{\mathrm{e}^{-st}}_{dv} \\ &= 0 - \int_0^\infty \mathrm{d}t\, f(t) \left(-\frac{\mathrm{e}^{-st}}{s}\right) \end{aligned}$$

$$= +\frac{1}{s}\int_0^\infty \mathrm{d}t\, f(t)\ \mathrm{e}^{-st}$$
$$= \frac{1}{s}F(s)\,.$$

The boundary term $uv|_0^\infty = 0$ because of the integral term at $t = 0$ and the exponential at $t = \infty$.

Problem A.7.1 Show that $\int_{-\infty}^{\infty} \mathrm{d}x \exp\left[-\frac{1}{2}(ax^2 + bx)\right] = \sqrt{\frac{2\pi}{a}}\ \mathrm{e}^{b^2/8a}$. (Complete the square.)

Solution.

$$\begin{aligned}
\int_{-\infty}^{\infty} \mathrm{d}x \exp\left[-\tfrac{1}{2}(ax^2 + bx)\right] &= \int_{-\infty}^{\infty} \mathrm{d}x \exp\left[-\tfrac{1}{2}a\left(x^2 + \frac{b}{a}x\right)\right] \\
&= \int_{-\infty}^{\infty} \mathrm{d}x \exp\left[-\tfrac{1}{2}a\left(x^2 + \frac{b}{a}x + \frac{b^2}{4a^2} - \frac{b^2}{4a^2}\right)\right] \\
&= \int_{-\infty}^{\infty} \mathrm{d}x \exp\left[-\tfrac{1}{2}a\left(x + \frac{b}{2a}\right)^2\right] \mathrm{e}^{b^2/8a} \\
&= \int_{-\infty}^{\infty} \mathrm{d}y \exp\left[-\tfrac{1}{2}ay^2\right] \mathrm{e}^{b^2/8a} \\
&= \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} \mathrm{d}z \exp\left[-\tfrac{z^2}{2}\right] \mathrm{e}^{b^2/8a} \\
&= \sqrt{\tfrac{2\pi}{a}}\ \mathrm{e}^{b^2/8a}\ .
\end{aligned}$$

Problem A.7.2 **Characteristic function of a Gaussian.** Derive the characteristic function of a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. The result is quoted in Eq. (A.183).

Solution.

The characteristic function of a Gaussian-distributed variable $x \sim \mathcal{N}(\mu, \sigma^2)$ is

$$\varphi_x(k) = \langle \mathrm{e}^{\mathrm{i}kx} \rangle = \int_{-\infty}^{\infty} \mathrm{d}x\, \frac{1}{\sqrt{2\pi}\sigma}\ \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\ \mathrm{e}^{\mathrm{i}kx}\ .$$

We change variables to $z = \frac{x-\mu}{\sigma}$, or $x = \mu + \sigma z$, with $\mathrm{d}x = \sigma\, \mathrm{d}z$. Then,

$$\begin{aligned}
\varphi_x(k) &= \frac{\sigma}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \mathrm{d}z\, \mathrm{e}^{-z^2/2}\ \mathrm{e}^{\mathrm{i}k(\mu+\sigma z)} \\
&= \frac{\mathrm{e}^{\mathrm{i}k\mu}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}z\, \mathrm{e}^{-\frac{1}{2}(z^2 - 2\,\mathrm{i}k\sigma z)} \\
&= \frac{\mathrm{e}^{\mathrm{i}k\mu}}{\sqrt{2\pi}} \sqrt{2\pi}\, \mathrm{e}^{-\frac{4k^2\sigma^2}{8}} \\
&= \mathrm{e}^{\mathrm{i}k\mu}\ \mathrm{e}^{-k^2\sigma^2/2}\,,
\end{aligned}$$

which is the result in Eq. (A.183).

Problem A.7.3 Higher moments of a Gaussian. For a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, use the characteristic function to calculate the first four moments, $\langle x^n \rangle$ ($n = 1, 2, 3, 4$). Verify that the skewness γ_1 and kurtosis γ_2 both vanish.

Solution.

We recall that the characteristic function $\varphi_k(x)$ can generate moments via

$$\langle x^m \rangle = (-\mathrm{i})^m \left. \frac{\mathrm{d}^m}{\mathrm{d}k^m} \varphi_x(k) \right|_{k=0} .$$

For a Gaussian random variable $x \sim \mathcal{N}(\mu, \sigma^2)$,

$$\varphi_x(k) = \mathrm{e}^{\mathrm{i}k\mu} \, \mathrm{e}^{-k^2\sigma^2/2} .$$

The first four derivatives $\frac{\mathrm{d}^m}{\mathrm{d}k^m}\varphi_x(k)$ with $m = 1, 2, 3, 4$ are, with $\alpha \equiv (\mathrm{i}\mu - k\sigma^2)$,

$$\begin{aligned}
\varphi_x'(k) &= \varphi_x(k)\alpha \\
\varphi_x''(k) &= \varphi_x(k)\left[-\sigma^2 + \alpha^2\right] \\
\varphi_x'''(k) &= \varphi_x(k)\left[-3\sigma^2\alpha + \alpha^3\right] \\
\varphi_x''''(k) &= \varphi_x(k)\left[3\sigma^4 - 6\sigma^2\alpha^2 + \alpha^4\right] .
\end{aligned}$$

To find the moments, we note that $\varphi_x'(0) = 1$ and $\alpha(0) = \mathrm{i}\mu$. Then,

$$\begin{aligned}
\langle x \rangle &= (-\mathrm{i})(\mathrm{i}\mu) = \mu \\
\left\langle x^2 \right\rangle &= (-\mathrm{i})^2\left(-\sigma^2 - \mu^2\right) = \sigma^2 + \mu^2 \\
\left\langle x^3 \right\rangle &= (-\mathrm{i})^3\left(-3\,\mathrm{i}\sigma^2\mu - \mathrm{i}\mu^3\right) = 3\sigma^2\mu + \mu^3 \\
\left\langle x^4 \right\rangle &= (-\mathrm{i})^4\left(3\sigma^4 + 6\sigma^2\mu^2 + \mu^4\right) = 3\sigma^4 + 6\sigma^2\mu^2 + \mu^4 .
\end{aligned}$$

The skewness and kurtosis are best found by changing coordinates to $z = \frac{x-\mu}{\sigma}$, but we will calculate them the "hard way" by directly evaluating the moments. For skewness,

$$\begin{aligned}
\gamma_1 &= \left\langle \left(\frac{x-\mu}{\sigma}\right)^3 \right\rangle \\
&= \frac{1}{\sigma^3}\left(\langle x^3 \rangle - 3\langle x^2 \rangle \mu + 3\langle x \rangle \mu^2 - \mu^3\right) \\
&= \frac{1}{\sigma^3}\left((3\sigma^2\mu + \mu^3) - 3(\sigma^2 + \mu^2)\mu + 3\mu^3 - \mu^3\right) \\
&= 0 .
\end{aligned}$$

For the excess kurtosis,

$$\begin{aligned}
\gamma_2 + 3 &= \left\langle \left(\frac{x-\mu}{\sigma}\right)^4 \right\rangle \\
&= \frac{1}{\sigma^4}\left(\langle x^4 \rangle - 4\langle x^3 \rangle \mu + 6\langle x^2 \rangle \mu^2 - 3\langle x \rangle \mu^3 + \mu^4\right)
\end{aligned}$$

$$= \frac{1}{\sigma^4}\left((3\sigma^4 + 6\sigma^2\mu^2 + \mu^4) - 4(3\sigma^2\mu + \mu^3)\mu + 6(\sigma^2 + \mu^2)\mu^2 - 4\mu^4 + \mu^4\right)$$
$$= 3\,,$$

so that $\gamma_2 = 0$, as well.

Problem A.7.4 Central-limit theorem (CLT), via cumulants. Consider N independent, identically distributed (i.i.d.) variables, each with mean zero and variance σ^2.

a. *Homogeneity*: Show that the cumulant $\kappa_m(\lambda x) = \lambda^m \kappa_m(x)$, where $\lambda > 0$.
b. Using additivity and homogeneity, find $\kappa_m(z_N)$ for $z_N \equiv \sum_{i=1}^{N}(x_i/\sqrt{N\sigma^2})$.
c. Argue that, for $N \to \infty$, the only non-zero cumulant is $m = 2$.
d. Conclude that $\lim_{N\to\infty} p(z_N) \sim \mathcal{N}(0, 1)$.
e. Define $\bar{x}_N = \frac{1}{N}\sum_i x_i$. Find $\lim_{N\to\infty} p(\bar{x}_N)$.

This is the essence of the CLT proof and can be generalized to the case where the x_i all have different distributions, each with its own mean μ_i and variance σ_i^2.

Solution.

a. Homogeneity: From the definition, $h_x(k) = \ln\left\langle e^{ikx}\right\rangle$, we can write

$$h_{\lambda x}(k) = \ln\left\langle e^{ik(\lambda x)}\right\rangle = \ln\left\langle e^{i(\lambda k)x}\right\rangle = \sum_{m=0}^{\infty}\frac{(i\lambda k)^m \kappa_m(x)}{m!} = \sum_{m=0}^{\infty}\frac{(ik)^m[\lambda^m\kappa_m(x)]}{m!}\,.$$

Alternatively,

$$h_{\lambda x}(k) = \sum_{m=0}^{\infty}\frac{(ik)^m \kappa_m(\lambda x)}{m!}\,,$$

which implies $\kappa_m(\lambda x) = \lambda^m \kappa_m(x)$.

b. With $z_N \equiv \sum_{i=1}^{N}(x_i/\sqrt{N\sigma^2})$, we have

$$\kappa_m(z_N) = \underbrace{N\kappa(x_i/\sqrt{N\sigma^2})}_{\text{additivity}} = N\,\underbrace{(N\sigma^2)^{-m/2}\kappa_m(x_i)}_{\text{homogeneity}} = N^{1-m/2}\sigma^{-m}\kappa_m(x_i)\,.$$

c. When $N \to \infty$, the factor $N^{1-m/2}$ diverges for $m = 1$, is finite for $m = 2$, and goes to zero for $m \geq 3$. The $m = 1$ divergence is tamed because $\kappa_1(x_i) = \mu = 0$. Thus, only the $m = 2$ cumulant is non-zero.

d. Since $\kappa_2(x_i) = \sigma^2$,

$$\kappa_2(z_N) = \underbrace{N^{1-2/2}}_{N^0=1}\left(\sigma^{-2}\right)\left(\sigma^2\right) = 1\,.$$

Thus, as $N \to \infty$, the cumulants are $\kappa_m(z_N) = \delta_{m,2}$. The cumulant generating function is then $h_{z_N}(x) = -\frac{1}{2}k^2$, which implies a characteristic function $\varphi_{z_N}(k) = \exp\left[-\frac{1}{2}k^2\right]$, and corresponds to a normal distribution: $\lim_{N\to\infty} p(z_N) \sim \mathcal{N}(0, 1)$.

e. We can rephrase the conclusion for the N^{th} approximation to the average. If

$$\bar{x}_N \equiv \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}\left(\sqrt{N\sigma^2}z_N\right) = \sqrt{\frac{\sigma^2}{N}}\,z_N\,,$$

then, for $N \to \infty$,

$$p(z_N) \sim \mathcal{N}(0,1) \quad \Longrightarrow \quad p(\bar{x}_N) \sim p\left(\sqrt{\frac{\sigma^2}{N}}\, z_N\right) \sim \mathcal{N}(0, \sigma^2/N)\,.$$

Thus, the average of N random variables tends to a Gaussian whose width decreases as $N^{-1/2}$.

I learned about this way to prove the CLT from Prof. Haye Hinrichsen (Univ. Würzburg, Germany).

Problem A.7.5 **Multiple measurements lead to Gaussian states-of-knowledge**. The Central Limit Theorem also explains why the state-of-knowledge for a quantity x tends to be Gaussian after many independent measurements are made.

a. Use Bayes' theorem, a uniform prior, the relations between characteristic functions and repeated convolution and the CLT, and that a Fourier transform of a Gaussian is also Gaussian to argue this point. (See Jacobs, 2014, Section 1.2.2.)
b. Explain why this claim is true but rather trivial when the individual measurements have Gaussian errors.
c. Explain why this claim is *not* true when the individual measurements have a uniform error distribution in the interval $[x - \frac{1}{2},\, x + \frac{1}{2}]$.

Solution.

a. Define $y^N = \{y_1, y_2, \ldots, y_N\}$ to be a set of N independent measurements of the same quantity x. Then, using Bayes' theorem and assuming a uniform prior for x,

$$p(x|y^N) \propto p(y^N|x)\, p(x) \propto p(y|x) = \prod_{i=1}^{N} p(y_i|x)\,.$$

Next, consider the CLT from the point of view of characteristic functions. The CLT claims that if we make N measurements y^N of the same quantity x that the average,

$$\bar{y} \equiv \frac{1}{N} \sum y_i \to \mathcal{N}(x, \sigma^2)\,,$$

as $N \to \infty$. The variance of the Gaussian, σ^2, depends on the individual variances of measurements (e.g., σ_0^2/N for measurements that all have individual variances σ_0^2).

The next step is to show that the distribution of the sum of N random variables is the repeated convolution of the individual distributions. The argument generalizes Example A.18. Let $z_N \equiv \sum_{i=1}^{N} y_i$. Then

$$\begin{aligned} p(z_2) &= \int \mathrm{d}y_1\, \mathrm{d}y_2\; \delta\,(z - y_1 - y_2)\, p(y_1)\, p(y_2) \\ &= \int \mathrm{d}y_1\, p(y_1) p(z - y_1) \equiv p(y_1) * p(y_2)\,. \end{aligned}$$

Similarly, with $z_3 = z_2 + y_3$, we have

$$p(z_3) = p(z_2) * p(y_3) = [p(y_1) * p(y_2)] * p(y_3) \equiv p(y_1) * p(y_2) * p(y_3)\,.$$

Repeating, we conclude that $p(z_N) = p(y_1) * \ldots * p(y_N)$. We can extend this result to $p(\overline{y}) = p(z_N/N) = N\,p(z_N)$. Thus the probability distribution of the average is proportional to the repeated convolution of the individual distributions.

Now, we combine these two results: the CLT says that $p(y^N)$ tends to a Gaussian. But $p(y^N)$ is also a repeated convolution. Thus, the repeated convolution of N variables must tend to a Gaussian probability distribution.

Finally, in our original problem, the state-of-belief $p(x|y^N)$ is the product of N probability distributions. The characteristic function of such a product is the repeated convolution of the N characteristic functions. By the previous argument, this convolution must tend to a Gaussian (in the variable k). Since the Fourier transform of a Gaussian in k space is another Gaussian (in x), we have our result.

b. If the likelihood of each measurement, $p(y_i|x)$ is Gaussian, then the claim is true for all N because the product of two Gaussian functions is another Gaussian.

c. The claim is <u>not</u> true for uniform distributions. We can see this two ways. First, by direct calculation. The likelihood function $p(y_i|x)$ is the interval $[x - \frac{1}{2}, x + \frac{1}{2}]$. This means that, with a uniform prior,

$$p(x|y^N) \propto \prod_{i=1}^{N} 1\{y_i \in [x - \tfrac{1}{2}, x + \tfrac{1}{2}]\}\,.$$

This is distribution is also uniform (i.e., not Gaussian) and ranges over the common intersection of the N intervals about each y_i. For example, given measurements y_1 and y_2, we would form the intersection of $y_1 \pm \frac{1}{2}$ and $y_2 \pm \frac{1}{2}$, and so on. Thus, the product never converges to a Gaussian.

To understand more deeply what is going on, we recall that the characteristic function of a uniform distribution is the sinc function, $\sin x/x$. This goes to zero asympotitically as $1/x$, meaning that its variance diverges and that the CLT cannot be applied. Loosely, we can view this as a case where the high frequencies always are important (the distribution always cuts off very quickly), rather than having ever-reduced importance when the CLT applies. The requirement for this claim to be valid is then that the tail of each measurement decay *slowly* enough. A local parabola would suffice.

Problem A.7.6 Marginal and conditional distributions.

a. Show that if $p(x, y)$ is a bivariate Gaussian, then the marginal distributions are $p(x) = \mathcal{N}(\mu_x, \sigma_x^2)$ and $p(y) = \mathcal{N}(\mu_y, \sigma_y^2)$.

b. Show $p(x|y) = \mathcal{N}(\mu_{x|y}, \sigma^2_{x|y})$, with $\mu_{x|y} = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(y - \mu_y)$ and $\sigma^2_{x|y} = \sigma^2_x(1 - \rho^2)$. Hint: Simplify by first defining $x' = (x - \mu_x)/\sigma_x$ and $y' = (y - \mu_y)/\sigma_y$.

Solution.

We change variables as suggested. Dropping the primes, the joint distribution is

$$p(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right].$$

a. For the marginal distribution $p(x)$, we have

$$\begin{aligned} p(x) &= \int_{-\infty}^{\infty} \mathrm{d}y\, p(x, y) \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{x^2}{2(1-\rho^2)}\right] \int_{-\infty}^{\infty} \mathrm{d}y \exp\left[-\frac{y^2 - 2\rho xy}{2(1-\rho^2)}\right] \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{x^2}{2(1-\rho^2)}\right] \exp\left[\frac{\rho^2 x^2}{2(1-\rho^2)}\right] \int_{-\infty}^{\infty} \mathrm{d}y \exp\left[-\frac{(y-\rho x)^2}{2(1-\rho^2)}\right] \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{x^2}{2(1-\rho^2)}\right] \exp\left[\frac{\rho^2 x^2}{2(1-\rho^2)}\right] \sqrt{2\pi(1-\rho^2)} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right]. \end{aligned}$$

Thus, $p(x) = \mathcal{N}(0, 1)$, which, transforming back to the original variables, gives $p(x) = \mathcal{N}(\mu_x, \sigma^2_x)$. The calculation for $p(y)$ is identical, interchanging $x \leftrightarrow y$.

b. For the conditional distribution $p(x|y)$, we have

$$\begin{aligned} p(x|y) = \frac{p(x, y)}{p(y)} &\propto \exp\left[-\frac{1}{2}\left(\frac{x^2 - 2\rho xy + y^2}{1-\rho^2} - y^2\right)\right] \\ &= \exp\left[-\frac{1}{2}\left(\frac{x^2 - 2\rho xy + y^2}{1-\rho^2} - \frac{(1-\rho^2)y^2}{1-\rho^2}\right)\right] \\ &= \exp\left[-\frac{1}{2}\left(\frac{x^2 - 2\rho xy + \rho^2 y^2}{1-\rho^2}\right)\right] \\ &= \exp\left[-\frac{1}{2}\left(\frac{(x-\rho y)^2}{1-\rho^2}\right)\right]. \end{aligned}$$

When normalized properly, the above expression describes a Gaussian, with

$$\mu_{x|y} = \rho y\,, \qquad \sigma^2_{x|y} = 1 - \rho^2\,.$$

Going back to the original variables proves the relations in Eq. (A.192).

Problem A.7.7 Check that the expressions for mean and variance for the bivariate distribution (Eq. A.192) are compatible with their n-dimensional generalizations (Eq. A.197).

Solution.

We first compute mean and covariance for the scaled distribution

$$p(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right],$$

where we have substituted

$$x = \frac{x' - \mu_x}{\sigma_x}, \qquad y = \frac{y' - \mu_y}{\sigma_y}.$$

Then it is straightforward to see (perhaps using Mathematica) that

$$\langle x \rangle = \langle y \rangle = 0, \qquad \left\langle x^2 \right\rangle = \left\langle y^2 \right\rangle = 1, \qquad \langle xy \rangle = \rho.$$

Transforming back to the original coordinates then gives, for the means

$$\langle x \rangle = \mu_x, \qquad \langle y \rangle = \mu_y,$$

and, for the covariance elements

$$\Sigma_{xx} = \left\langle (x-\mu_x)^2 \right\rangle = \sigma_x^2, \qquad \Sigma_{yy} = \left\langle (y-\mu_y)^2 \right\rangle = \sigma_y^2,$$
$$\Sigma_{xy} = \langle (x-\mu_x)(y-\mu_y) \rangle = \rho\,\sigma_x\sigma_y.$$

In Eq. (A.192), we saw that $p(x|y)$ is a Gaussian whose mean is

$$\begin{aligned}\mu_{x|y} &= \mu_x + \rho\frac{\sigma_x}{\sigma_y}(y-\mu_y) \\ &= \mu_x + \frac{\rho\sigma_x\sigma_y}{\sigma_y^2}(y-\mu_y) \\ &= \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y).\end{aligned}$$

and variance

$$\begin{aligned}\sigma_{x|y}^2 &= \sigma_x^2(1-\rho^2) \\ &= \sigma_x^2 - \frac{\rho^2\sigma_x^2\sigma_y^2}{\sigma_y^2} \\ &= \Sigma_{xx} - \Sigma_{xy}^2/\Sigma_{yy} \\ &= \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}.\end{aligned}$$

Thus, marginalizing the bivariate Gaussian distribution gives a scalar version of the more general formula from Eq. (A.197), which holds for arbitrary dimensions d_1 and d_2 for $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

Problem A.7.8 Propagating means and covariances. Prove Property A.7.5 for multivariate Gaussians: if $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the linear coordinate transformation $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x}$ is Gaussian with mean $\boldsymbol{A}\boldsymbol{\mu}$ and covariance $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\mathsf{T}$. More succinctly, $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\mathsf{T})$. Hint: You can use Eq. (A.179), but characteristic functions are simpler.

Solution.

The characteristic function is

$$\left\langle e^{i\boldsymbol{k}\cdot\boldsymbol{z}}\right\rangle = \left\langle e^{i\boldsymbol{k}\cdot\boldsymbol{A}\boldsymbol{x}}\right\rangle = \left\langle e^{i(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{k})\cdot\boldsymbol{x}}\right\rangle = e^{i(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{k})\cdot\boldsymbol{\mu}-(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{k})\cdot\boldsymbol{\Sigma}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{k}/2}$$
$$= e^{i\boldsymbol{k}\cdot\boldsymbol{A}\boldsymbol{\mu}-\boldsymbol{k}\cdot(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathsf{T}})\boldsymbol{k}/2}\,.$$

This last expression is the characteristic of a Gaussian with mean $\boldsymbol{A}\boldsymbol{\mu}$ and covariance $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathsf{T}}$.

Problem A.8.1 Two Gaussian measurements. Often, we need to combine information from independent measurements with different precision. Assume that there are two independent measurements x_1 and x_2, distributed as $x_1 \sim \mathcal{N}(\mu, \sigma_1^2)$ and $x_2 \sim \mathcal{N}(\mu, \sigma_2^2)$. The variances σ_1 and σ_2 are known, and we wish to infer the mean, μ. Assume a uniform prior for μ.

a. Using Bayes' theorem, show that you can estimate μ as $\hat{\mu} \pm \sigma_\mu$, with

$$\hat{\mu} = \sigma_\mu^2\left(\frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2}\right), \qquad \frac{1}{\sigma_\mu^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\,.$$

b. Show that $p(\mu|x_1, x_2, \sigma_1^2, \sigma_2^2)$ is in fact a Gaussian, with the mean and variance given above. If you have a computer algebra program, do this in general. If you are doing the problem by hand, show the claim assuming that $\sigma_1^2 = \sigma_2^2 = 1$.

Solution.

a. With a uniform prior for μ, we can write

$$p(\mu|x_1, x_2, \sigma_1^2, \sigma_2^2) \propto p(x_1, x_2|\mu, \sigma_1^2, \sigma_2^2) \propto p(x_1|\mu, \sigma_1^2)\, p(x_2|\mu, \sigma_2^2) \equiv L$$

Since $p(x_1|\mu, \sigma_1^2) = \mathcal{N}(\mu, \sigma_1^2)$ and $p(x_2|\mu, \sigma_2^2) = \mathcal{N}(\mu, \sigma_1^2)$, we can write

$$-\ln L = \frac{(x_1-\mu)^2}{2\sigma_1^2} + \frac{(x_2-\mu)^2}{2\sigma_2^2}\,,$$

We find $\hat{\mu}$ by maximizing L and hence minimizing $-\ln L$. Taking $\partial_\mu(-\ln L) = 0$ gives

$$\partial_\mu(-\ln L) = -\frac{x_1-\mu}{\sigma_1^2} - \frac{x_2-\mu}{\sigma_2^2} = 0\,.$$

Solving for μ then gives

$$\hat{\mu} = \sigma_\mu^2\left(\frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2}\right), \qquad \text{where} \qquad \frac{1}{\sigma_\mu^2} \equiv \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}$$

Taking the second derivative $\partial_{\mu\mu}(-\ln L)$ gives

$$\partial_{\mu\mu}(-\ln L) = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\,.$$

Then, since $\sigma_\mu^2 \equiv \partial_{\mu\mu}(-\ln L)^{-1}$, we have

$$\frac{1}{\sigma_\mu^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\,,$$

which confirms the identification made in the estimate for $\hat{\mu}$.

b. i. Using *Mathematica*, I find that

$$\frac{(x_1-\mu)^2}{2\sigma_1^2} + \frac{(x_2-\mu)^2}{2\sigma_2^2} - \frac{(\hat{\mu}-\mu)^2}{2\sigma_\mu^2} = \frac{(x_1-x_2)^2}{2(\sigma_1^2+\sigma_2^2)}\,.$$

Since the right-hand side is independent of μ, we conclude that since the exponential of $-\frac{(\hat{\mu}-\mu)^2}{2\sigma_\mu^2}$ describes a Gaussian, so does the exponential of $-\frac{(x_1-\mu)^2}{2\sigma_1^2} - \frac{(x_2-\mu)^2}{2\sigma_2^2}$. That is,

$$\exp\left[-\frac{(x_1-\mu)^2}{2\sigma_1^2}\right]\exp\left[-\frac{(x_2-\mu)^2}{2\sigma_2^2}\right] \propto \exp\left[-\frac{(\hat{\mu}-\mu)^2}{2\sigma_\mu^2}\right],$$

where we omit constants that are independent of μ.

ii. For hand calculations, simplify by setting $\sigma_1 = \sigma_2 = 1$. We then have $\hat{\mu} = \frac{1}{2}(x_1+x_2)$ and $\sigma_\mu^2 = \frac{1}{2}$. We can then easily verify that

$$\frac{(x_1-\mu)^2}{2} + \frac{(x_2-\mu)^2}{2} - \frac{\left[\left(\frac{1}{2}(x_1+x_2)-\mu\right)\right]^2}{2(\frac{1}{2})} = \frac{(x_1-x_2)^2}{4}\,.$$

Problem A.8.2 Show that if $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbb{I})$, then the best estimate $\hat{\boldsymbol{\theta}}$ of a linear fit is Gaussian distributed about the true values $\hat{\boldsymbol{\theta}}^*$, with variance $\sigma^2(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi})^{-1}$. Notice how the variance of the parameter estimates is proportional to the variance of the original data. Qualitatively, what happens if the measurement noise is colored, so that $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}_\xi)$, with a covariance matrix that has off-diagonal elements?

Solution.

The best estimate is Gaussian distributed since

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{y} \tag{A.1}$$

shows that $\hat{\boldsymbol{\theta}}$ is a linear transformation of $\boldsymbol{y}$. Then, since $\boldsymbol{y}$ is Gaussian, so is $\hat{\boldsymbol{\theta}}$.

To show that $\hat{\boldsymbol{\theta}}$ is unbiased, we have

$$\begin{aligned}\left\langle\hat{\boldsymbol{\theta}}\right\rangle &= \left\langle\left(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{y}\right\rangle \\ &= \left\langle\left(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{\Phi}\boldsymbol{\theta}^* + \boldsymbol{\xi})\right\rangle \\ &= \left(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\,\boldsymbol{\theta}^* + 0 \\ &= \boldsymbol{\theta}^*\,.\end{aligned}$$

In the last step, we assume that

$$\left\langle\left(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\xi}\right\rangle = \mathbf{0}\,,$$

which is true only if the elements of $\boldsymbol{\xi}$ are not correlated with the measurements (elements Φ_{ij} of the design matrix). The condition is satisfied if $\boldsymbol{\xi}$ is white noise.

We next find the variance, recalling from Section A.7.5 that if $\hat{\boldsymbol{\theta}} = \boldsymbol{A}\boldsymbol{y}$, then Var $(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{y}\boldsymbol{A}^{\mathsf{T}}$. Thus,

$$\begin{aligned}\text{Var}\,\hat{\boldsymbol{\theta}} &= \text{Var}\left[\left(\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{y}\right] \\ &= \underbrace{\left(\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathsf{T}}}_{\boldsymbol{A}}\,(\text{Var}\,\boldsymbol{y})\,\underbrace{\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}\right)^{-1}}_{\boldsymbol{A}^{\mathsf{T}}} \\ &= \sigma^2\left(\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}\right)^{-1}.\end{aligned}$$

Problem A.8.3 Show that for an orthonormal basis in function space, $e_j \cdot e_k = \delta_{jk}$ that

$$\hat{\theta}_k = \boldsymbol{y}\cdot e_k\,, \qquad \text{or } \boldsymbol{y} = \sum_{k=1}^{K}(\boldsymbol{y}\cdot e_k)e_k\,, \tag{A.2}$$

justifying the interpretation of $\boldsymbol{P}^{(K)}$ as a projector matrix in Eq. (A.219).

Solution.

The model $\hat{\boldsymbol{y}} = \sum_{k=1}^{K}\theta_k e_k$, and we want to minimize $\|\boldsymbol{y}-\hat{\boldsymbol{y}}\|^2$. We have

$$\|\boldsymbol{y}-\hat{\boldsymbol{y}}\|^2 \;=\; \boldsymbol{y}^2 - 2\boldsymbol{y}\cdot\hat{\boldsymbol{y}} + \hat{\boldsymbol{y}}^2 \;=\; \boldsymbol{y}^2 - 2\boldsymbol{y}\cdot\sum_{k=1}^{K}\theta_k e_k + \sum_{k=1}^{K}\theta_k^2\,.$$

Taking a derivative with respect to θ_ℓ then gives

$$-2(\boldsymbol{y}\cdot e_\ell) + 2\theta_\ell = 0\,.$$

Solving for θ_ℓ, we have

$$\theta_\ell = \boldsymbol{y}\cdot e_\ell = \sum_{i=1}^{N} y_i\, e_\ell(x_i)\,,$$

as claimed. Thus, each parameter is obtained by projecting the data vector onto the associated basis function.

Problem A.8.4 For the cost function $J(\hat{x}) = \langle|x-\hat{x}|\rangle = \int \mathrm{d}x\,|\hat{x}-x|\,p(x|y)$, show that minimizing J implies that $\hat{x}^*$ is the median of $p(x|y)$. Investigate $J''(\hat{x}^*)$, too.

Solution.

The derivative of the absolute value function is the sign function. Thus,

$$\begin{aligned}\left.\frac{\mathrm{d}J}{\mathrm{d}\hat{x}}\right|_{\hat{x}^*} &= \int_{-\infty}^{\infty}\mathrm{d}x\,\text{sign}(\hat{x}^*-x)\,p(x|y) \\ &= (+1)\int_{-\infty}^{\hat{x}^*}\mathrm{d}x\,p(x|y) + (-1)\int_{\hat{x}^*}^{\infty}\mathrm{d}x\,p(x|y) = 0\,.\end{aligned}$$

Equating the two terms means that

$$\int_{-\infty}^{\hat{x}^*}\mathrm{d}x\,p(x|y) = \int_{\hat{x}^*}^{\infty}\mathrm{d}x\,p(x|y)\,,$$

which defines the median ($\hat{x}^*$). Note that $J''(\hat{x}^*) = 2p(\hat{x}^*|y) \geq 0$, implying that the median minimizes J. (The case $J''(\hat{x}^*) = 0$ would require further investigation.)

Problem A.8.5 For the bottom "box shaped" cost function with a small width Δ, show that minimizing $J(\hat{x})$ implies that $\hat{x}^*$ is the mode of $p(x|y)$.

Solution.

The cost function $J(\hat{x})$ is given by

$$J(\hat{x}) = 1 - \int_{\hat{x}-\Delta}^{\hat{x}+\Delta} \mathrm{d}x\, p(x|y)\,,$$

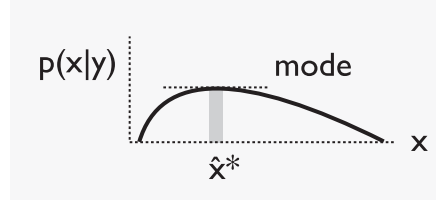
where the value 1 is just an arbitrary positive constant and where Δ is some small interval. In words, we slide a small band centered on $\hat{x}$, of width 2Δ, over the posterior $p(x|y)$ and select the $\hat{x}$ that minimizes J. Clearly, this is the value that *maximizes* the integral. But for small Δ, the integral is

$$\int_{\hat{x}-\Delta}^{\hat{x}+\Delta} \mathrm{d}x\, p(x|y) \approx p(\hat{x}|y)\, 2\Delta\,,$$

which is maximized by

$$\hat{x}^* = \arg\max_x p(x|y)\,,$$

meaning that $\hat{x}^*$ is the *mode* of the posterior distribution, as illustrated below.



Problem A.8.6 **Important details about importance sampling.** We fill in some gaps in our discussion of importance sampling. Assume a scalar variable x.

a. As a warmup, use Eq. (A.227) to show that, for N independent draws x^i from $p(x)$, if $\hat{\varphi} = \frac{1}{N}\sum_{i=1}^{N} \varphi_i$, then $\mathrm{Var}\,\hat{\varphi} = \frac{1}{N}\,\mathrm{Var}\,\varphi$.

b. Importance sampling estimates the average of $\varphi(x)$ over the distribution $p(x)$ using a second, proposal distribution $q(x)$. In general, p and q need not be normalized, but here we assume they are. Then, $\langle\varphi\rangle_p \approx \hat{\varphi} = \sum_{i=1}^{N} w_i\, \varphi(x^i)$, with weights $w_i = \frac{p(x^i)/q(x^i)}{\sum_i p(x^i)/q(x^i)}$. The x^i are N independent draws $x^i \sim q(x)$.

   Show that $\langle\hat{\varphi}\rangle = \langle\varphi\rangle_p$ and $\mathrm{Var}_q[\hat{\varphi}] = \frac{1}{N}(\langle\frac{p\varphi^2}{q}\rangle_p - \langle\varphi\rangle_p^2)$. Show that the variance vanishes when we pick $q(x) = p(x)\,\varphi(x)$, a choice that is valid only for $\varphi(x) > 0$.

c. *Bias of the ratio.* $\hat{\varphi}$ is biased for finite N. To see how this can arise, consider a case of N random variables n_i and d_i (numerator and denominator, in this case). Assume that they are correlated, as this is true for the importance-sampling case. (Why?) Let $\langle n\rangle$ and $\langle d\rangle$ be the mean values of the random

variables (e.g., $\langle n\rangle = \int dn\, n\, p(n)$). Let bars denote arithmetic averages (e.g. $\overline{n} = \frac{1}{N}\sum_{i=1}^{N} n_i$). Then show $\left\langle \frac{\bar{n}}{\bar{d}} \right\rangle \neq \frac{\langle n\rangle}{\langle d\rangle}$ by writing $n_i = \langle n\rangle + \delta n_i$ and giving the lowest-order corrections.

Solution.

a. We first recall that

$$\langle\hat{\varphi}\rangle = \left\langle \frac{1}{N}\sum_{i=1}^{N}\varphi(x^i)\right\rangle = \frac{1}{N}\sum_i \langle\varphi(x^i)\rangle = \left(\frac{1}{N}\right) N\langle\varphi\rangle = \langle\varphi\rangle\,.$$

For the second moment,

$$\begin{aligned}
\langle\hat{\varphi}^2\rangle &= \left\langle \left(\frac{1}{N}\sum_{i=1}^{N}\varphi(x^i)\right)^2 \right\rangle \\
&= \frac{1}{N^2}\left[\underbrace{N\langle\varphi^2\rangle}_{\text{diagonal}} + \underbrace{(N^2-N)\,\langle\varphi(x^i)\,\varphi(x^j)\rangle}_{\text{off-diagonal}}\right] \\
&= \frac{1}{N}\langle\varphi^2\rangle + \frac{N^2-N}{N^2}\underbrace{\langle\varphi(x^i)\rangle\langle\varphi(x^j)\rangle}_{\text{independent}} \\
&= \frac{1}{N}\left(\langle\varphi^2\rangle - \langle\varphi\rangle^2\right) + \langle\varphi\rangle^2\,,
\end{aligned}$$

so that

$$\text{Var}\,\hat{\varphi} = \langle\hat{\varphi}^2\rangle - \langle\hat{\varphi}\rangle^2 = \frac{1}{N}\text{Var}\,\varphi\,, \qquad \text{as claimed.}$$

b. With weights $w(x) = \frac{p(x)}{q(x)}$, we can go through the same steps as in part (a) to show that the $\text{Var}\,\hat{\varphi} = \frac{1}{N}\,\text{Var}\,[\varphi w]$. Thus,

$$\begin{aligned}
\text{Var}_q\,[\hat{\varphi}] &= \frac{1}{N}\text{Var}_q[\varphi(x)\,w(x)] \\
&= \frac{1}{N}\text{Var}_q\left[\frac{\varphi(x)\,p(x)}{q(x)}\right] \\
&= \frac{1}{N}\int \mathrm{d}x\left(\frac{\varphi p}{Q}\right)^2 q(x) - \left(\int \mathrm{d}x\,\frac{\varphi p}{q}q(x)\right)^2 \\
&= \frac{1}{N}\int \mathrm{d}x\left[\frac{\varphi^2 p^2}{q}\right] - \frac{1}{N}\langle\varphi\rangle_p^2
\end{aligned}$$

$$= \frac{1}{N}\int \mathrm{d}x \left[\frac{p\varphi^2}{q}\right] p - \frac{1}{N}\langle\varphi\rangle_p^2$$
$$= \frac{1}{N}\left(\left\langle\frac{p\varphi^2}{q}\right\rangle_p - \langle\varphi\rangle_p^2\right),$$

Selecting $q = p\varphi$, or, equivalently, writing $\varphi = q/p$, we note that

$$\langle\varphi\rangle_p = \int \mathrm{d}x \, \frac{q}{p}\, p = \int \mathrm{d}x \, q = 1\,.$$

Similarly,

$$\left\langle\frac{p\varphi^2}{q}\right\rangle_p = \int \mathrm{d}x \, \frac{pq^2}{qp^2}\, p = \int \mathrm{d}x \, q = 1\,,$$

which implies that $\mathrm{Var}_q[\hat{\varphi}] \propto 1 - 1 = 0$. Intuitively, the variance vanishes because we have selected a distribution $q(x)$ from which every draw, $x \sim q(x)$, will give a value of 1.

As stated in the text, this is not a useful result in itself. But it does mean that you can reduce the variance of your estimator by matching, as well as possible, q to $p\varphi$, so that the range of values x that you find upon drawing from $q(x)$ is as small as possible. Finally, selecting $q = p$ just gives $\mathrm{Var}_q[\hat{\varphi}] = \mathrm{Var}_p[\hat{\varphi}]$, as it must.

c. We have

$$\bar{n} = \frac{1}{N}\sum_{i=1}^{N} n_i$$
$$= \frac{1}{N}\sum_{i=1}^{N} \langle n\rangle + \delta n_i$$
$$= \langle n\rangle + \frac{1}{N}\sum_{i=1}^{N} \delta n_i\,.$$

Then, the ratio of $\bar{n}$ to $\bar{d}$ is

$$\frac{\bar{n}}{\bar{d}} = \frac{\langle n\rangle + \frac{1}{N}\sum_{i=1}^{N}\delta n_i}{\langle d\rangle + \frac{1}{N}\sum_{j=1}^{N}\delta d_j}$$
$$= \frac{\langle n\rangle}{\langle d\rangle}\frac{1 + \frac{1}{N}\sum_{i=1}^{N}\frac{\delta n_i}{\langle n\rangle}}{1 + \frac{1}{N}\sum_{j=1}^{N}\frac{\delta d_j}{\langle d\rangle}}$$
$$= \frac{\langle n\rangle}{\langle d\rangle}\left(1 + \frac{1}{N}\sum_{i=1}^{N}\frac{\delta n_i}{\langle n\rangle}\right)\left(1 + \frac{1}{N}\sum_{j=1}^{N}\frac{\delta d_j}{\langle d\rangle}\right)^{-1}$$
$$= \frac{\langle n\rangle}{\langle d\rangle}\left(1 + \frac{1}{N}\sum_{i=1}^{N}\frac{\delta n_i}{\langle n\rangle}\right)\left(1 - \frac{1}{N}\sum_{j=1}^{N}\frac{\delta d_j}{\langle d\rangle} + \sum_{j=1}^{N}\frac{\delta d_j}{\langle d\rangle}\sum_{k=1}^{N}\frac{\delta d_k}{\langle d\rangle} + \cdots\right)$$

Then, taking ensemble averages we have, to second order,

$$\left\langle \frac{\bar{n}}{\bar{d}} \right\rangle \approx \frac{\langle n \rangle}{\langle d \rangle} \left[1 + \frac{1}{N} \left(\frac{\langle\, \delta d^2 \rangle}{\langle d \rangle^2} - \frac{\langle\, \delta n \delta d \rangle}{\langle n \rangle \langle d \rangle} \right) \right] ,$$

where we have used an argument similar to (a) in writing the averages. More intuitively, we are just using the fact that δn is the standard error of the mean, which is $\sigma_n / \sqrt{N}$.

Finally, note that the calculation depends on having $\delta n \ll \bar{n}$. On the other hand, the general statement that the ratio of finite-N estimators is biased still clearly holds. But you may need to evaluate the bias more carefully in some cases.

Problem A.10.1 **Entropy of a function of a stochastic process**. For any deterministic function $f(\cdot)$ and discrete random variable X, show that $H[f(X)] \leq H(X)$. Hint: Apply the chain rule to $H[X, f(X)]$.

Solution.

From the chain rule for entropy, Eq. (A.259),

$$H[X, f(X)] = H(X) + H[f(X)|X] = H(X) .$$

The last equation follows because if X is given, then the value of the deterministic function $f(X)$ is known and thus has no entropy.

We can also write the chain rule the other way:

$$H[X, f(X)] = H[f(X)] + H[X|f(X)] \geq H[f(X)] .$$

The last identity occurs because $f(X)$ may have multiple inputs. For example, if $f(X) = X^2$, then knowing $f(X)$ means that there are two possible values for X. The multiple values for X means that $H[X|f(X)] \geq 0$ and hence gives the stated inequality.

Putting these two identities together, we have

$$H[f(X)] \leq H(X) .$$

The direct, intuitive statement of the result is that $f(X)$ may map different values of X onto the same functional value, thereby reducing the entropy of $f(X)$. If $f(\cdot)$ is an invertible function, then $H[f(X)] = H(X)$.

Problem A.10.2 **Entropy of a multivariate Gaussian distribution**. The N-dimensional vector $\boldsymbol{X}$ has realizations $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Show that $H(\boldsymbol{X}) = \frac{1}{2} \log \det(2\pi \, \mathrm{e} \, \boldsymbol{\Sigma})$.

Solution.

For $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the distribution $p(\boldsymbol{x})$ is

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^N \det \boldsymbol{\Sigma}}} \, \mathrm{e}^{-\frac{1}{2} \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}} .$$

Then,

$$-\log p(\boldsymbol{x}) = \tfrac{N}{2}\log 2\pi + \tfrac{1}{2}\log\det\boldsymbol{\Sigma} + \tfrac{1}{2}\log\mathrm{e}\;\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\,.$$

The entropy is $H = \langle -\log p\rangle$, which gives

$$H = \tfrac{N}{2}\log 2\pi + \tfrac{1}{2}\log\det\boldsymbol{\Sigma} + \tfrac{1}{2}\log\mathrm{e}\,\langle\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\rangle\,.$$

To evaluate the last term, we transform to coordinates in which

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_N^2 \end{pmatrix}.$$

In that coordinate system,

$$\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x} = \begin{pmatrix} x_1 & \cdots & x_N \end{pmatrix}\begin{pmatrix} 1/\sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_N^2 \end{pmatrix}\cdot\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} = \sum_{i=1}^{N}\frac{x_i^2}{\sigma_i^2}\,,$$

and

$$\langle\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\rangle = \sum_{i=1}^{N}\frac{\sigma_i^2}{\sigma_i^2} = N\,.$$

Substituting this result into the differential entropy expression gives

$$\begin{aligned} H(\boldsymbol{X}) &= \tfrac{N}{2}\log 2\pi + \tfrac{1}{2}\log\det\boldsymbol{\Sigma} + \tfrac{1}{2}(\log\mathrm{e})\,N \\ &= \tfrac{N}{2}\log(2\pi\,\mathrm{e}) + \tfrac{1}{2}\log\det\boldsymbol{\Sigma} \\ &= \tfrac{1}{2}\log(2\pi\,\mathrm{e})^N + \tfrac{1}{2}\log\det\boldsymbol{\Sigma} \\ &= \tfrac{1}{2}\log\det(2\pi\,\mathrm{e}\,\boldsymbol{\Sigma})\,, \end{aligned}$$

where $(2\pi\,\mathrm{e})^N = \det(2\pi\,\mathrm{e}\,\mathbb{I})$ and $\mathbb{I}$ is the $N \times N$ identity matrix. Note that $\det\boldsymbol{\Sigma} > 0$, since the covariance matrix $\boldsymbol{\Sigma}$ is positive definite.

Problem A.10.3 **Chain rule for relative entropy**. Show that $D[p(x,y)\parallel q(x,y)] = D[p(x)\parallel q(x)] + D[p(y|x)\parallel q(y|x)]$; i.e., *coarse graining reduces relative entropy*.

Solution.

We first prove the chain rule for relative entropy:

$$\begin{aligned} D[p(x,y)\parallel q(x,y)] &= \sum_{x,y} p(x,y)\log\frac{p(x,y)}{q(x,y)} \\ &= \sum_{x,y} p(x,y)\log\frac{p(x)\,p(y|x)}{q(x)\,q(y|x)} \\ &= \sum_{x,y} p(x,y)\log\frac{p(x)}{q(x)} + \sum_{x,y} p(x,y)\log\frac{p(y|x)}{q(y|x)} \\ &= D[p(x)\parallel q(x)] + D[p(y|x)\parallel q(y|x)]\,. \end{aligned}$$

This proof is from Cover and Thomas (2006), Theorem 2.5.3.

Because of the Gibbs inequality, Eq. A.277, we have $D[p(y|x) \parallel q(y|x)] \geq 0$. Thus,

$$D[p(x,y) \parallel q(x,y)] \geq D[p(x) \parallel q(x)] \,.$$

That is, the relative entropy is reduced, or held constant, whenever we "integrate out" variables from *both* probability distributions. The result clearly applies when x represents N_x variables and y represents N_y variables.

In practice, the equality condition is seldom encountered. For equality, we must have $D[p(y|x) \parallel q(y|x)] = 0$. From the Gibbs inequality, the relative entropy is zero only when the two probability distributions are the same. The most likely scenario for this to occur is that X and Y are independent random variables, so that $p(y|x) = p(y)$ and $q(y|x) = q(y)$. Then we would further require $p(y) = q(y)$. Since these are fairly artificial circumstances, we can conclude that in almost all cases of interest, coarse graining does indeed reduce relative entropy.

Problem A.10.4 Relative entropy for Gaussians.

a. If $x_1 \sim \mathcal{N}(\mu_1, \sigma^2)$ and $x_2 \sim \mathcal{N}(\mu_2, \sigma^2)$, show $D(p_{x_1} \| p_{x_2}) = (\mu_1 - \mu_2)^2 / 2\sigma^2$.

b. If $x_1 \sim \mathcal{N}(0, \sigma_1^2)$ and $x_2 \sim \mathcal{N}(0, \sigma_2^2)$, show $D(p_{x_1} \| p_{x_2}) = (\sigma_1^2 - \sigma_2^2)/(2\sigma_2^2) + \log \frac{\sigma_2}{\sigma_1}$. Verify that $D(p_{x_1} \| p_{x_2})$ is non-negative.

Solution.

a. We have $\langle x \rangle = \mu_1$, since we average over x_1.

$$\begin{aligned} D(p_{x_1} \| p_{x_2}) &= \left\langle -\frac{(x-\mu_1)^2}{2\sigma^2} + \frac{(x-\mu_2)^2}{2\sigma^2} \right\rangle \\ &= \frac{1}{2\sigma^2} \langle 2x\mu_1 - \mu_1^2 - 2x\mu_2 + \mu_2^2 \rangle \\ &= \frac{1}{2\sigma^2} (2\mu_1^2 - \mu_1^2 - 2\mu_1\mu_2 + \mu_2^2) \\ &= \boxed{\tfrac{1}{2\sigma^2}(\mu_1 - \mu_2)^2} \,. \end{aligned}$$

b. We have $\langle x^2 \rangle = \sigma_1^2$, since $\mu = 0$.

$$\begin{aligned} D(p_{x_1} \| p_{x_2}) &= \left\langle -\frac{x^2}{2\sigma_1^2} + \frac{x^2}{2\sigma_2^2} + \log \frac{\sigma_2}{\sigma_1} \right\rangle \\ &= -\frac{1}{2} + \frac{\sigma_1^2}{\sigma_2^2} + \log \frac{\sigma_2}{\sigma_1} \\ &= \boxed{\tfrac{\sigma_1^2 - \sigma_2^2}{2\sigma_2^2} + \log \tfrac{\sigma_2}{\sigma_1}} \,. \end{aligned}$$

A graph of $D(\sigma_2/\sigma_1)$ is at left (solid trace). Notice that it is indeed non-negative, with a minimum of 0 at $\sigma_1 = \sigma_2$, as claimed.

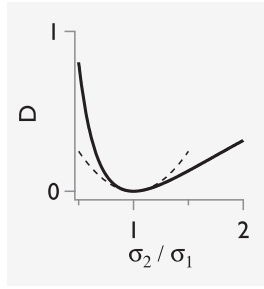
Algebraically, we can define $s = \sigma_2/\sigma_1$, in terms of which

$$D = \frac{1 - s^2}{2s^2} + \log s\,,$$

A series expansion about $s = 1$ gives

$$D = (s-1)^2 + O(s-1)^3$$

and shows, locally, that $D \geq 0$ for $\sigma_1 \approx \sigma_2$. The quadratic approximation is included (dashed line) in the plot below.



Problem A.10.5 Relative entropy and Fisher information. Consider a one-parameter family of probability distributions $P_\theta(x)$ over a discrete variable x. Show that the relative entropy between $P_\theta(x)$ and $P_{\theta+\Delta\theta}(x)$ is given, to lowest order in $\Delta\theta$, by

$$D(P_\theta \| P_{\theta+\Delta\theta}) \approx \frac{(\Delta\theta)^2}{2} \left\langle \left(\frac{d \log P_\theta}{\mathrm{d}\theta} \right)^2 \right\rangle \equiv \frac{(\Delta\theta)^2}{2} F\,,$$

where F is the *Fisher information*.

Solution.

The relative entropy, in nats, is

$$\begin{aligned}
D &= \sum_i P_\theta(x_i) \ln \frac{P_\theta(x_i)}{P_{\theta+\Delta\theta}(x_i)} \\
&= -\sum_i P_\theta(x_i) \ln \frac{P_{\theta+\Delta\theta}(x_i)}{P_\theta(x_i)} \\
&= -\sum_i P_\theta(x_i) \ln \frac{P_\theta(x_i) + (\Delta\theta) P'_\theta(x_i) + \frac{1}{2}(\Delta\theta)^2 P''_\theta(x_i) + \cdots}{P_\theta(x_i)} \\
&= -\sum_i P_\theta(x_i) \ln \left[1 + (\Delta\theta) \frac{P'_\theta(x_i)}{P_\theta(x_i)} + \tfrac{1}{2}(\Delta\theta)^2 \frac{P''_\theta(x_i)}{P_\theta(x_i)} \right] \\
&= -\sum_i P_\theta(x_i) \left[(\Delta\theta) \frac{P'_\theta(x_i)}{P_\theta(x_i)} + \tfrac{1}{2}(\Delta\theta)^2 \left(\frac{P''_\theta(x_i)}{P_\theta(x_i)} - \left(\frac{P'_\theta(x_i)}{P_\theta(x_i)} \right)^2 \right) \right] + \cdots \\
&= \sum_i \left\{ -(\Delta\theta) P'_\theta(x_i) - \tfrac{1}{2}(\Delta\theta)^2 P''_\theta(x_i) + \tfrac{1}{2}(\Delta\theta)^2 P_\theta(x_i) \left(\frac{P'_\theta(x_i)}{P_\theta(x_i)} \right)^2 \right\}
\end{aligned}$$

$$= -(\Delta\theta)\frac{d}{\mathrm{d}\theta}\sum_i P_\theta(x_i) - \tfrac{1}{2}(\Delta\theta)^2\frac{d^2}{\mathrm{d}\theta^2}\sum_i P_\theta(x_i) + \tfrac{1}{2}(\Delta\theta)^2\sum_i P_\theta(x_i)\left(\frac{P'_\theta(x_i)}{P_\theta(x_i)}\right)^2$$

$$= 0 + 0 + \tfrac{1}{2}(\Delta\theta)^2\left\langle\left(\frac{d\ln P_\theta}{\mathrm{d}\theta}\right)^2\right\rangle,$$

where $\sum_i P_\theta(x_i) = 1$, and the first two terms vanish because we take $\frac{d}{\mathrm{d}\theta}$ of a constant. Note that we first Taylor expanded P_θ and then used $\ln(1+x) = x - \frac{1}{2}x^2 + \cdots$.

Problem A.10.6 **Numerical example**. Consider two variables X and Y with two-letter alphabets $\{1, 2\}$. The joint probabilities $P(x, y)$ are given in the table at left. Calculate

$P(x,y)$	$x=1$	$x=2$
$y=1$	0.3	0.3
$y=2$	0.4	0

a. the marginal distributions $P(x)$ and $P(y)$;
b. the entropies $H(X)$, $H(Y)$, and $H(X, Y)$;
c. conditional entropies $H(Y|x = 1)$ and $H(Y|x = 2)$. Show that $H(Y|x = 1) > H(Y)$. What does this mean?
d. the average conditional entropy $H(Y|X)$. Show that $H(Y|X) < H(Y)$, and reconcile this result with the previous part.
e. the mutual information $I(X; Y)$.

Solution.

a. $P(x = 1) = 0.3 + 0.4 = 0.7$. $P(x = 2) = 0.3 + 0 = 0.3$. $P(y = 1) = 0.6$. $P(y = 2) = 0.4$.

b. $H(X) = -0.7\log 0.7 - 0.3\log 0.3 \approx 0.88$ bits. $H(Y) \approx 0.97$ bits. Finally, the joint entropy $H(X, Y) \approx 1.57$ bits.

c. $H(Y|x = 1) = -\sum_y P(y|1)\log P(y|1)$. Then use $P(y|1) = P(y, 1)/P(x = 1)$. This gives $H(Y|x = 1) \approx 0.98$ bits, which is indeed bigger than $H(Y)$. We also calculate $H(Y|x = 2) = 0$ bits.

   What does it mean to have $H(Y|x = 1) > H(Y)$? Informally, the particular observation $x = 1$ is *confusing*, making us more uncertain about Y than before the observation. The conditional probabilities are more "even" than the unconditional ones. By contrast, the result that $H(Y|x = 2) = 0$ means that if we observe $x = 2$, then we *know* that $y = 1$, since there is zero probability to observe $y = 2$, given $x = 2$.

d. We can calculate $H(Y|X)$ two ways. First, we have the previous result. $H(Y|X) = P(x = 1)H(Y|x = 1) + P(x = 2)H(Y|x = 2) = 0.7 \times 0.98 + 0 \approx 0.69$ bits.
   Alternatively, $H(Y|X) = H(Y, X) - H(X) = 1.57 - 0.88 \approx 0.69$ bits.

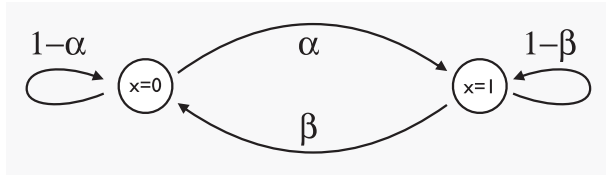
e. Finally, $I(X; Y) = H(X) + H(Y) - H(X, Y) \approx 0.28$ bits.

Problem A.10.7 **Entropy rate of a Markov chain**. Generalize the result for the entropy rate of a symmetric 2-state Markov chain (Eq. A.262) to the asymmetric case, with transition probabilities α and β, using Eq. (A.271).

a. Show that $\mathcal{H}(X) = \left(\frac{\beta}{\alpha+\beta}\right) H_2(\alpha) + \left(\frac{\alpha}{\alpha+\beta}\right) H_2(\beta)$.
b. Which values of (α, β) maximize and which minimize $\mathcal{H}(X)$?
c. What is wrong with the following argument: For $\alpha = \beta = \varepsilon$, the single-symbol entropy is 1 bit. The time spent in each state before jumping to the other is typically ε^{-1}. Therefore, the entropy rate should be $\mathcal{H} \approx 1$ bit/$(\varepsilon^{-1}) = \varepsilon$ bits/trial.

Solution.

a. The state diagram is shown below.



The transition matrix is then

$$\boldsymbol{P} = \begin{pmatrix} 1-\alpha & \beta \\ \alpha & 1-\beta \end{pmatrix}.$$

The entropy rate is

$$\mathcal{H}(X) = H(X_2|X_1) = -\sum_{i,j} P(x_i)\, P(x_j|x_i)\, \log P(x_j|x_i)\,.$$

The probabilities $P(x_i)$ are just the steady-state occupation probabilities and satisfy the marginal probability law $P(x_i) = \sum_j P(x_i, x_j) = \sum_j P(x_j)P(x_i|x_j)$.

$$P(0) = P(0)\,(1-\alpha) + P(1)\beta\,, \qquad P(1) = P(0)\alpha + P(1)\,(1-\beta)\,.$$

Using $P(0) + P(1) = 1$, we find

$$P(0) = \frac{\beta}{\alpha+\beta}\,, \qquad P(1) = \frac{\alpha}{\alpha+\beta}\,.$$

Then

$$\begin{aligned} \mathcal{H}(X) &= -\left(\frac{\beta}{\alpha+\beta}\right)[(1-\alpha)\log(1-\alpha) + \alpha\log\alpha] \\ &\quad -\left(\frac{\alpha}{\alpha+\beta}\right)[(1-\beta)\log(1-\beta) + \beta\log\beta] \\ &= \left(\tfrac{\beta}{\alpha+\beta}\right) H_2(\alpha) + \left(\tfrac{\alpha}{\alpha+\beta}\right) H_2(\beta)\,. \end{aligned}$$

b. It is easy to see that $\alpha = \beta = 0.5$ maximizes $\mathcal{H}(X) = \log 2$. We can understand this limit by noting $x = 0$ and 1 occur with equal probability (0.5) and that we learn the full 1 bit of information each time we measure them, regardless of what state they were previously.

The minimum value is $\mathcal{H}(X) = 0$, which occurs when either α or $\beta = 0$. This also makes sense: If $\alpha = 0$, then $x = 0$ is an absorbing state. Once the system

gets to $x = 0$, it will stay there. The entropy rate, an average over infinite time, is zero because we know that the system is eventually always in $x = 0$. The value of β merely determines the mean lifetime of an initial state that happens to be $x = 1$. Note that the same statements apply (with $0 \leftrightarrow 1$) when $\beta = 0$.

The situation $\alpha = 1$ is different. This case means that if the system is in $x = 0$, we are certain that it will be in $x = 1$ the next time; however, depending on the value of β, the system will take a greater or lesser time to return to $x = 0$.

c. For $\alpha = \beta = \varepsilon \ll 1$, the simple argument gives $\mathcal{H} = \varepsilon$ bits/trial. The exact answer is $\mathcal{H} = -\varepsilon \ln \varepsilon - (1 - \varepsilon) \ln(1 - \varepsilon) \approx -\varepsilon \ln \varepsilon$ is greater by a factor $\ln 1/\varepsilon$.

The simple, "mean-field" solution neglects the information contained in the fluctuations of transition times. This is easily understood by comparing graphical representations of representative time series generated from the "mean-field" solution and from the exact solution. The top plot shows a Markov sequence of 1000 points, with a transition probability $\varepsilon = 0.1$. Thus, there is, on average, a jump every 10 steps. The bottom plot shows a similar sequence, but not grouped in blocks of $\varepsilon^{-1} = 10$ steps. Each block is chosen with $P(0) = P(1) = 0.5$. Qualitatively, it is obvious that there is less information contained in the sequence (fewer transitions).

Problem A.10.8 **Mutual information of a bivariate Gaussian**. Derive Eq. (A.298).

Solution.

We calculate the mutual information as $I(X;Y) = H(X) + H(Y) - H(X,Y)$.

For $x \sim \mathcal{N}(0, \sigma_x^2)$, the entropy $H(X) = \sqrt{2\pi\,\mathrm{e}} + \frac{1}{2} \log \sigma_x^2$.

Likewise, for $y \sim \mathcal{N}(0, \sigma_y^2)$, the entropy $H(Y) = \sqrt{2\pi\,\mathrm{e}} + \frac{1}{2} \log \sigma_y^2$.

For a bivariate Gaussian $(x, y) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. From Eq. (A.193), we have,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2, \end{pmatrix}.$$

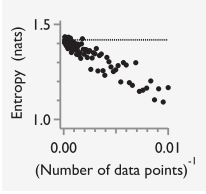
Then

$$\begin{aligned} H(X,Y) &= 2 \log \sqrt{2\pi\,\mathrm{e}} + \tfrac{1}{2} \log \det \boldsymbol{\Sigma} \\ &= 2 \log \sqrt{2\pi\,\mathrm{e}} + \tfrac{1}{2} \log[\sigma_x^2 \sigma_y^2 (1 - \rho^2)]\,. \end{aligned}$$

Finally,

$$I(X;Y) = \left(\log\sqrt{2\pi\,\mathrm{e}} + \tfrac{1}{2}\log\sigma_x^2\right) + \left(\log\sqrt{2\pi\,\mathrm{e}} + \tfrac{1}{2}\log\sigma_y^2\right) - \left(2\log\sqrt{2\pi\,\mathrm{e}} + \tfrac{1}{2}\log[\sigma_x^2\sigma_y^2(1-\rho^2)]\right)$$
$$= -\tfrac{1}{2}\log\left(1-\rho^2\right) .$$

Problem A.10.9 Estimating entropy from limited experimental data is tricky.



a. Simulate N draws $x \sim \mathcal{N}(0,1)$ of the random variable X, and histogram the results. Let N_i be the number of data points in bin i and let $f_i = N_i/N$ be the corresponding frequency estimates. If p_i are the exact frequencies, then $H = -\sum p_i \ln p_i$ and a "naive" estimator is $\hat{H} = -\sum f_i \ln f_i$, in nats. Confirm the plot of $\hat{H}$ versus N^{-1} at left, where the dashed line is the expected result, $\frac{1}{2}\ln 2\pi\,\mathrm{e} \approx 1.42$. Use 100 bins, over $x \in [-5,5]$. The linearity of the plot implies an $O(1/N)$ bias. The bias is large: more than 20% for 100 points.
b. Use Jensen's inequality to show that the bias is general: $\langle\hat{H}\rangle \leq H$.
c. Taylor expand $\hat{H}$ about the exact H. Assuming that N_i is Poisson distributed about the expected value, Np_i, show that $\langle\hat{H}\rangle = H(X) - (1/2)(N_{\text{bin}}/N) + O(N^{-2})$, which has the $O(N^{-1})$ dependence noted in part (a). An obvious way to deal with finite-N bias is to extrapolate N^{-1} to zero, as shown in the figure. Appendix A.8 of Bialek (2012) discusses more sophisticated estimators with smaller bias.

Solution.

a. See the book website for code.
b. The frequencies f_i are unbiased estimators of the probabilities p_i. That is, $\langle f_i\rangle = p_i$. Then, $\langle\hat{H}\rangle = \langle H(f)\rangle \leq H(\langle f\rangle) = H(p)$, where we use the Jensen inequality and the fact that the entropy is a convex function of its arguments.
c. Let $N_i \sim \text{Poisson}(Np_i)$. Then $\langle N_i\rangle = \langle(\delta N_i)^2\rangle = Np_i$. The empirical frequencies for bin i are $f_i = N_i/N$. Then $\langle f_i\rangle = p_i$ and $\langle(\delta f_i)^2\rangle = Np_i/N^2 = p_i/N$ and

$$\hat{H} = H(X) + \sum_{i=1}^{N_{\text{bin}}} \left.\frac{\partial\hat{H}}{\partial f_i}\right|_{p_i} (\delta f_i) + \frac{1}{2}\left.\frac{\partial^2\hat{H}}{\partial f_i^2}\right|_{p_i} (\delta f_i)^2 + \cdots .$$

We have

$$-\frac{\partial\hat{H}}{\partial f_i} = \ln f_i + 1\,, \qquad -\frac{\partial^2\hat{H}}{\partial f_i^2} = \frac{1}{f_i}\,.$$

Substituting and taking the expectation value then gives

$$\langle\hat{H}\rangle = H(X) + 0 - \frac{1}{2}\sum_{i=1}^{N_{\text{bin}}} \frac{1}{p_i}\frac{p_i}{N} + \cdots$$
$$\approx H(X) - \frac{1}{2}\frac{N_{\text{bin}}}{N}\,,$$

where $\langle \delta f_i \rangle = 0$ eliminates the linear term. The next term would be $O\langle (\delta f)^3 \rangle = O(N^{-2})$, since the Poisson distribution is skew, and the central 3rd moment is $\langle (\delta f_i)^3 \rangle = Np_i/N^3$.

Note that while reducing the number of bins would seem to lower the bias, the situation is more complicated because coarsening bins increases discretization bias. Accurate estimation with the naive estimator $\hat{H}$ requires taking a large N_{bin} and then an even larger N. Although extrapolation ($N \to \infty$) helps, as mentioned in the problem statement, other estimators can do a better job. See Appendix A.8 of Bialek (2012) for a discussion of all these issues.

Problem A.10.10 Mutual information for a nonlinear relationship. Let $y = x^2$ and let $x \sim \mathcal{N}(0, 1)$. Show the following:

a. The linear correlation coefficient $\rho = 0$.
b. $I(X; Y) \approx 1.13$ bits. Why is $I < H(X) \approx 2.05$ bits? Hint: Show that $H(Y|X) = 0$.

Solution.

a. The linear covariance ρ is proportional to

$$\begin{aligned} &\left\langle (x - \langle x \rangle)(x^2 - \langle x^2 \rangle) \right\rangle \\ &= \left\langle (x - 0)(x^2 - \sigma_x^2) \right\rangle \\ &= \left\langle x^3 \right\rangle - \langle x \rangle \sigma_x^2 \\ &= 0 - 0 = 0 \,. \end{aligned}$$

More simply, x is odd and x^2 even. The covariance between even and odd functions of a random variable with even probability distribution function vanishes.

b. From Eq. (A.288), a definition of mutual information is $I(X; Y) = H(Y) - H(Y|X)$. But the latter term is zero, because if we know x exactly, we know $y = x^2$ exactly. Thus, a measurement of X does not tell us anything more about Y.

We recall that $H(X) = \log \sqrt{2\pi \mathrm{e}} \approx 2.05$ bits. We also recall, from Eq. (A.161), that

$$p(y) = \begin{cases} \frac{1}{\sqrt{2\pi y}} \, \mathrm{e}^{-\frac{y}{2}} \,. & y > 0 \\ 0 & y \le 0 \end{cases}$$

Then, using Mathematica to do the integral gives

$$H(Y) = -\int_0^\infty \mathrm{d}y \, p(y) \log p(y) = \frac{1 - \gamma + \log \pi}{2 \log 2} \approx 1.13 \text{ bits} \,,$$

where $\gamma \approx 0.577$ is the Euler-Mascheroni constant. Intuitively, a measurement y tells us only "half as much" about x, because $x = \pm\sqrt{y}$, and there is a roughly one-bit ambiguity. The relationship is not exact because it also depends on local stretching of coordinates.

Problem A.10.11 Noisy measurements and Gaussian channels. Derive Eq. (A.293).

Solution.

We first note that

$$p(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}}\, \mathrm{e}^{-\frac{x^2}{2\sigma_x^2}} \qquad p(y) = \frac{1}{\sqrt{2\pi\sigma_y^2}}\, \mathrm{e}^{-\frac{y^2}{2\sigma_y^2}} \qquad p(y|x) = \frac{1}{\sqrt{2\pi\sigma_\xi^2}}\, \mathrm{e}^{-\frac{(y-x)^2}{2\sigma_\xi^2}},$$

where $\sigma_y^2 = \sigma_x^2 + \sigma_\xi^2$. The mutual information is then

$$\begin{aligned} I(X;Y) &= \iint \mathrm{d}x\, \mathrm{d}y\, p(x,y)\, \log\left(\frac{p(y|x)}{p(y)}\right) \\ &= \frac{1}{\ln 2}\left\langle \ln\frac{\sqrt{2\pi\sigma_y^2}}{\sqrt{2\pi\sigma_\xi^2}} - \frac{(y-x)^2}{2\sigma_\xi^2} + \frac{y^2}{2\sigma_y^2}\right\rangle \\ &= \frac{1}{\ln 2}\left(\frac{1}{2}\ln\frac{\sigma_y^2}{\sigma_\xi^2} - \frac{1}{2} + \frac{1}{2}\right) \\ &= \frac{1}{\ln 2}\left(\frac{1}{2}\right)\ln\left(1 + \frac{\sigma_x^2}{\sigma_\xi^2}\right) \\ &= \tfrac{1}{2}\log\left(1 + \tfrac{\sigma_x^2}{\sigma_\xi^2}\right). \end{aligned}$$

Problem A.10.12 Measurements of noisy, correlated variables. Derive Eq. (A.294). Hint: Follow Problem A.10.11, and use Tr log = log det (see Problem A.1.21).

Solution

This problem is just the multivariate analog of Problem A.10.11, and the solution parallels the steps in that case. We first recall (Eq. A.194) the following multivariate Gaussian distributions:

$$\begin{aligned} p(X^N) = p(\boldsymbol{x}) &= \frac{1}{\sqrt{(2\pi)^N \det \boldsymbol{S}_{xx}}} \exp\left[-\frac{1}{2}\boldsymbol{x}^\mathsf{T}\left(\boldsymbol{S}_{xx}^{-1}\right)\boldsymbol{x}\right], \\ p(Y^N) = p(\boldsymbol{y}) &= \frac{1}{\sqrt{(2\pi)^N \det \boldsymbol{S}_{yy}}} \exp\left[-\frac{1}{2}\boldsymbol{y}^\mathsf{T}\left(\boldsymbol{S}_{yy}^{-1}\right)\boldsymbol{y}\right], \\ p(Y^N|X^N) = p(\boldsymbol{y}|\boldsymbol{x}) &= \frac{1}{\sqrt{(2\pi)^N \det \boldsymbol{\Xi}}} \exp\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{G}\boldsymbol{x})^\mathsf{T}\,\boldsymbol{\Xi}^{-1}\,(\boldsymbol{y}-\boldsymbol{G}\boldsymbol{x})\right], \end{aligned}$$

where $(\boldsymbol{S}_{xx})_{ij} = \langle x_i\, x_j\rangle$, $(\boldsymbol{S}_{yy})_{ij} = \langle y_i\, y_j\rangle$, and $\boldsymbol{\Xi}_{ij} = \langle \xi_i\, \xi_j\rangle$. We also introduce, for this problem, the vector notation $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$, etc.

The mutual information is then

$$\begin{aligned} I(X^N; Y^N) &= \iint \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{y}\; p(\boldsymbol{x}, \boldsymbol{y})\, \log\left(\frac{p(\boldsymbol{y}|\boldsymbol{x})}{p(\boldsymbol{y})}\right) \\ &= \frac{1}{\ln 2}\left\langle \ln\, \sqrt{\frac{\det \boldsymbol{S}_{yy}}{\det \boldsymbol{\Xi}}} - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{G}\boldsymbol{x})^{\mathsf{T}}\, \boldsymbol{\Xi}^{-1}\, (\boldsymbol{y} - \boldsymbol{G}\boldsymbol{x}) + \frac{1}{2}\boldsymbol{y}^{\mathsf{T}}\, \left(\boldsymbol{S}_{yy}^{-1}\right)\, \boldsymbol{y}\right\rangle \\ &= \frac{1}{\ln 2}\left(\frac{1}{2}\ln\, \frac{\det \boldsymbol{S}_{yy}}{\det \boldsymbol{\Xi}} - \frac{1}{2}(\boldsymbol{\Xi}^{-1}\, \boldsymbol{\Xi}) + \frac{1}{2}(\boldsymbol{S}_{yy}^{-1}\, \boldsymbol{S}_{yy})\right) \\ &= \frac{1}{\ln 2}\left(\frac{1}{2}\ln\, \frac{\det \boldsymbol{S}_{yy}}{\det \boldsymbol{\Xi}}\right) \\ &= \frac{1}{2}\log\, \frac{\det \boldsymbol{S}_{yy}}{\det \boldsymbol{\Xi}}\,. \end{aligned}$$

We can express $\boldsymbol{S}_{yy}$ in terms of $\boldsymbol{S}_{xx}$ and $\boldsymbol{\Xi}$ via

$$\begin{aligned} \boldsymbol{S}_{yy} = \langle \boldsymbol{y}\, \boldsymbol{y}^{\mathsf{T}}\rangle &= \left\langle (\boldsymbol{G}\boldsymbol{x} + \boldsymbol{\xi})\, (\boldsymbol{G}\boldsymbol{x} + \boldsymbol{\xi})^{\mathsf{T}}\right\rangle \\ &= \left\langle (\boldsymbol{G}\boldsymbol{x} + \boldsymbol{\xi})\, (\boldsymbol{x}^{\mathsf{T}}\boldsymbol{G}^{\mathsf{T}} + \boldsymbol{\xi}^{\mathsf{T}})\right\rangle \\ &= \boldsymbol{G}\langle \boldsymbol{x}\, \boldsymbol{x}^{\mathsf{T}}\rangle \boldsymbol{G}^{\mathsf{T}} + \langle \boldsymbol{\xi}\, \boldsymbol{\xi}^{\mathsf{T}}\rangle + \cancelto{0}{\text{crossterms}} \\ &= \boldsymbol{G}\, \boldsymbol{S}_{xx}\, \boldsymbol{G}^{\mathsf{T}} + \boldsymbol{\Xi}\,. \end{aligned}$$

Then,

$$\begin{aligned} I(X^N; Y^N) &= \frac{1}{2}\left[\log\det\left(\boldsymbol{G}\, \boldsymbol{S}_{xx}\, \boldsymbol{G}^{\mathsf{T}} + \boldsymbol{\Xi}\right) - \log\det\, \boldsymbol{\Xi}\right] \\ &= \frac{1}{2}\left\{\mathrm{Tr}\left[\log\left(\boldsymbol{G}\, \boldsymbol{S}_{xx}\, \boldsymbol{G}^{\mathsf{T}} + \boldsymbol{\Xi}\right)\right] - \mathrm{Tr}\left[\log\, \boldsymbol{\Xi}\right]\right\} \\ &= \frac{1}{2}\mathrm{Tr}\left[\log\left(\boldsymbol{G}\, \boldsymbol{S}_{xx}\, \boldsymbol{G}^{\mathsf{T}} + \boldsymbol{\Xi}\right) - \log\, \boldsymbol{\Xi}\right] \\ &= \frac{1}{2}\mathrm{Tr}\log\left[\boldsymbol{\Xi}^{-1}\left(\boldsymbol{G}\, \boldsymbol{S}_{xx}\, \boldsymbol{G}^{\mathsf{T}} + \boldsymbol{\Xi}\right)\right] \\ &= \tfrac{1}{2}\mathrm{Tr}\log\left[\mathbb{I} + \boldsymbol{\Xi}^{-1}\left(\boldsymbol{G}\, \boldsymbol{S}_{xx}\, \boldsymbol{G}^{\mathsf{T}}\right)\right]\,, \end{aligned}$$

where Tr log = log det (see Problem A.1.21).

Problem A.10.13 **Discrete memoryless channel (DMC)**. Prove Eq. (A.304).

Solution.

The solution follows Massey (1990).

$$\begin{aligned} P(x^N, y^N) &= \prod_{k=1}^{N} P(x_k, y_k | x^{k-1}, y^{k-1}) && \text{chain rule for probabilities} \\ &= \prod_{k=1}^{N} P(x_k | x^{k-1}, y^{k-1})\, P(y_k | x^k, y^{k-1}) \end{aligned}$$

$$= \prod_{k=1}^{N} P(x_k|x^{k-1})\, P(y_k|x^k, y^{k-1}) \qquad \text{used without feedback}$$

$$= P(x^N) \prod_{k=1}^{N} P(y_k|x^k, y^{k-1}) \qquad \text{chain rule}\,.$$

If $P(x^N) \neq 0$, we can divide by it and express the relation in terms of conditional probabilities:

$$P(y^N|x^N) = \prod_{k=1}^{N} P(y_k|x^k, y^{k-1})$$

$$= \prod_{k=1}^{N} P(y_k|x_k) \qquad \text{memoryless}\,,$$

Thus, the intuitive definition of a discrete memoryless channel also assumes that it is used without feedback. If not, we are led to concepts such as *directed information*, as discussed in Chapter 15.