Contents

Appendix A

Mathematical foundations

Our goal here is to present the basic results and definitions from linear algebra, probability theory, information theory and computational complexity that serve as the mathematical foundations for pattern recognition. We will try to give intuitive insight whenever appropriate, but do not attempt to prove these results; systematic expositions can be found in the references.

A.1 Notation

Here are the terms and notation used throughout the book. In addition, there are numerous specialized variables and functions whose definitions and usage should be clear from the text.

variables, symbols and operations	
$\simeq$	approximately equal to
$\equiv$	equivalent to (or defined to be)
$\propto$	proportional to
∞	infinity
$x \rightarrow a$	x approaches a
$t \leftarrow t+1$	in an algorithm: assign to variable t the new value $t+1$
$\lim\limits_{x \rightarrow a} f(x)$	the value of $f(x)$ in the limit as x approaches a
$\arg\max\limits_{x} f(x)$	the value of x that leads to the maximum value of $f(x)$
$\arg\min\limits_{x} f(x)$	the value of x that leads to the minimum value of $f(x)$
$\lceil x \rceil$	ceiling of x, i.e., the least integer not smaller than x (e.g., $\lceil 3.5 \rceil = 4$)
$\lfloor x \rfloor$	floor of x, i.e., the greatest integer not larger than x (e.g., $\lfloor 3.5 \rfloor = 3$)
$m \bmod n$	m modulo n, the remainder when m is divided by n (e.g., 7 mod 5 = 2)
$\ln(x)$	logarithm base e, or natural logarithm of x
$\log(x)$	logarithm base 10 of x
$\log_2(x)$	logarithm base 2 of x

$\exp[x]$ or e^x	exponential of x, i.e., e raised to the power of x
$\partial f(x)/\partial x$	partial derivative of f with respect to x
$\int_a^b f(x)dx$	the integral of $f(x)$ between a and b. If no limits are written, the full space is assumed
$F(x;\ \theta)$	function of x, with implied dependence upon θ
$\blacksquare$	Q.E.D., quod erat demonstrandum ("which was to be proved") — used to signal the end of a proof

mathematical operations

$<x>$	expected value of random variable x
$\bar{x}$	mean or average value of x
$\mathcal{E}[f(x)]$	the expected value of function $f(x)$ where x is a random variable
$\mathcal{E}_y[f(x,y)]$	the expected value of function over several variables, $f(x,y)$, taken over a subset y of them
$\mathrm{Var}_f[\cdot]$	the variance, i.e., $\mathcal{E}_f[(x-\mathcal{E}_f[x])^2]$
$\sum_{i=1}^{n} a_i$	the sum from $i=1$ to n: $a_1+a_2+...+a_n$
$\prod_{i=1}^{n} a_i$	the product from $i=1$ to n: $a_1 \times a_2 \times ... \times a_n$
$f(x) \star g(x)$	convolution of $f(x)$ with $g(x)$

vectors and matrices

$\mathbf{R}^d$	d-dimensional Euclidean space
$\mathbf{x}, \boldsymbol{A}, ...$	boldface is used for (column) vectors and matrices
$\mathbf{f}(x)$	vector-valued function (note the boldface) of a scalar
$\mathbf{f}(\mathbf{x})$	vector-valued function (note the boldface) of a vector
$\mathbf{I}$	identity matrix, square matrix having 1s on the diagonal and 0 everywhere else
$\mathbf{1}_i$	vector of length i consisting solely of 1's
$diag(a_1, a_2, ..., a_d)$	matrix whose diagonal elements are $a_1, a_2, ..., a_d$, and off-diagonal elements 0
$\mathbf{x}^t$	transpose of vector $\mathbf{x}$
$\|\mathbf{x}\|$	Euclidean norm of vector $\mathbf{x}$
$\mathbf{\Sigma}$	covariance matrix
$\mathrm{tr}[\mathbf{A}]$	the trace of $\mathbf{A}$, i.e., the sum of its diagonal components: $\mathrm{tr}[\mathbf{A}] = \sum_{i=1}^{d} a_{ii}$
$\mathbf{A}^{-1}$	the inverse of matrix $\mathbf{A}$
$\mathbf{A}^{\dagger}$	pseudoinverse of matrix $\mathbf{A}$
$\|\mathbf{A}\|$ or $\mathrm{Det}[\mathbf{A}]$	determinant of $\mathbf{A}$
λ	eigenvalue
$\mathbf{e}$	eigenvector
$\mathbf{u}_i$	unit vector in the ith direction in Euclidean space

sets

$\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, ...$	"Calligraphic" font generally denotes sets or lists, e.g., data set $\mathcal{D} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$
$\mathbf{x} \in \mathcal{D}$	$\mathbf{x}$ is an element of set $\mathcal{D}$
$\mathbf{x} \notin \mathcal{D}$	$\mathbf{x}$ is not an element of set $\mathcal{D}$
$\mathcal{A} \cup \mathcal{B}$	union of two sets, i.e., the set containing all elements of $\mathcal{A}$ and $\mathcal{B}$
$\lvert\mathcal{D}\rvert$	the cardinality of set $\mathcal{D}$, i.e., the number of (possibly non-distinct) elements in it; occassionally written card$\lvert\mathcal{D}\rvert$
$\max_x[\mathcal{D}]$	the maximum x value in set $\mathcal{D}$

probability, distributions and complexity

ω	state of nature
$P(\cdot)$	probability
$p(\cdot)$	probability density
$P(a, b)$	the joint probability , i.e., the probability of having both a and b
$p(a, b)$	the joint probability density, i.e., the probability density of having both a and b
$\Pr\{\cdot\}$	the probability of a condition being met, e.g., $\Pr\{x < x_0\}$ means the probability that x is less than x_0
$p(\mathbf{x}\|\boldsymbol{\theta})$	the conditional probability density of $\mathbf{x}$ given $\boldsymbol{\theta}$
$\mathbf{w}$	weight vector
$\lambda(\cdot, \cdot)$	loss function
$\nabla = \begin{pmatrix} \frac{d}{dx_1} \\ \frac{d}{dx_2} \\ \vdots \\ \frac{d}{dx_d} \end{pmatrix}$	gradient operator in $\mathbf{R}^d$, sometimes written $grad$
$\nabla_{\boldsymbol{\theta}} = \begin{pmatrix} \frac{d}{d\theta_1} \\ \frac{d}{d\theta_2} \\ \vdots \\ \frac{d}{d\theta_d} \end{pmatrix}$	gradient operator in $\boldsymbol{\theta}$ coordinates, sometimes written $grad_{\boldsymbol{\theta}}$
$\hat{\boldsymbol{\theta}}$	maximum likelihood value of $\boldsymbol{\theta}$
$\sim$	"has the distribution," e.g., $p(x) \sim N(\mu, \sigma^2)$ means that the density of x is normal, with mean μ and variance σ^2
$N(\mu, \sigma^2)$	normal or Gaussian distribution with mean μ and variance σ^2
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	multidimensional normal or Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$U(x_l, x_u)$	a one-dimensional uniform distribution between x_l and x_u
$U(\mathbf{x}_l, \mathbf{x}_u)$	a d-dimensional uniform density, i.e., uniform density within the smallest axes-aligned bounding box that contains both $\mathbf{x}_l$ and $\mathbf{x}_u$, and zero elsewhere
$T(\mu, \delta)$	triangle distribution, having center μ and full half-width δ
$\delta(x)$	Dirac delta function
$\Gamma(\cdot)$	Gamma function
$n!$	n factorial $= n \times (n-1) \times (n-2) \times ... \times 1$
$\binom{n}{k} = \frac{n!}{k!(n-k)!}$	binomial coefficient, n choose k for n and k integers

$O(h(x))$	big oh order of $h(x)$
$\mathbf{\Theta}(h(x))$	big theta order of $h(x)$
$\mathbf{\Omega}(h(x))$	big omega order of $h(x)$
$\sup_x f(x)$	the supremum value of $f(x)$ — the global maximum of $f(x)$ over all values of x

A.2 Linear algebra

A.2.1 Notation and preliminaries

A d-dimensional column vector $\mathbf{x}$ and its transpose $\mathbf{x}^t$ can be written as

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \qquad \text{and} \quad \mathbf{x}^t = (x_1 \ x_2 \ \dots \ x_d), \tag{1}$$

where all components can take on real values. We denote an $n \times d$ (rectangular) matrix $\mathbf{M}$ and its $d \times n$ transpose $\mathbf{M}^t$ as

$$\mathbf{M} = \begin{pmatrix} m_{11} & m_{12} & m_{13} & \dots & m_{1d} \\ m_{21} & m_{22} & m_{23} & \dots & m_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n2} & m_{n3} & \dots & m_{nd} \end{pmatrix} \text{and} \tag{2}$$

$$\mathbf{M}^t = \begin{pmatrix} m_{11} & m_{21} & \dots & m_{n1} \\ m_{12} & m_{22} & \dots & m_{n2} \\ m_{13} & m_{23} & \dots & m_{n3} \\ \vdots & \vdots & \ddots & \vdots \\ m_{1d} & m_{2d} & \dots & m_{nd} \end{pmatrix}. \tag{3}$$

In other words, the jith entry of $\mathbf{M}^t$ is the ijth entry of $\mathbf{M}$.

A square ($d \times d$) matrix is called symmetric if its entries obey $m_{ij} = m_{ji}$; it is called skew-symmetric (or anti-symmetric) if $m_{ij} = -m_{ji}$. A general matrix is called non-negative if $m_{ij} \geq 0$ for all i and j. A particularly important matrix is the *identity matrix*, $\mathbf{I}$ — a $d \times d$ (square) matrix whose diagonal entries are 1's, and all other entries 0. The *Kronecker delta* function or Kronecker symbol, defined as

IDENTITY MATRIX

KRONECKER DELTA

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

can serve to define the entries of an identity matrix. A general diagonal matrix (i.e., one having 0 for all off diagonal entries) is denoted $diag(m_{11}, m_{22}, ..., m_{dd})$, the entries being the successive elements $m_{11}, m_{22}, \dots, m_{dd}$. Addition of vectors and of matrices is component by component.

We can multiply a vector by a matrix, $\mathbf{Mx} = \mathbf{y}$, i.e.,

$$\begin{pmatrix} m_{11} & m_{12} & \dots & m_{1d} \\ m_{21} & m_{22} & \dots & m_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n2} & \dots & m_{nd} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{pmatrix}, \tag{5}$$

where

$$y_j = \sum_{i=1}^{d} m_{ji} x_i. \tag{6}$$

Note that the number of columns of $\mathbf{M}$ must equal the number of rows of $\mathbf{x}$. Also, if $\mathbf{M}$ is not square, the dimensionality of $\mathbf{y}$ differs from that of $\mathbf{x}$.

A.2.2 Inner product

The *inner product* of two vectors having the same dimensionality will be denoted here as $\mathbf{x}^t\mathbf{y}$ and yields a scalar:

INNER PRODUCT

$$\mathbf{x}^t\mathbf{y} = \sum_{i=1}^{d} x_i y_i = \mathbf{y}^t\mathbf{x}. \tag{7}$$

It is sometimes also called the *scalar product* or *dot product* and denoted $\mathbf{x} \bullet \mathbf{y}$, or more rarely (x, y). The *Euclidean norm* or length of the vector is

EUCLIDEAN NORM

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^t\mathbf{x}}. \tag{8}$$

we call a vector "normalized" if $\|\mathbf{x}\| = 1$. The angle between two d-dimensional vectors obeys

$$cos\ \theta = \frac{\mathbf{x}^t\mathbf{y}}{||\mathbf{x}||\ ||\mathbf{y}||}, \tag{9}$$

and thus the inner product is a measure of the colinearity of two vectors — a natural indication of their similarity. In particular, if $\mathbf{x}^t\mathbf{y} = 0$, then the vectors are orthogonal, and if $||\mathbf{x}^t\mathbf{y}|| = ||\mathbf{x}||\ ||\mathbf{y}||$, the vectors are colinear. From Eq. 9, we have immediately the Cauchy-Schwarz inequality, which states

$$\|\mathbf{x}^t\mathbf{y}\| \leq ||\mathbf{x}||\ ||\mathbf{y}||. \tag{10}$$

We say a set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ is *linearly independent* if no vector in the set can be written as a linear combination of any of the others. Informally, a set of d linearly independent vectors spans an d-dimensional vector space, i.e., any vector in that space can be written as a linear combination of such spanning vectors.

LINEAR INDEPENDENCE

A.2.3 Outer product

The outer product (sometimes called *matrix product* or *dyadic product*) of two vectors yields a matrix

MATRIX PRODUCT

$$\mathbf{M} = \mathbf{x}\mathbf{y}^t = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \begin{pmatrix} y_1 & y_2 & \dots & y_n \end{pmatrix} = \begin{pmatrix} x_1y_1 & x_1y_2 & \dots & x_1y_n \\ x_2y_1 & x_2y_2 & \dots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_dy_1 & x_dy_2 & \dots & x_dy_n \end{pmatrix}, \tag{11}$$

that is, the components of $\mathbf{M}$ are $m_{ij} = x_iy_j$. Of course, if the dimensions of $\mathbf{x}$ and $\mathbf{y}$ are not the same, then $\mathbf{M}$ is not square.

A.2.4 Derivatives of matrices

Suppose $f(\mathbf{x})$ is a scalar-valued function of d variables x_i, $i = 1, 2, ...d$, which we represent as the vector $\mathbf{x}$. Then the derivative or gradient of f with respect to this vector is computed component by component, i.e.,

$$\nabla f(\mathbf{x}) = \text{grad} f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_d} \end{pmatrix}. \tag{12}$$

If we have an n-dimensional vector-valued function $\mathbf{f}$ (note the use of boldface), of a d-dimensional vector $\mathbf{x}$, we calculate the derivatives and represent them as the *Jacobian matrix*

JACOBIAN MATRIX

$$\mathbf{J}(\mathbf{x}) = \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial f_1(\mathbf{x})}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial f_n(\mathbf{x})}{\partial x_d} \end{pmatrix}. \tag{13}$$

If this matrix is square, its determinant (Sect. A.2.5) is called simply the *Jacobian* or occassionally the *Jacobian determinant*.

If the entries of $\mathbf{M}$ depend upon a scalar parameter θ, we can take the derivative of $\mathbf{M}$ component by component, to get another matrix, as

$$\frac{\partial \mathbf{M}}{\partial \theta} = \begin{pmatrix} \frac{\partial m_{11}}{\partial \theta} & \frac{\partial m_{12}}{\partial \theta} & \dots & \frac{\partial m_{1d}}{\partial \theta} \\ \frac{\partial m_{21}}{\partial \theta} & \frac{\partial m_{22}}{\partial \theta} & \dots & \frac{\partial m_{2d}}{\partial \theta} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial m_{n1}}{\partial \theta} & \frac{\partial m_{n2}}{\partial \theta} & \dots & \frac{\partial m_{nd}}{\partial \theta} \end{pmatrix}. \tag{14}$$

In Sect. A.2.6 we shall discuss matrix inversion, but for convenience we give here the derivative of the inverse of a matrix, $\mathbf{M}^{-1}$:

$$\frac{\partial}{\partial \theta}\mathbf{M}^{-1} = -\mathbf{M}^{-1}\frac{\partial \mathbf{M}}{\partial \theta}\mathbf{M}^{-1}. \tag{15}$$

Consider a matrix $\mathbf{M}$ that is independent of $\mathbf{x}$. The following vector derivative identities can be verified by writing out the components:

$$\frac{\partial}{\partial \mathbf{x}}[\mathbf{M}\mathbf{x}] = \mathbf{M} \tag{16}$$

$$\frac{\partial}{\partial \mathbf{x}}[\mathbf{y}^t\mathbf{x}] = \frac{\partial}{\partial \mathbf{x}}[\mathbf{x}^t\mathbf{y}] = \mathbf{y} \tag{17}$$

$$\frac{\partial}{\partial \mathbf{x}}[\mathbf{x}^t\mathbf{M}\mathbf{x}] = [\mathbf{M}+\mathbf{M}^t]\mathbf{x}. \tag{18}$$

In the case where $\mathbf{M}$ is symmetric (as for instance a covariance matrix, cf. Sect. A.4.10), then Eq. 18 simplifies to

$$\frac{\partial}{\partial \mathbf{x}}[\mathbf{x}^t\mathbf{M}\mathbf{x}] = 2\mathbf{M}\mathbf{x}. \tag{19}$$

We first recall the use of second derivatives of a scalar function of a scalar x in writing a Taylor series (or Taylor expansion) about a point:

$$f(x) = f(x_0) + \left.\frac{df(x)}{dx}\right|_{x=x_0}(x-x_0) + \frac{1}{2!}\left.\frac{d^2f(x)}{dx^2}\right|_{x=x_0}(x-x_0)^2 + O((x-x_0)^3). \tag{20}$$

Analogously, if our scalar-valued f is a instead function of a vector $\mathbf{x}$, we can expand $f(\mathbf{x})$ in a Taylor series around a point $\mathbf{x}_0$:

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \underbrace{\left[\frac{\partial f}{\partial \mathbf{x}}\right]}_{\mathbf{J}}{}^{t}_{\mathbf{x}=\mathbf{x}_0}(\mathbf{x}-\mathbf{x}_0) + \frac{1}{2!}(\mathbf{x}-\mathbf{x}_0)^t\underbrace{\left[\frac{\partial^2 f}{\partial \mathbf{x}^2}\right]}_{\mathbf{H}}{}^{t}_{\mathbf{x}=\mathbf{x}_0}(\mathbf{x}-\mathbf{x}_0) + O(||\mathbf{x}-\mathbf{x}_0||^3), \tag{21}$$

where $\mathbf{H}$ is the *Hessian* matrix, the matrix of second-order derivatives of $f(\cdot)$, here evaluated at $\mathbf{x}_0$. (We shall return in Sect. A.8 to consider the $O(\cdot)$ notation and the order of a function used in Eq. 21 and below.) HESSIAN MATRIX

A.2.5 Determinant and trace

The determinant of a $d \times d$ (square) matrix is a scalar, denoted $|\mathbf{M}|$, and reveals properties of the matrix. For instance, if we consider the columns of $\mathbf{M}$ as vectors, if these vectors are not linearly independent, then the determinant vanishes. In pattern recognition, we have particular interest in the covariance matrix $\mathbf{\Sigma}$, which contains the second moments of a sample of data. In this case the absolute value of the determinant of a covariance matrix is a measure of the d-dimensional hypervolume of the data that yielded $\mathbf{\Sigma}$. (It can be shown that the determinant is equal to the product of the eigenvalues of a matrix, as mentioned in Sec. A.2.7.) If the data lies in a subspace of the full d-dimensional space, then the columns of $\mathbf{\Sigma}$ are not linearly independent, and the determinant vanishes. Further, the determinant must be non-zero for the inverse of a matrix to exist (Sec. A.2.6).

The calculation of the determinant is simple in low dimensions, and a bit more involved in high dimensions. If $\mathbf{M}$ is itself a scalar (i.e., a 1×1 matrix M), then $|M| = M$. If $\mathbf{M}$ is 2×2, then $|\mathbf{M}| = m_{11}m_{22} - m_{21}m_{12}$. The determinant of a general square matrix can be computed by a method called *expansion by minors*, and this EXPANSION BY MINORS leads to a recursive definition. If $\mathbf{M}$ is our $d \times d$ matrix, we define $\mathbf{M}_{i|j}$ to be the $(d-1) \times (d-1)$ matrix obtained by deleting the i^{th} row and the j^{th} column of $\mathbf{M}$:

$$\begin{array}{cc} & \begin{matrix} \phantom{m_{11}} & \phantom{m_{12}} & & j & & & \phantom{m_{1d}} \end{matrix} \\ \begin{matrix} \\ \\ \\ \\ i \\ \\ \\ \end{matrix} & \begin{pmatrix} m_{11} & m_{12} & \cdots & \otimes & \cdots & \cdots & m_{1d} \\ m_{21} & m_{22} & \cdots & \otimes & \cdots & \cdots & m_{2d} \\ \vdots & \vdots & \ddots & \otimes & \cdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \otimes & \cdots & \cdots & \vdots \\ \otimes & \otimes & \otimes & \otimes & \otimes & \otimes & \otimes \\ \vdots & \vdots & \cdots & \otimes & \cdots & \ddots & \vdots \\ m_{d1} & m_{d2} & \cdots & \otimes & \cdots & \cdots & m_{dd} \end{pmatrix} \end{array} = \mathbf{M}_{i|j}. \tag{22}$$

Given the determinants $|\mathbf{M}_{x|1}|$, we can now compute the determinant of $\mathbf{M}$ the expansion by minors on the first column giving

$$|\mathbf{M}| = m_{11}|\mathbf{M}_{1|1}| - m_{21}|\mathbf{M}_{2|1}| + m_{31}|\mathbf{M}_{3|1}| - \cdots \pm m_{d1}|\mathbf{M}_{d|1}|, \tag{23}$$

where the signs alternate. This process can be applied recursively to the successive (smaller) matrixes in Eq. 23.

Only for a 3×3 matrix, this determinant calculation can be represented by "sweeping" the matrix, i.e., taking the sum of the products of matrix terms along a diagonal, where products from upper-left to lower-right are added with a positive sign, and those from the lower-left to upper-right with a minus sign. That is,

$$\begin{aligned} |\mathbf{M}| &= \begin{vmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{vmatrix} \\ &= m_{11}m_{22}m_{33} + m_{13}m_{21}m_{32} + m_{12}m_{23}m_{31} \\ &\quad -m_{13}m_{22}m_{31} - m_{11}m_{23}m_{32} - m_{12}m_{21}m_{33}. \end{aligned} \tag{24}$$

Again, this "sweeping" mnemonic does not work for matrices larger than 3×3.

For any matrix we have $|\mathbf{M}| = |\mathbf{M}^t|$. Furthermore, for two square matrices of equal size $\mathbf{M}$ and $\mathbf{N}$, we have $|\mathbf{MN}| = |\mathbf{M}|\ |\mathbf{N}|$.

The *trace* of a $d \times d$ (square) matrix, denoted $\mathrm{tr}[\mathbf{M}]$, is the sum of its diagonal elements:

$$\mathrm{tr}[\mathbf{M}] = \sum_{i=1}^{d} m_{ii}. \tag{25}$$

Both the determinant and trace of a matrix are invariant with respect to rotations of the coordinate system.

A.2.6 Matrix inversion

So long as its determinant does not vanish, the inverse of a $d \times d$ matrix $\mathbf{M}$, denoted $\mathbf{M}^{-1}$, is the $d \times d$ matrix such that

$$\mathbf{M}\mathbf{M}^{-1} = \mathbf{I}. \tag{26}$$

COFACTOR We call the scalar $C_{ij} = (-1)^{i+j}|M_{i|j}|$ the i, j *cofactor* or equivalently the cofactor of the i, j entry of $\mathbf{M}$. As defined in Eq. 22, $\mathbf{M}_{i|j}$ is the $(d-1)\times(d-1)$ matrix formed by deleting the ith row and jth column of $\mathbf{M}$. The *adjoint* of $\mathbf{M}$, written $\mathrm{Adj}[\mathbf{M}]$, is the matrix whose i, j entry is the j, i cofactor of $\mathbf{M}$. Given these definitions, we can write the inverse of a matrix as

ADJOINT

$$\mathbf{M}^{-1} = \frac{\mathrm{Adj}[\mathbf{M}]}{|\mathbf{M}|}. \tag{27}$$

If $\mathbf{M}$ is not square (or if $\mathbf{M}^{-1}$ in Eq. 27 does not exist because the columns of $\mathbf{M}$ are not linearly independent) we typically use instead the *pseudoinverse* $\mathbf{M}^\dagger$, defined as

PSEUDO-INVERSE

$$\mathbf{M}^\dagger = [\mathbf{M}^t\mathbf{M}]^{-1}\mathbf{M}^t. \tag{28}$$

The pseudoinverse is useful because it insures $\mathbf{M}^\dagger\mathbf{M} = \mathbf{I}$.

A.2.7 Eigenvectors and eigenvalues

Given a $d \times d$ matrix $\mathbf{M}$ a very important class of linear equations is of the form

$$\mathbf{M}\mathbf{x} = \lambda\mathbf{x} \tag{29}$$

for scalar λ, which can be rewritten

$$(\mathbf{M} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}, \tag{30}$$

where $\mathbf{I}$ the identity matrix, and $\mathbf{0}$ the zero vector. The solution vector $\mathbf{x} = \mathbf{e}_i$ and corresponding scalar $\lambda = \lambda_i$ to Eq. 29 are called the *eigenvector* and associated *eigenvalue*. There are d (possibly non-distinct) solution vectors $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_d\}$ each with an associated eigenvalue $\{\lambda_1, \lambda_2, \ldots, \lambda_d\}$. Under multiplication by $\mathbf{M}$ the eigenvectors are changed only in magnitude — not direction:

$$\mathbf{M}\mathbf{e}_j = \lambda_j\mathbf{e}_j. \tag{31}$$

If $\mathbf{M}$ is diagonal, then the eigenvectors are parallel to the coordinate axes.

One method of finding the eigenvectors and eigenvalues is to solve the *characteristic equation* (or *secular equation*),

CHARACTER-ISTIC EQUATION

$$|\mathbf{M} - \lambda\mathbf{I}| = \lambda^d + a_1\lambda^{d-1} + \ldots + a_{d-1}\lambda + a_d = 0, \tag{32}$$

SECULAR EQUATION

for each of its d (possibly non-distinct) roots λ_j. For each such root, we then solve a set of linear equations to find its associated eigenvector $\mathbf{e}_j$.

Finally, it can be shown that the determinant of a matrix is just the product of its eigenvalues:

$$|\mathbf{M}| = \prod_{i=1}^{d} \lambda_i. \tag{33}$$

A.3 Lagrange optimization

Suppose we seek the position $\mathbf{x}_0$ of an extremum of a scalar-valued function $f(\mathbf{x})$, subject to some constraint. If a constraint can be expressed in the form $g(\mathbf{x}) = 0$, then we can find the extremum of $f(\mathbf{x})$ as follows. First we form the Lagrangian function

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \underbrace{\lambda g(\mathbf{x})}_{=0}, \tag{34}$$

UNDETERMINED MULTIPLIER

where λ is a scalar called the Lagrange *undetermined multiplier*. We convert this constrained optimization problem into an unconstrained problem by taking the derivative,

$$\frac{\partial L(\mathbf{x}, \lambda)}{\partial \mathbf{x}} = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} + \lambda \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} = 0, \tag{35}$$

and using standard methods from calculus to solve the resulting equations for λ and the extremizing value of $\mathbf{x}$. (Note that the last term on the left hand side does not vanish, in general.) The solution gives the $\mathbf{x}$ position of the extremum, and it is a simple matter of substitution to find the extreme value of $f(\cdot)$ under the constraints.

A.4 Probability Theory

A.4.1 Discrete random variables

Let x be a discrete random variable that can assume any of the finite number m of different values in the set $\mathcal{X} = \{v_1, v_2, \ldots, v_m\}$. We denote by p_i the probability that x assumes the value v_i:

$$p_i = \Pr\{x = v_i\}, \quad i = 1, \ldots, m. \tag{36}$$

Then the probabilities p_i must satisfy the following two conditions:

$$\begin{aligned} p_i &\geq 0 \quad \text{and} \\ \sum_{i=1}^{m} p_i &= 1. \end{aligned} \tag{37}$$

PROBABILITY MASS FUNCTION

Sometimes it is more convenient to express the set of probabilities $\{p_1, p_2, \ldots, p_m\}$ in terms of the *probability mass function* $P(x)$, which must satisfy the following two conditions:

$$\begin{aligned} P(x) &\geq 0 \quad \text{and} \\ \sum_{x \in \mathcal{X}} P(x) &= 1\text{and} \qquad (38) \\ \sum_{\mathbf{x} \notin \mathcal{X}} P(x) &= 0. \qquad (39) \end{aligned}$$

A.4.2 Expected values

MEAN

The *expected value*, *mean* or *average* of the random variable x is defined by

$$\mathcal{E}[x] = \mu = \sum_{x \in \mathcal{X}} xP(x) = \sum_{i=1}^{m} v_i p_i. \tag{40}$$

If one thinks of the probability mass function as defining a set of point masses, with p_i being the mass concentrated at $x = v_i$, then the expected value μ is just the center of mass. Alternatively, we can interpret μ as the arithmetic average of the values in a large random sample. More generally, if $f(x)$ is any function of x, the expected value of f is defined by

$$\mathcal{E}[f(x)] = \sum_{x \in \mathcal{X}} f(x)P(x). \tag{41}$$

Note that the process of forming an expected value is *linear*, in that if α_1 and α_2 are arbitrary constants,

$$\mathcal{E}[\alpha_1 f_1(x) + \alpha_2 f_2(x)] = \alpha_1 \mathcal{E}[f_1(x)] + \alpha_2 \mathcal{E}[f_2(x)]. \tag{42}$$

It is sometimes convenient to think of $\mathcal{E}$ as an operator — the (linear) *expectation operator*. Two important special-case expectations are the *second moment* and the *variance*:

EXPECTATION OPERATOR

SECOND MOMENT

VARIANCE

$$\mathcal{E}[x^2] = \sum_{x \in \mathcal{X}} x^2 P(x) \tag{43}$$

$$\mathrm{Var}[x] \equiv \sigma^2 = \mathcal{E}[(x-\mu)^2] = \sum_{x \in \mathcal{X}} (x-\mu)^2 P(x), \tag{44}$$

where σ is the *standard deviation* of x. The variance can be viewed as the moment of inertia of the probability mass function. The variance is never negative, and is zero if and only if all of the probability mass is concentrated at one point.

STANDARD DEVIATION

The standard deviation is a simple but valuable measure of how far values of x are likely to depart from the mean. Its very name suggests that it is the standard or typical amount one should expect a randomly drawn value for x to deviate or differ from μ. *Chebyshev's inequality* (or Bienaymé-Chebyshev inequality) provides a mathematical relation between the standard deviation and $|x - \mu|$:

CHEBYSHEV'S INEQUALITY

$$\Pr\{|x - \mu| > n\sigma\} \leq \frac{1}{n^2}. \tag{45}$$

This inequality is not a tight bound (and it is useless for $n < 1$); a more practical rule of thumb, which strictly speaking is true only for the normal distribution, is that 68% of the values will lie within one, 95% within two, and 99.7% within three standard deviations of the mean (Fig. A.1). Nevertheless, Chebyshev's inequality shows the strong link between the standard deviation and the spread of a distribution. In addition, it suggests that $|x-\mu|/\sigma$ is a meaningful normalized measure of the distance from x to the mean (cf. Sect. A.4.12).

By expanding the quadratic in Eq. 44, it is easy to prove the useful formula

$$\mathrm{Var}[x] = \mathcal{E}[x^2] - (\mathcal{E}[x])^2. \tag{46}$$

Note that, unlike the mean, the variance is *not* linear. In particular, if $y = \alpha x$, where α is a constant, then $\mathrm{Var}[y] = \alpha^2 \mathrm{Var}[x]$. Moreover, the variance of the sum of two random variables is usually *not* the sum of their variances. However, as we shall see below, variances do add when the variables involved are statistically independent.

In the simple but important special case in which x is binary valued (say, $v_1 = 0$ and $v_2 = 1$), we can obtain simple formulas for μ and σ. If we let $p = \Pr\{x = 1\}$, then it is easy to show that

$$\begin{aligned} \mu &= p \quad \text{and} \\ \sigma &= \sqrt{p(1-p)}. \end{aligned} \tag{47}$$

A.4.3 Pairs of discrete random variables

Let x and y be random variables which can take on values in $\mathcal{X} = \{v_1, v_2, \ldots, v_m\}$, and $\mathcal{Y} = \{w_1, w_2, \ldots, w_n\}$, respectively. We can think of (x, y) as a vector or a point in the *product space* of x and y. For each possible pair of values (v_i, w_j) we have a *joint probability* $p_{ij} = \Pr\{x = v_i, y = w_j\}$. These mn joint probabilities p_{ij} are non-negative and sum to 1. Alternatively, we can define a *joint probability mass function* $P(x, y)$ for which

PRODUCT SPACE

$$\begin{aligned} P(x,y) &\geq 0 \quad \text{and} \\ \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P(x,y) &= 1. \end{aligned} \tag{48}$$

The joint probability mass function is a complete characterization of the pair of random variables (x, y); that is, everything we can compute about x and y, individually or together, can be computed from $P(x, y)$. In particular, we can obtain the separate *marginal distributions* for x and y by summing over the unwanted variable:

MARGINAL DISTRIBUTION

$$\begin{aligned} P_x(x) &= \sum_{y\in\mathcal{Y}} P(x,y) \\ P_y(y) &= \sum_{x\in\mathcal{X}} P(x,y). \end{aligned} \tag{49}$$

We will occassionally use subscripts, as in Eq. 49, to emphasize the fact that $P_x(x)$ has a different functional form than $P_y(y)$. It is common to omit them and write simply $P(x)$ and $P(y)$ whenever the context makes it clear that these are in fact two different functions — rather than the same function merely evaluated with different variables.

A.4.4 Statistical independence

Variables x and y are said to be *statistically independent* if and only if

$$P(x, y) = P_x(x)P_y(y). \tag{50}$$

We can understand such independence as follows. Suppose that $p_i = \Pr\{x = v_i\}$ is the fraction of the time that $x = v_i$, and $q_j = \Pr\{y = w_j\}$ is the fraction of the time that $y = w_j$. Consider those situations where $x = v_i$. If it is still true that the fraction of those situations in which $y = w_j$ is the same value q_j, it follows that knowing the value of x did not give us any additional knowledge about the possible values of y; in that sense y is independent of x. Finally, if x and y are statistically independent, it is clear that the fraction of the time that the specific pair of values (v_i, w_j) occurs must be the product of the fractions $p_i q_j = P(v_i)P(w_j)$.

A.4.5 Expected values of functions of two variables

In the natural extension of Sect. A.4.2, we define the expected value of a function $f(x, y)$ of two random variables x and y by

$$\mathcal{E}[f(x, y)] = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} f(x, y)P(x, y), \tag{51}$$

and as before the expectation operator $\mathcal{E}$ is linear:

$$\mathcal{E}[\alpha_1 f_1(x, y) + \alpha_2 f_2(x, y)] = \alpha_1\mathcal{E}[f_1(x, y)] + \alpha_2\mathcal{E}[f_2(x, y)]. \tag{52}$$

The means (first moments) and variances (second moments) are:

$$\begin{aligned}
\mu_x = \mathcal{E}[x] &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} xP(x, y) \\
\mu_y = \mathcal{E}[y] &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} yP(x, y) \\
\sigma_x^2 = V[x] = \mathcal{E}[(x - \mu_x)^2] &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} (x - \mu_x)^2 P(x, y) \\
\sigma_y^2 = V[y] = \mathcal{E}[(y - \mu_y)^2] &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} (y - \mu_y)^2 P(x, y). \qquad (53)
\end{aligned}$$

An important new "cross-moment" can now be defined, the *covariance* of x and y:

COVARIANCE

$$\sigma_{xy} = \mathcal{E}[(x - \mu_x)(y - \mu_y)] = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} (x - \mu_x)(y - \mu_y)P(x, y). \tag{54}$$

We can summarize Eqs. 53 & 54 using vector notation as:

$$\begin{aligned}
\boldsymbol{\mu} &= \mathcal{E}[\mathbf{x}] = \sum_{\mathbf{x}\in\{\mathcal{X}\mathcal{Y}\}} \mathbf{x}P(\mathbf{x}) \qquad (55) \\
\boldsymbol{\Sigma} &= \mathcal{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t], \qquad (56)
\end{aligned}$$

where $\{\mathcal{X}\mathcal{Y}\}$ respresents the space of all possible values for all components of $\mathbf{x}$ and $\mathbf{\Sigma}$ is the covariance matrix (cf., Sect. A.4.9).

The covariance is one measure of the degree of statistical dependence between x and y. If x and y are statistically independent, then $\sigma_{xy} = 0$. If $\sigma_{xy} = 0$, the variables x and y are said to be *uncorrelated*. It does *not* follow that uncorrelated variables must be statistically independent — covariance is just one measure of dependence. However, it is a fact that uncorrelated variables are statistically independent if they have a multivariate normal distribution, and in practice statisticians often treat uncorrelated variables as if they were statistically independent. If α is a constant and $y = \alpha x$, which is a case of strong statistical dependence, it is also easy to show that $\sigma_{xy} = \alpha\sigma_x^2$. Thus, the covariance is positive if x and y both increase or decrease together, and is negative if y decreases when x increases.

UNCORRELATED

CAUCHY-SCHWARZ INEQUALITY

There is an important *Cauchy-Schwarz inequality* for the variances σ_x and σ_y and the covariance σ_{xy}. It can be derived by observing that the variance of a random variable is never negative, and thus the variance of $\lambda x + y$ must be non-negative no matter what the value of the scalar λ. This leads to the famous inequality

$$\sigma_{xy}^2 \leq \sigma_x^2 \sigma_y^2, \tag{57}$$

which is analogous to the vector inequality $(\mathbf{x}^t\mathbf{y})^2 \leq \|\mathbf{x}\|^2 \, \|\mathbf{y}\|^2$ given in Eq. 8.

CORRELATION COEFFICIENT

The *correlation coefficient*, defined as

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \tag{58}$$

is a normalized covariance, and must always be between -1 and $+1$. If $\rho = +1$, then x and y are maximally positively correlated, while if $\rho = -1$, they are maximally negatively correlated. If $\rho = 0$, the variables are uncorrelated. It is common for statisticians to consider variables to be uncorrelated for practical purposes if the magnitude of their correlation coefficient is below some threshold, such as 0.05, although the threshold that makes sense does depend on the actual situation.

If x and y are statistically independent, then for any two functions f and g

$$\mathcal{E}[f(x)g(y)] = \mathcal{E}[f(x)]\mathcal{E}[g(y)], \tag{59}$$

a result which follows from the definition of statistical independence and expectation. Note that if $f(x) = x - \mu_x$ and $g(y) = y - \mu_y$, this theorem again shows that $\sigma_{xy} = \mathcal{E}[(x - \mu_x)(y - \mu_y)]$ is zero if x and y are statistically independent.

A.4.6 Conditional probability

When two variables are statistically dependent, knowing the value of one of them lets us get a better estimate of the value of the other one. This is expressed by the following definition of the *conditional probability* of x given y:

$$\Pr\{x = v_i | y = w_j\} = \frac{\Pr\{x = v_i, y = w_j\}}{\Pr\{y = w_j\}}, \tag{60}$$

or, in terms of mass functions,

$$P(x|y) = \frac{P(x, y)}{P(y)}. \tag{61}$$

Note that if x and y are statistically independent, this gives $P(x|y) = P(x)$. That is, when x and y are independent, knowing the value of y gives you no information about x that you didn't already know from its marginal distribution $P(x)$.

Consider a simple illustration of a two-variable binary case where both x and y are either 0 or 1. Suppose that a large number n of pairs of xy-values are randomly produced. Let n_{ij} be the number of pairs in which we find $x = i$ and $y = j$, i.e., we see the $(0, 0)$ pair n_{00} times, the $(0, 1)$ pair n_{01} times, and so on, where $n_{00} + n_{01} + n_{10} + n_{11} = n$. Suppose we pull out those pairs where $y = 1$, i.e., the $(0, 1)$ pairs and the $(1, 1)$ pairs. Clearly, the fraction of those cases in which x is also 1 is

$$\frac{n_{11}}{n_{01} + n_{11}} = \frac{n_{11}/n}{(n_{01} + n_{11})/n}. \tag{62}$$

Intuitively, this is what we would like to get for $P(x|y)$ when $y = 1$ and n is large. And, indeed, this is what we do get, because n_{11}/n is approximately $P(x, y)$ and $\frac{n_{11}/n}{(n_{01}+n_{11})/n}$ is approximately $P(y)$ for large n.

A.4.7 The Law of Total Probability and Bayes' rule

The *Law of Total Probability* states that if an event A can occur in m different ways $A_1, A_2, \ldots, A_m$, and if these m subevents are *mutually exclusive* — that is, cannot occur at the same time — then the probability of A occurring is the sum of the probabilities of the subevents A_i. In particular, the random variable y can assume the value y in m different ways — with $x = v_1$, with $x = v_2$, ..., and $x = v_m$. Because these possibilities are mutually exclusive, it follows from the Law of Total Probability that $P(y)$ is the sum of the joint probability $P(x, y)$ over all possible values for x. Formally we have

$$P(y) = \sum_{x \in \mathcal{X}} P(x, y). \tag{63}$$

But from the definition of the conditional probability $P(y|x)$ we have

$$P(x, y) = P(y|x)P(x), \tag{64}$$

and after rewriting Eq. 64 with x and y exchanged and a trivial math, we obtain

$$P(x|y) = \frac{P(y|x)P(x)}{\sum\limits_{x \in \mathcal{X}} P(y|x)P(x)}, \tag{65}$$

or in words,

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}},$$

where these terms are discussed more fully in Chapt. **??**.

Equation 65 is called *Bayes' rule*. Note that the denominator, which is just $P(y)$, is obtained by summing the numerator over all x values. By writing the denominator in this form we emphasize the fact that everything on the right-hand side of the equation is conditioned on x. If we think of x as the important variable, then we can say that the shape of the distribution $P(x|y)$ depends only on the numerator $P(y|x)P(x)$; the denominator is just a normalizing factor, sometimes called the *evidence*, needed to insure that the $P(x|y)$ sum to one. EVIDENCE

The standard interpretation of Bayes' rule is that it "inverts" statistical connections, turning $P(y|x)$ into $P(x|y)$. Suppose that we think of x as a "cause" and y as an "effect" of that cause. That is, we assume that if the cause x is present, it is easy to determine the probability of the effect y being observed; the conditional probability function $P(y|x)$ — the *likelihood* — specifies this probability explicitly. If we observe the effect y, it might not be so easy to determine the cause x, because there might be several different causes, each of which could produce the same observed effect. However, Bayes' rule makes it easy to determine $P(x|y)$, provided that we know both $P(y|x)$ and the so-called *prior probability* $P(x)$, the probability of x before we make any observations about y. Said slightly differently, Bayes' rule shows how the probability distribution for x changes from the *prior distribution* $P(x)$ before anything is observed about y to the *posterior* $P(x|y)$ once we have observed the value of y.

LIKELIHOOD

PRIOR

POSTERIOR

A.4.8 Vector random variables

To extend these results from two variables x and y to d variables $x_1, x_2, \ldots, x_d$, it is convenient to employ vector notation. As given by Eq. 48, the joint probability mass function $P(\mathbf{x})$ satisfies $P(\mathbf{x}) \geq 0$ and $\sum P(\mathbf{x}) = 1$, where the sum extends over all possible values for the vector $\mathbf{x}$. Note that $P(\mathbf{x})$ is a function of d variables, and can be a very complicated, multi-dimensional function. However, if the random variables x_i are statistically independent, it reduces to the product

$$\begin{aligned} P(\mathbf{x}) &= P_{x_1}(x_1)P_{x_2}(x_2)\cdots P_{x_d}(x_d) \\ &= \prod_{i=1}^{d} P_{x_i}(x_i). \end{aligned} \tag{66}$$

where we have used the subscripts just to emphasize the fact that the marginal distributions will generally have a different form. Here the separate marginal distributions $P_{x_i}(x_i)$ can be obtained by summing the joint distribution over the other variables. In addition to these univariate marginals, other marginal distributions can be obtained by this use of the Law of Total Probability. For example, suppose that we have $P(x_1, x_2, x_3, x_4, x_5)$ and we want $P(x_1, x_4)$, we merely calculate

$$P(x_1, x_4) = \sum_{x_2} \sum_{x_3} \sum_{x_5} P(x_1, x_2, x_3, x_4, x_5). \tag{67}$$

One can define many different conditional distributions, such as $P(x_1, x_2|x_3)$ or $P(x_2|x_1, x_4, x_5)$. For example,

$$P(x_1, x_2|x_3) = \frac{P(x_1, x_2, x_3)}{P(x_3)}, \tag{68}$$

where all of the joint distributions can be obtained from $P(\mathbf{x})$ by summing out the unwanted variables. If instead of scalars we have vector variables, then these conditional distributions can also be written as

$$P(\mathbf{x}_1|\mathbf{x}_2) = \frac{P(\mathbf{x}_1, \mathbf{x}_2)}{P(\mathbf{x}_2)}, \tag{69}$$

and likewise, in vector form, Bayes' rule becomes

$$P(\mathbf{x}_1|\mathbf{x}_2) = \frac{P(\mathbf{x}_2|\mathbf{x}_1)P(\mathbf{x}_1)}{\sum_{\mathbf{x}_1} P(\mathbf{x}_2|\mathbf{x}_1)P(\mathbf{x}_1)}. \tag{70}$$

A.4.9 Expectations, mean vectors and covariance matrices

The expected value of a vector is defined to be the vector whose components are the expected values of the original components. Thus, if $\mathbf{f}(\mathbf{x})$ is an n-dimensional, vector-valued function of the d-dimensional random vector $\mathbf{x}$,

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_n(\mathbf{x}) \end{bmatrix}, \tag{71}$$

then the expected value of $\mathbf{f}$ is defined by

$$\mathcal{E}[\mathbf{f}] = \begin{bmatrix} \mathcal{E}[f_1(\mathbf{x})] \\ \mathcal{E}[f_2(\mathbf{x})] \\ \vdots \\ \mathcal{E}[f_n(\mathbf{x})] \end{bmatrix} = \sum_{\mathbf{x}} \mathbf{f}(\mathbf{x})P(\mathbf{x}). \tag{72}$$

In particular, the d-dimensional *mean vector* $\boldsymbol{\mu}$ is defined by MEAN VECTOR

$$\boldsymbol{\mu} = \mathcal{E}[\mathbf{x}] = \begin{bmatrix} \mathcal{E}[x_1] \\ \mathcal{E}[x_2] \\ \vdots \\ \mathcal{E}[x_d] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{bmatrix} = \sum_{\mathbf{x}} \mathbf{x}P(\mathbf{x}). \tag{73}$$

Similarly, the *covariance matrix* $\boldsymbol{\Sigma}$ is defined as the (square) matrix whose ijth element σ_{ij} is the covariance of x_i and x_j: COVARIANCE MATRIX

$$\sigma_{ij} = \sigma_{ji} = \mathcal{E}[(x_i - \mu_i)(x_j - \mu_j)] \qquad i, j = 1 \ldots d, \tag{74}$$

as we saw in the two-variable case of Eq. 54. Therefore, in expanded form we have

$$\begin{aligned}
\boldsymbol{\Sigma} &= \begin{bmatrix} \mathcal{E}[(x_1-\mu_1)(x_1-\mu_1)] & \mathcal{E}[(x_1-\mu_1)(x_2-\mu_2)] & \ldots & \mathcal{E}[(x_1-\mu_1)(x_d-\mu_d)] \\ \mathcal{E}[(x_2-\mu_2)(x_1-\mu_1)] & \mathcal{E}[(x_2-\mu_2)(x_2-\mu_2)] & \ldots & \mathcal{E}[(x_2-\mu_2)(x_d-\mu_d)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{E}[(x_d-\mu_d)(x_1-\mu_1)] & \mathcal{E}[(x_d-\mu_d)(x_2-\mu_2)] & \ldots & \mathcal{E}[(x_d-\mu_d)(x_d-\mu_d)] \end{bmatrix} \\
&= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \ldots & \sigma_{dd} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \ldots & \sigma_d^2 \end{bmatrix}.
\end{aligned} \tag{75}$$

We can use the vector product $(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t$, to write the covariance matrix as

$$\boldsymbol{\Sigma} = \mathcal{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t]. \tag{76}$$

Thus, $\mathbf{\Sigma}$ is symmetric, and its diagonal elements are just the variances of the individual elements of $\mathbf{x}$, which can never be negative; the off-diagonal elements are the covariances, which can be positive or negative. If the variables are statistically independent, the covariances are zero, and the covariance matrix is diagonal. The analog to the Cauchy-Schwarz inequality comes from recognizing that if $\mathbf{w}$ is any d-dimensional vector, then the variance of $\mathbf{w}^t\mathbf{x}$ can never be negative. This leads to the requirement that the quadratic form $\mathbf{w}^t\mathbf{\Sigma}\mathbf{w}$ never be negative. Matrices for which this is true are said to be *positive semi-definite*; thus, the covariance matrix $\mathbf{\Sigma}$ must be positive semi-definite. It can be shown that this is equivalent to the requirement that none of the eigenvalues of $\mathbf{\Sigma}$ can be negative.

A.4.10 Continuous random variables

When the random variable x can take values in the continuum, it no longer makes sense to talk about the probability that x has a particular value, such as 2.5136, because the probability of any particular exact value will almost always be zero. Rather, we talk about the probability that x falls in some interval (a, b); instead of having a probability mass function $P(x)$ we have a *probability mass density function* $p(x)$. The mass density has the property that

MASS DENSITY

$$\Pr\{x \in (a,b)\} = \int_a^b p(x)\,dx. \tag{77}$$

The name *density* comes by analogy with material density. If we consider a small interval $(a, a+\Delta x)$ over which $p(x)$ is essentially constant, having value $p(a)$, we see that $p(a) = \Pr\{x \in (a, a+\Delta x)\}/\Delta x$. That is, the probability mass density at $x = a$ is the probability mass $\Pr\{x \in (a, a+\Delta x)\}$ per unit distance. It follows that the probability density function must satisfy

$$\begin{aligned} &p(x) \geq 0 \quad \text{and} \\ &\int_{-\infty}^{\infty} p(x)\,dx = 1. \end{aligned} \tag{78}$$

In general, most of the definitions and formulas for discrete random variables carry over to continuous random variables with sums replaced by integrals. In particular, the expected value, mean and variance for a continuous random variable are defined by

$$\begin{aligned} \mathcal{E}[f(x)] &= \int_{-\infty}^{\infty} f(x)p(x)\,dx \\ \mu = \mathcal{E}[x] &= \int_{-\infty}^{\infty} xp(x)\,dx \\ \mathrm{Var}[x] = \sigma^2 = \mathcal{E}[(x-\mu)^2] &= \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\;dx, \end{aligned} \tag{79}$$

and, as in Eq. 46, the variance obeys $\sigma^2 = \mathcal{E}[x^2] - (\mathcal{E}[x])^2$.

The multivariate situation is similarly handled with continuous random vectors $\mathbf{x}$. The probability density function $p(\mathbf{x})$ must satisfy

$$\begin{aligned} & p(\mathbf{x}) \geq 0 \quad \text{and} \\ & \int_{-\infty}^{\infty} p(\mathbf{x})\, d\mathbf{x} = 1, \end{aligned} \tag{80}$$

where the integral is understood to be a d-fold, multiple integral, and where $d\mathbf{x}$ is the element of d-dimensional volume $d\mathbf{x} = dx_1 dx_2 \cdots dx_d$. The corresponding moments for a general n-dimensional vector-valued function are

$$\mathcal{E}[\mathbf{f}(\mathbf{x})] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbf{f}(\mathbf{x}) p(\mathbf{x})\ dx_1 dx_2 \ldots dx_d = \int_{-\infty}^{\infty} \mathbf{f}(\mathbf{x}) p(\mathbf{x})\, d\mathbf{x} \tag{81}$$

and for the particular d-dimensional functions as above, we have

$$\begin{aligned} \boldsymbol{\mu} = \mathcal{E}[\mathbf{x}] &= \int_{-\infty}^{\infty} \mathbf{x} p(\mathbf{x})\, d\mathbf{x} \qquad (82) \\ \boldsymbol{\Sigma} = \mathcal{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t] &= \int_{-\infty}^{\infty} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t p(\mathbf{x})\ d\mathbf{x}. \end{aligned}$$

If the components of $\mathbf{x}$ are statistically independent, then the joint probability density function factors as

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i) \tag{83}$$

and the covariance matrix is diagonal.

Conditional probability density functions are defined just as conditional mass functions. Thus, for example, the density for x given y is given by

$$p(x|y) = \frac{p(x, y)}{p(y)} \tag{84}$$

and Bayes' rule for density functions is

$$p(x|y) = \frac{p(y|x)p(x)}{\int_{-\infty}^{\infty} p(y|x)p(x)\, dx}, \tag{85}$$

and likewise for the vector case.

Occassionally we will need to take the expectation with respect to a subset of the variables, and in that case we must show this as a subscript, for instance

$$\mathcal{E}_{x_1}[f(x_1, x_2)] = \int_{-\infty}^{\infty} f(x_1, x_2) p(x_1)\ dx_1. \tag{86}$$

A.4.11 Distributions of sums of independent random variables

It frequently happens that we know the densities for two independent random variables x and y, and we need to know the density of their sum $z = x + y$. It is easy to obtain the mean and the variance of this sum:

$$\begin{aligned}
\mu_z &= \mathcal{E}[z] = \mathcal{E}[x+y] = \mathcal{E}[x] + \mathcal{E}[y] = \mu_x + \mu_y, \\
\sigma_z^2 &= \mathcal{E}[(z-\mu_z)^2] = \mathcal{E}[(x+y-(\mu_x+\mu_y))^2] = \mathcal{E}[((x-\mu_x)+(y-\mu_y))^2] \\
&= \mathcal{E}[(x-\mu_x)^2] + 2\underbrace{\mathcal{E}[(x-\mu_x)(y-\mu_y)]}_{=0} + \mathcal{E}[(y-\mu_y)^2] \qquad (87) \\
&= \sigma_x^2 + \sigma_y^2,
\end{aligned}$$

where we have used the fact that the cross-term factors into $\mathcal{E}[x-\mu_x]\mathcal{E}[y-\mu_y]$ when x and y are independent; in this case the product is manifestly zero, since each of the component expectations vanishes. Thus, in words, the mean of the sum of two independent random variables is the sum of their means, and the variance of their sum is the sum of their variances. If the variables are random *yet not independent* — for instance $y = -x$, where x is randomly distributed — then the variance is not the sum of the component variances.

It is only slightly more difficult to work out the exact probability density function for $z = x+y$ from the separate density functions for x and y. The probability that z is between ζ and $\zeta + \Delta z$ can be found by integrating the joint density $p(x, y) = p(x)p(y)$ over the thin strip in the xy-plane between the lines $x + y = \zeta$ and $x + y = \zeta + \Delta z$. It follows that, for small Δz,

$$\Pr\{\zeta < z < \zeta + \Delta z\} = \left\{ \int_{-\infty}^{\infty} p(x)p(\zeta - x)\, dx \right\} \Delta z, \qquad (88)$$

CONVOLUTION and hence that the probability density function for the sum is the *convolution* of the probability density functions for the components:

$$p(z) = p(x) \star p(y) = \int_{-\infty}^{\infty} p(x)p(z-x)\, dx. \qquad (89)$$

As one would expect, these results generalize. It is not hard to show that:

- The mean of the sum of d independent random variables $x_1, x_2, \ldots, x_d$ is the sum of their means. (In fact the variables need not be independent for this to hold.)
- The variance of the sum is the sum of their variances.
- The probability density function for the sum is the convolution of the separate density functions:

$$p(z) = p(x_1) \star p(x_2) \star \ldots \star p(x_d). \qquad (90)$$

A.4.12 Univariate normal density

Central Limit Theorem

One of the most important results of probability theory is the *Central Limit Theorem*, which states that, under various conditions, the distribution for the sum of d independent random variables approaches a particular limiting form known as the *normal distribution*. As such, the *normal* or *Gaussian* probability density function is very important, both for theoretical and practical reasons. In one dimension, it is defined by

Gaussian

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-1/2((x-\mu)/\sigma)^2}. \tag{91}$$

The normal density is traditionally described as a "bell-shaped curve"; it is completely determined by the numerical values for two parameters, the mean μ and the variance σ^2. This is often emphasized by writing $p(x) \sim N(\mu, \sigma^2)$, which is read as "$x$ is distributed normally with mean μ and variance σ^2." The distribution is symmetrical about the mean, the peak occurring at $x = \mu$ and the width of the "bell" is proportional to the standard deviation σ. The parameters of a normal density in Eq. 91 satisfy the following equations:

$$\begin{aligned}
\mathcal{E}[1] &= \int_{-\infty}^{\infty} p(x)\,dx = 1 \\
\mathcal{E}[x] &= \int_{-\infty}^{\infty} x\,p(x)\,dx = \mu \\
\mathcal{E}[(x-\mu)^2] &= \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\,dx = \sigma^2.
\end{aligned} \tag{92}$$

Normally distributed data points tend to cluster about the mean. Numerically, the probabilities obey

$$\begin{aligned}
\Pr\{|x-\mu| \leq \sigma\} &\simeq 0.68 \\
\Pr\{|x-\mu| \leq 2\sigma\} &\simeq 0.95 \\
\Pr\{|x-\mu| \leq 3\sigma\} &\simeq 0.997,
\end{aligned} \tag{93}$$

as shown in Fig. A.1.

A natural measure of the distance from x to the mean μ is the distance $|x-\mu|$ measured in units of standard deviations:

$$r = \frac{|x-\mu|}{\sigma}, \tag{94}$$

the *Mahalanobis distance* from x to μ. (In the one-dimensional case, this is sometimes called the z-score.) Thus for instance the probability is 0.95 that the Mahalanobis distance from x to μ will be less than 2. If a random variable x is modified by (a) subtracting its mean and (b) dividing by its standard deviation, it is said to be *standardized*. Clearly, a standardized normal random variable $u = (x-\mu)/\sigma$ has zero mean and unit standard deviation, that is,

Mahalanobis distance

standardized

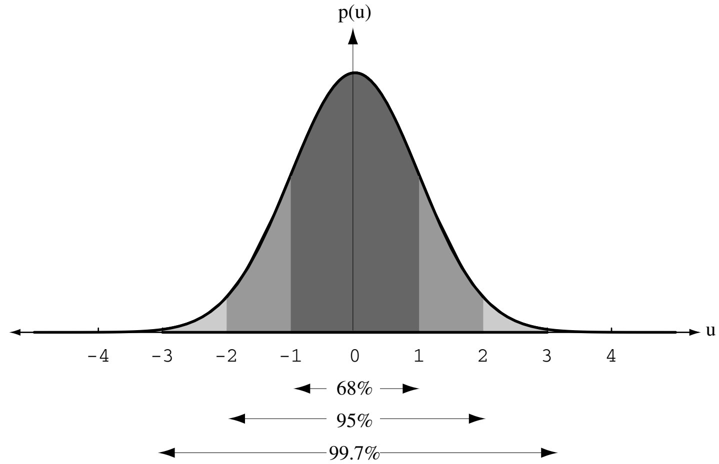

Figure A.1: A one-dimensional Gaussian distribution, $p(u) \sim N(0,1)$, has 68% of its probability mass in the range $|u| \leq 1$, 95% in the range $|u| \leq 2$, and 99.7% in the range $|u| \leq 3$.

$$p(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}, \tag{95}$$

which can be written as $p(u) \sim N(0,1)$. Table A.1 shows the probability that a value, chosen at random according to $p(u) \sim N(0,1)$, differs from the mean value by less than a criterion z.

Table A.1: The probability a sample drawn from a standardized Gaussian has absolute value less than a criterion, i.e., $Pr[|u| \leq z]$

z	$Pr[\|u\| \leq z]$	z	$Pr[\|u\| \leq z]$	z	$Pr[\|u\| \leq z]$
0.0	0.0	1.0	0.682	2.0	0.954
0.1	0.080	1.1	0.728	2.1	0.963
0.2	0.158	1.2	0.770	2.326	0.980
0.3	0.236	1.3	0.806	2.5	0.988
0.4	0.310	1.4	0.838	2.576	0.990
0.5	0.382	1.5	0.866	3.0	0.9974
0.6	0.452	1.6	0.890	3.090	0.9980
0.7	0.516	1.7	0.910	3.291	0.999
0.8	0.576	1.8	0.928	3.5	0.9996
0.9	0.632	1.9	0.942	4.0	0.99994

A.5 Gaussian derivatives and integrals

Because of the prevalence of Gaussian functions throughout statistical pattern recognition, we often have occassion to integrate and differentiate them. The first three derivatives of a one-dimensional (standardized) Gaussian are

$$\begin{aligned}
\frac{\partial}{\partial x}\left[\frac{1}{\sqrt{2\pi}\sigma}e^{-x^2/(2\sigma^2)}\right] &= \frac{-x}{\sqrt{2\pi}\sigma^3}e^{-x^2/(2\sigma^2)} &&= \frac{-x}{\sigma^2}p(x) \\
\frac{\partial^2}{\partial x^2}\left[\frac{1}{\sqrt{2\pi}\sigma}e^{-x^2/(2\sigma^2)}\right] &= \frac{1}{\sqrt{2\pi}\sigma^5}\left(-\sigma^2+x^2\right)e^{-x^2/(2\sigma^2)} &&= \frac{-\sigma^2+x^2}{\sigma^4}p(x) \qquad (96)\\
\frac{\partial^3}{\partial x^3}\left[\frac{1}{\sqrt{2\pi}\sigma}e^{-x^2/(2\sigma^2)}\right] &= \frac{1}{\sqrt{2\pi}\sigma^7}\left(3x\sigma^2-x^3\right)e^{-x^2/(2\sigma^2)} &&= \frac{-3x\sigma^2-x^3}{\sigma^6}p(x),
\end{aligned}$$

and are shown in Fig. A.2.

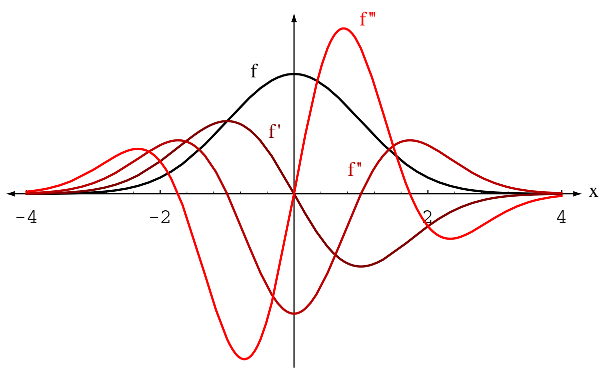

Figure A.2: A one-dimensional Gaussian distribution and its first three derivatives, shown for $f(x) \sim N(0,1)$.

An important finite integral of the Gaussian is the so-called *error function*, defined as

ERROR FUNCTION

$$\mathrm{erf}(u) = \sqrt{\frac{2}{\pi}}\int_0^u e^{-x^2/2}dx. \qquad (97)$$

As can be seen from Fig. A.1, $\mathrm{erf}(0) = 0$, $\mathrm{erf}(1) = 0.68$ and $\lim_{x\to\infty}\mathrm{erf}(x) = 1$. There is no closed analytic form for the error function, and thus we typically use tables, approximations or numerical integration for its evaluation (Fig. A.3).

In calculating moments of Gaussians, we need the general integral of powers of x weighted by a Gaussian. Recall first the definition of a *gamma function*

GAMMA FUNCTION

$$\Gamma(n+1) = \int_0^\infty x^n e^{-x}dx, \qquad (98)$$

where the gamma function obeys

$$\Gamma(n) = n\Gamma(n-1) \qquad (99)$$

and $\Gamma(1/2) = \sqrt{\pi}$. For n an integer we have $\Gamma(n+1) = n \times (n-1) \times (n-2) \dots 1 = n!$, read "n factorial."

FACTORIAL

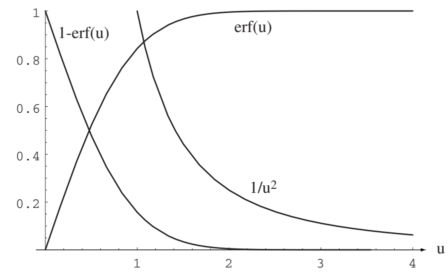

Figure A.3: The error function corresponds to the area under a standardized Gaussian (Eq. 97) between $-u$ and u, i.e., it describes the probability that a sample drawn from a standardized Gaussian obeys $|x| \leq u$. Thus, the complementary probability, $1 - \text{erf}(u)$ is the probability that a sample is chosen with $|x| > u$. Chebyshev's inequality states that for an *arbitrary* distribution having standard deviation = 1, this latter probability is bounded by $1/u^2$. As shown, this bound is quite loose for a Gaussian.

Changing variables in Eq. 98, we find the moments of a (normalized) Gaussian distribution as

$$2 \int_0^\infty x^n \frac{e^{-x^2/(2\sigma^2)}}{\sqrt{2\pi}\sigma} dx = \frac{2^{n/2}\sigma^n}{\sqrt{\pi}} \Gamma\left(\frac{n+1}{2}\right), \tag{100}$$

where again we have used a pre-factor of 2 and lower integration limit of 0 in order give non-trivial (i.e., non-vanishing) results for odd n.

A.5.1 Multivariate normal densities

Normal random variables have many desirable theoretical properties. For example, it turns out that the convolution of two Gaussian functions is again a Gaussian function, and thus the distribution for the sum of two independent normal random variables is again normal. In fact, sums of dependent normal random variables also have normal distributions. Suppose that each of the d random variables x_i is normally distributed, each with its own mean and variance: $p(x_i) \sim N(\mu_i, \sigma_i^2)$. If these variables are independent, their joint density has the form

$$\begin{aligned} p(\mathbf{x}) &= \prod_{i=1}^d p(x_i) = \prod_{i=1}^d \frac{1}{\sqrt{2\pi}\sigma_i} e^{-1/2((x_i-\mu_i)/\sigma_i)^2} \\ &= \frac{1}{(2\pi)^{d/2} \prod\limits_{i=1}^d \sigma_i} \exp\left[-\frac{1}{2} \sum_{i=1}^d \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 \right]. \end{aligned} \tag{101}$$

This can be written in a compact matrix form if we observe that for this case the covariance matrix is diagonal, i.e.,

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_d^2 \end{bmatrix}, \tag{102}$$

and hence the inverse of the covariance matrix is easily written as

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 & \dots & 0 \\ 0 & 1/\sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/\sigma_d^2 \end{bmatrix}. \tag{103}$$

Thus, the exponent in Eq. 101 can be rewritten using

$$\sum_{i=1}^{d} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 = (\mathbf{x} - \boldsymbol{\mu})^t \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}). \tag{104}$$

Finally, by noting that the determinant of $\mathbf{\Sigma}$ is just the product of the variances, we can write the joint density compactly in terms of the quadratic form

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^t \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}. \tag{105}$$

This is the general form of a *multivariate normal density function*, where the covariance matrix $\mathbf{\Sigma}$ is no longer required to be diagonal. With a little linear algebra, it can be shown that if $\mathbf{x}$ obeys this density function, then

$$\begin{aligned} \boldsymbol{\mu} = \mathcal{E}[\mathbf{x}] &= \int_{-\infty}^{\infty} \mathbf{x}\, p(\mathbf{x})\, d\mathbf{x} \\ \mathbf{\Sigma} = \mathcal{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t] &= \int_{-\infty}^{\infty} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t p(\mathbf{x})\, d\mathbf{x}, \end{aligned} \tag{106}$$

just as one would expect. Multivariate normal data tend to cluster about the mean vector, $\boldsymbol{\mu}$, falling in an ellipsoidally-shaped cloud whose principal axes are the eigenvectors of the covariance matrix. The natural measure of the distance from $\mathbf{x}$ to the mean $\boldsymbol{\mu}$ is provided by the quantity

$$r^2 = (\mathbf{x} - \boldsymbol{\mu})^t \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}), \tag{107}$$

which is the square of the Mahalanobis distance from $\mathbf{x}$ to $\boldsymbol{\mu}$. It is not as easy to standardize a vector random variable (reduce it to zero mean and unit covariance matrix) as it is in the univariate case. The expression analogous to $u = (x - \mu)/\sigma$ is $\mathbf{u} = \mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$, which involves the "square root" of the inverse of the covariance matrix. The process of obtaining $\mathbf{\Sigma}^{-1/2}$ requires finding the eigenvalues and eigenvectors of $\mathbf{\Sigma}$, and is just a bit beyond the scope of this Appendix.

A.5.2 Bivariate normal densities

It is illuminating to look at the bivariate normal density, that is, the case of two normally distributed random variables x_1 and x_2. It is convenient to define $\sigma_1^2 = \sigma_{11}, \sigma_2^2 = \sigma_{22}$, and to introduce the correlation coefficient ρ defined by

$$\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2}. \tag{108}$$

With this notation, the covariance matrix becomes

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \tag{109}$$

and its determinant simplifies to

$$|\mathbf{\Sigma}| = \sigma_1^2 \sigma_2^2 (1 - \rho^2). \tag{110}$$

Thus, the inverse covariance matrix is given by

$$\begin{aligned} \mathbf{\Sigma}^{-1} &= \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix} \\ &= \frac{1}{1-\rho^2} \begin{bmatrix} 1/\sigma_1^2 & -\rho/(\sigma_1\sigma_2) \\ -\rho/(\sigma_1\sigma_2 & 1/\sigma_2^2 \end{bmatrix}. \end{aligned} \tag{111}$$

Next we explicitly expand the quadratic form in the normal density:

$$\begin{aligned} & (\mathbf{x} - \boldsymbol{\mu})^t \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \\ = \; & [(x_1 - \mu_1)\,(x_2 - \mu_2)] \frac{1}{1-\rho^2} \begin{bmatrix} 1/\sigma_1^2 & -\rho/(\sigma_1\sigma_2) \\ -\rho/(\sigma_1\sigma_2) & 1/\sigma_2^2 \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \\ = \; & \frac{1}{1-\rho^2} \left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho \left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 \right]. \end{aligned} \tag{112}$$

Thus, the general bivariate normal density has the form

$$\begin{aligned} p_{x_1 x_2}(x_1, x_2) = & \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \\ & \exp\left[-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho \left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 \right] \right]. \end{aligned} \tag{113}$$

As we can see from Fig. A.4, $p(x_1, x_2)$ is a hill-shaped surface over the $x_1 x_2$ plane. The peak of the hill occurs at the point $(x_1, x_2) = (\mu_1, \mu_2)$, i.e., at the mean vector $\boldsymbol{\mu}$. The shape of the hump depends on the two variances σ_1^2 and σ_2^2, and the correlation coefficient ρ. If we slice the surface with horizontal planes parallel to the $x_1 x_2$ plane, we obtain the so-called *level curves*, defined by the locus of points where the quadratic form

$$\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho \left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 \tag{114}$$

is constant. It is not hard to show that $|\rho| \leq 1$, and that this implies that the level curves are ellipses. The x and y extent of these ellipses are determined by the variances σ_1^2 and σ_2^2, and their eccentricity is determined by ρ. More specifically, the *principal axes* of the ellipse are in the direction of the eigenvectors $\mathbf{e}_i$ of $\mathbf{\Sigma}$, and the different widths in these directions $\sqrt{\lambda_i}$. For instance, if $\rho = 0$, the principal axes of the ellipses are parallel to the coordinate axes, and the variables are statistically independent. In the special cases where $\rho = 1$ or $\rho = -1$, the ellipses collapse to straight lines. Indeed, the joint density becomes singular in this situation, because there is really only one independent variable. We shall avoid this degeneracy by assuming that $|\rho| < 1$.

PRINCIPAL AXES

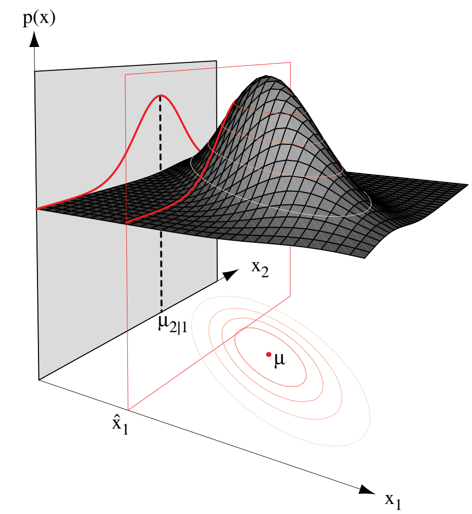

Figure A.4: A two-dimensional Gaussian having mean $\boldsymbol{\mu}$ and non-diagonal covariance $\mathbf{\Sigma}$. If the value on one variable is known, for instance $x_1 = \hat{x}_1$, the distribution over the other variable is Gaussian with mean $\mu_{2|1}$.

One of the important properties of the multivariate normal density is that all conditional and marginal probabilities are also normal. To find such a density explicitly, which we denote $p_{x_2|x_1}(x_2|x_1)$, we substitute our formulas for $p_{x_1x_2}(x_1, x_2)$ and $p_{x_1}(x_1)$ in the defining equation

$$\begin{aligned}
p_{x_2|x_1}(x_2|x_1) &= \frac{p_{x_1x_2}(x_1,x_2)}{p_{x_1}(x_1)} \\
&= \left[\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]}\right] \\
&\quad \times \left[\sqrt{2\pi}\sigma_1 e^{\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2}\right] \qquad (115) \\
&= \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left[\frac{x_2-\mu_2}{\sigma_2} - \rho\frac{x_1-\mu_1}{\sigma_1}\right]^2\right] \\
&= \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2}\left(\frac{x_2 - [\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)]}{\sigma_2\sqrt{1-\rho^2}}\right)^2\right].
\end{aligned}$$

Thus, we have verified that the conditional density $p_{x_1|x_2}(x_1|x_2)$ is a normal distribution. Moreover, we have explicit formulas for the *conditional mean* $\mu_{2|1}$ and the conditional variance $\sigma^2_{2|1}$:

CONDITIONAL MEAN

$$\begin{aligned} \mu_{2|1} &= \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1) \quad \text{and} \\ \sigma^2_{2|1} &= \sigma^2_2(1-\rho^2), \end{aligned} \tag{116}$$

as illustrated in Fig. A.4.

These formulas provide some insight into the question of how knowledge of the value of x_1 helps us to estimate x_2. Suppose that we know the value of x_1. Then a natural estimate for x_2 is the conditional mean, $\mu_{2|1}$. In general, $\mu_{2|1}$ is a linear function of x_1; if the correlation coefficient ρ is positive, the larger the value of x_1, the larger the value of $\mu_{2|1}$. If it happens that x_1 is the mean value μ_1, then the best we can do is to guess that x_2 is equal to μ_2. Also, if there is no correlation between x_1 and x_2, we ignore the value of x_1, whatever it is, and we always estimate x_2 by μ_2. Note that in that case the variance of x_2, given that we know x_1, is the same as the variance for the marginal distribution, i.e., $\sigma^2_{2|1} = \sigma^2_2$. If there is correlation, knowledge of the value of x_1, whatever the value is, reduces the variance. Indeed, with 100% correlation there is no variance left in x_2 when the value of x_1 is known.

A.6 Hypothesis testing

Suppose samples are drawn either from distribution D_0 or they are not. In pattern classification, we seek to determine which distribution was the source of any sample, and if it is indeed D_0, we would classify the point accordingly, into ω_1, say. Hypothesis testing addresses a somewhat different but related problem. We assume initially that distribution D_0 is the source of the patterns; this is called the *null hypothesis*, and often denoted H_0. Based on the value of any observed sample we ask whether we can *reject the null hypothesis*, that is, state with some degree of confidence (expressed as a probability) that the sample did *not* come from D_0.

For instance, D_0 might be a standardized Gaussian, $p(x) \sim N(0, 1)$, and our null hypothesis is that a sample comes from a Gaussian with mean $\mu = 0$. If the value of a particular sample is small (e.g., $x = 0.3$), it is likely that it came from the D_0; after all, 68% of the samples drawn from that distribution have absolute value less than $x = 1.0$ (cf. Fig. A.1). If a sample's value is large (e.g., $x = 5$), then we would be more confident that it did *not* come from D_0. At such a situation we merely conclude that (with some probability) the sample was drawn from a distribution with $\mu \neq 0$.

Viewed another way, for any confidence — expressed as a probability — there exists a criterion value such that if the sampled value differs from $\mu = 0$ by more than that criterion, we reject the null hypothesis. (It is traditional to use confidences of .01 or .05.) We then say that the difference of the sample from 0 is *statistically significant*. For instance, if our null hypothesis is a standardized Gaussian, then if our sample differs from the value $x = 0$ by more than 2.576, we could reject the null hypothesis "at the .01 confidence level," as can be deduced from Table A.1. A more sophisticated analysis could be applied if *several* samples are all drawn from D_0 or if the null hypothesis involved a distribution other than a Gaussian. Of course, this usage of "significance" applies only to the statistical properties of the problem — it implies nothing about whether the results are "important." Hypothesis testing is of

STATISTICAL SIGNIFICANCE

great generality, and useful when we seek to know whether something other than the assumed case (the null hypothesis) is the case.

A.6.1 Chi-squared test

Hypothesis testing can be applied to discrete problems too. Suppose we have n patterns — n_1 of which are known to be in ω_1, and n_2 in ω_2 — and we are interested in determining whether a particular decision rule is useful or informative. In this case, the null hypothesis is a random decision rule — one that selects a pattern and with some probability P places it in a category which we will call the "left" category, and otherwise in the "right" category. We say that a candidate rule is informative if it differs signficantly from such a random decision.

What we need is a clear mathematical definition of statistical significance under these conditions. The random rule (the null hypothesis) would place Pn_1 patterns from ω_1 and Pn_2 from ω_2 independently in the left category and the remainder in the right category. Our candidate decision rule would differ significantly from the random rule if the proportions differed significantly from those given by the random rule. Formally, we let n_{iL} denote the number of patterns from category ω_i placed in the left category by our candidate rule. The so-called *chi-squared* statistic for this case is

$$\chi^2 = \sum_{k=1}^{2} \frac{(n_{iL} - n_{ie})^2}{n_{ie}}. \tag{117}$$

where according to the null hypothesis, the number of patterns in category ω_i that we expect to be placed in the left category is $n_{ie} = Pn_i$. Clearly χ^2 is non-negative, and is zero if and only if all the observed match the expected numbers. The higher the χ^2, the less likely it is that the null hypothesis is true. Thus, for a sufficiently high χ^2, the difference between the expected and observed distributions is statistically significant, we can reject the null hypothesis, and can consider our candidate decision rule is "informative." For any desired level of significance — such as .01 or .05 — a table gives the critical values of χ^2 that allow us to reject the null hypothesis (Table A.2).

There is one detail that must be addressed: the number of degrees of freedom. In the situation described above, once the probability P is known, there is only one free variable needed to describe a candidate rule. For instance, once the number of patterns from ω_1 placed in the left category are known, all other values are determined uniquely. Hence in this case the number of degrees of freedom is 1. If there were more categories, or if the candidate decision rule had more possible outcomes, then *df* would be greater than 1. The higher the number of degrees of freedom, the higher must be the computed χ^2 to meet a disired level of significance.

We denote the critical values as, for instance, $\chi^2_{.01(1)} = 6.64$, where the subscript denotes the significance, here .01, and the integer in parentheses is the degrees of freedom. (In the Table, we conform to the usage in statistics, where this positive integer is denoted *df*, despite the possible confusion in calculus where it denotes an infinitessimal real number.) Thus if we have one degree of freedom, and the observed χ^2 is greater than 6.64, then we can reject the null hypothesis, and say that, at the .01 confidence level our results did not come from a (weighted) random decision.

Table A.2: Critical values of chi-square (at two confidence levels) for different degrees of freedom (*df*)

df	.05	.01	*df*	.05	.01	*df*	.05	.01
1	3.84	6.64	11	19.68	24.72	21	32.67	38.93
2	5.99	9.21	12	21.03	26.22	22	33.92	40.29
3	7.82	11.34	13	22.36	27.69	23	35.17	41.64
4	9.49	13.28	14	23.68	29.14	24	36.42	42.98
5	11.07	15.09	15	25.00	30.58	25	37.65	44.31
6	12.59	16.81	16	26.30	32.00	26	38.88	45.64
7	14.07	18.48	17	27.59	33.41	27	40.11	46.96
8	15.51	20.09	18	28.87	34.80	28	41.34	48.28
9	16.92	21.67	19	30.14	37.57	29	42.56	49.59
10	18.31	23.21	20	31.41	37.57	30	43.77	50.89

A.7 Information theory

A.7.1 Entropy and information

Assume we have a discrete set of symbols $\{v_1\ v_2\ \dots\ v_m\}$ with associated probabilities P_i. The entropy of the discrete distribution — a measure of the randomness or unpredictability of a sequence of symbols drawn from it — is

$$H = -\sum_{i=1}^{m} P_i \log_2 P_i, \tag{118}$$

BIT

where since we use the logarithm base 2 entropy is measured in *bits*. In case any of the probabilities vanish, we use the relation $0 \log 0 = 0$. One bit corresponds to the uncertainty that can be resolved by the answer to a single yes/no question. (For continuous distributions, we often use logarithm base e, denoted ln, in which case the unit is *nat*.) The expectation operator (cf. Eq. 41) can be used to write $H = \mathcal{E}[\log 1/P]$, where we think of P as being a random variable whose possible values are $P_1, P_2, \dots, P_m$. The term $\log_2 1/P$ is sometimes called the *surprise* — if $P_i = 0$ except for one i, then there is no surprise when the corresponding symbol occurs.

SURPRISE

Note that the entropy does not depend on the symbols themselves, just on their probabilities. For a given number of symbols m, the uniform distribution in which each symbol is equally likely, is the *maximum entropy distribution* (and $H = \log_2 m$ bits) — we have the maximum uncertainty about the identity of each symbol that will be chosen. Clearly if x is equally likely to take on integer values $0, 1, ..., 7$, we need 3 bits to describe the outcome and $H = \log_2 2^3 = 3$. Conversely, if all the p_i are 0 except one, we have the *minimum entropy distribution* ($H = 0$ bits) — we are certain as to the symbol that will appear.

For a continuous distribution, the entropy is

$$H = -\int_{-\infty}^{\infty} p(x) \ln p(x) dx, \tag{119}$$

and again $H = \mathcal{E}[\ln 1/p]$. It is worth mentioning that among all continuous density functions having a given mean μ and variance σ^2, it is the Gaussian that has the maximum entropy ($H = .5 + \log_2(\sqrt{2\pi}\sigma)$ bits). We can let σ approach zero to find that a probability density in the form of a *Dirac delta* function, i.e., DIRAC DELTA

$$\delta(x-a) = \begin{cases} 0 & \text{if } x \neq a \\ \infty & \text{if } x = a, \end{cases} \quad \text{with}$$

$$\int_{-\infty}^{\infty} \delta(x)dx = 1, \tag{120}$$

has the minimum entropy ($H = -\infty$ bits). For a Dirac function, we are sure that the value a will be selected each time.

Our use of entropy in continuous functions, such as in Eq. 119, belies some subtle issues which are worth pointing out. If x had units, such as meters, then the probability density $p(x)$ would have to have units of $1/x$. There would be something fundamentally wrong in taking the logarithm of $p(x)$ — the argument of the logarithm function should be dimensionless. What we should really be dealing with is a dimensionless quantity, say $p(x)/p_0(x)$, where $p_0(x)$ is some reference density function (cf., Sect. A.7.2).

For discrete variable x and arbitrary function $f(\cdot)$, we have $H(f(x)) \leq H(x)$, i.e., processing decreases entropy. For instance, if $f(x) = const$, the entropy will vanish. Another key property of the entropy of a discrete distribution is that it is invariant to "shuffling" the event labels. The related question with continuous variables concerns what happens when one makes a change of variables. In general, if we make a change of variables, such as $y = x^3$ or even $y = 10x$, we will get a different value for the integral of $\int q(y) \log q(y)\, dy$, where q is the induced density for y. If entropy is supposed to measure the intrinsic disorganization, it doesn't make sense that y would have a different amount of intrinsic disorganization than x, since one is always derivable from the other; only if there were some randomness (e.g., shuffling) incorporated into the mapping could we say that one is more disorganized than the other.

Fortunately, in practice these concerns do not present important stumbling blocks since relative entropy and differences in entropy are more fundamental than H taken by itself. Nevertheless, questions of the foundations of entropy measures for continuous variables are addressed in books listed in Bibliographical Remarks.

A.7.2 Relative entropy

Suppose we have two discrete distributions over the same variable x, $p(x)$ and $q(x)$. The relative entropy or *Kullback-Leibler distance* (which is closely related to cross entropy, information divergence and information for discrimination) is a measure of the "distance" between these distributions: KULLBACK-LEIBLER DISTANCE

$$D_{KL}(p(x), q(x)) = \sum_x q(x) \ln \frac{q(x)}{p(x)}. \tag{121}$$

The continuous version is

$$D_{KL}(p(x), q(x)) = \int_{-\infty}^{\infty} q(x) \ln \frac{q(x)}{p(x)} dx. \tag{122}$$

Although $D_{KL}(p(\cdot), q(\cdot)) \geq 0$ and $D_{KL}(p(\cdot), q(\cdot)) = 0$ if and only if $p(\cdot) = q(\cdot)$, the relative entropy is not a true metric, since D_{KL} is not necessarily symmetric in the interchange $p \leftrightarrow q$ and furthermore the triangle inequality need not be satisfied.

A.7.3 Mutual information

Now suppose we have two distributions over possibly *different* variables, e.g., $p(x)$ and $q(y)$. The mutual information is the reduction in uncertainty about one variable due to the knowledge of the other variable

$$I(p;\ q) = H(p) - H(p|q) = \sum_{x,y} r(x,y) \log \frac{r(x,y)}{p(x)q(y)}, \tag{123}$$

where $r(x, y)$ is the joint distribution of finding value x and y. Mutual information is simply the relative entropy between the joint distribution $r(x, y)$ and the product distribution $p(x)q(y)$ and as such it measures how much the distributions of the variables differ from statistical independence. Mutual information does not obey all the properties of a metric. In particular, the metric requirement that if $p(x) = q(y)$ then $I(x;\ y) = 0$ need not hold, in general. As an example, suppose we have two binary random variables with $r(0, 0) = r(1, 1) = 1/2$, so $r(0, 1) = r(1, 0) = 0$. According to Eq. 123, the mutual information between $p(x)$ and $q(y)$ is $\log 2 = 1$.

The relationships among the entropy, relative entropy and mutual information are summarized in Fig. A.5. The figure shows, for instance, that the joint entropy $H(p, q)$ is always larger than individual entropies $H(p)$ and $H(q)$; that $H(p) = H(p|q) + I(p;\ q)$, and so on.

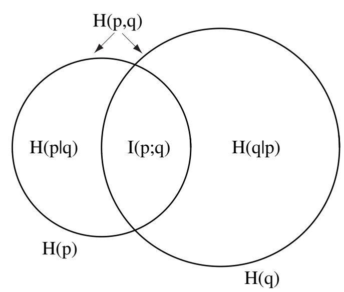

Figure A.5: The mathematical relationships among the entropy of distributions p and q, mutual information $I(p, q)$, and conditional entropies $H(p|q)$ and $H(q|p)$. From this figure one can quickly see relationships among the information functions. For instance we can see immediately that $I(p;\ p) = H(p)$; that if $I(p;\ q) = 0$ then $H(q|p) = H(q)$; that $H(p, q) = H(p|q) + H(q)$, and so forth.

A.8 Computational complexity

In order to analyze and describe the difficulty of problems and the algorithms designed to solve such problems, we turn now to the technical notion of computational complexity. For instance, calculating the covariance matrix for a samples is somehow "harder" than calculating the mean. Furthermore, some algorithms for computing some function may be faster or take less memory, than another algorithm. We seek to specify such differences, independent of the current computer hardware (which is always changing anyway).

To this end we use the concept of the order of a function and the asymptotic notations "big oh," "big omega," and "big theta." The three asymptotic bounds most often used are:

Asymptotic upper bound $O(g(x)) = \{f(x)$: there exist positive constants c and x_0 such that $0 \leq f(x) \leq cg(x)$ for all $x \geq x_0\}$

Asymptotic lower bound $\boldsymbol{\Omega}(g(x)) = \{f(x)$: there exist positive constants c and x_0 such that $0 \leq cg(x) \leq f(x)$ for all $x \geq x_0\}$

Asymptotically tight bound $\boldsymbol{\Theta}(g(x)) = \{f(x)$: there exist positive constants c_1, c_2, and x_0 such that $0 \leq c_1 g(x) \leq f(x) \leq c_2 g(x)$ for all $x \geq x_0\}$

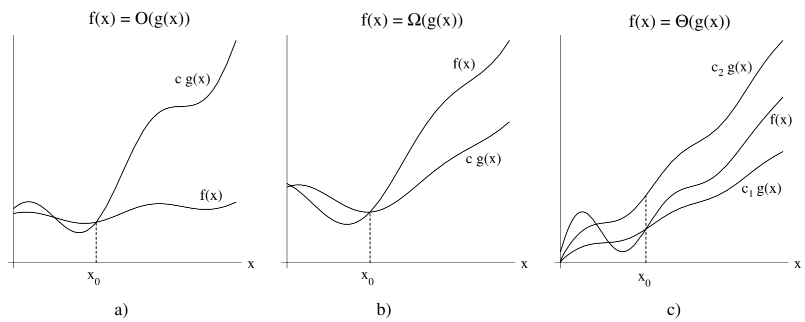

Figure A.6: Three types of asymptotic bounds: a) $f(x) = O(g(x))$. b) $f(x) = \boldsymbol{\Omega}(g(x))$. c) $f(x) = \boldsymbol{\Theta}(g(x))$.

Consider the asymptotic upper bound. We say that $f(x)$ is "of order big oh of $g(x)$" (written $f(x) = O(g(x))$ if there exist constants c_0 and x_0 such that $f(x) \leq c_0 g(x)$ for all $x > x_0$. We shall assume that all our functions are positive and dispense with taking absolute values. This means simply that for sufficiently large x, an upper bound on $f(x)$ grows no worse than $g(x)$. For instance, if $f(x) = a + bx + cx^2$ then $f(x) = O(x^2)$ because for sufficiently large x, the constant, linear and quadratic terms can be "overcome" by proper choice of c_0 and x_0. The generalization to functions of two or more variables is straightforward. It should be clear that by the definition above, the (big oh) order of a function is not unique. For instance, we can describe our particular $f(x)$ as being $O(x^2)$, $O(x^3)$, $O(x^4)$, $O(x^2 \ln x)$, and so forth. We use big omega notation, $\boldsymbol{\Omega}(\cdot)$, for lower bounds, and little omega, $\omega(\cdot)$, for the tightest lower bound. Of these, the big oh notation has proven to be most useful since we generally want an *upper* bound on the resources when solving a problem. BIG OH

The lower bound on the complexity of the *problem* is denoted $\boldsymbol{\Omega}(g(x))$, and is therefore the lower bound on any algorithm algorithm that solves that problem. Similarly, if the complexity of an algorithm is $O(g(x))$, it is an upper bound on the complexity of the problem it solves. The complexity of some problems — such as computing the mean of a discrete set — is known, and thus once we have found an algorithm having equal complexity, the only possible improvement could be on lowering the constants of proportionality. The complexity of other problems — such as inverting a matrix

— is not yet known, and if fundamental analysis cannot derive it, we must rely on algorithm developers who find algorithms whose complexity

Approximately.

Such a rough analysis does not tell us the constants c and x_0. For a finite size problem it is possible that a particular $O(x^3)$ algorithm is simpler than a particular $O(x^2)$ algorithm, and it is occasionally necessary for us to determine these constants to find which of several implemementations is the simplest. Nevertheless, for our purposes the big oh notation as just described is generally the best way to describe the computational complexity of an algorithm.

Suppose we have a set of n vectors, each of which is d-dimensional and we want to calculate the mean vector. Clearly, this requires $O(nd)$ multiplications. Sometimes we stress space and time complexities, which are particularly relevant when contemplating parallel hardware implementations. For instance, the d-dimensional sample mean could be calculated with d separate processors, each adding n sample values. Thus we can describe this implementation as $O(d)$ in *space* (i.e., the amount of memory or possibly the number of processors) and $O(n)$ in *time* (i.e., number of sequential steps). Of course for any particular algorithm there may be a number of time-space tradeoffs.

SPACE COMPLEXITY

TIME COMPLEXITY

Bibliographical Remarks

There are several good books on linear systems, such as [14], and matrix computations [8]. Lagrange optimization and related techniques are covered in the definitive book [2]. While [13] and [3] are of foundational and historic interest, readers seeking clear presentations of the central ideas in probability should consult [10, 7, 6, 21]. A handy reference to terms in probability and statistics is [20]. A number of hypothesis testing and statistical significance, elementary, such as [24], and more advanced [18, 25]. Shannon's foundational paper [22] should be read by all students of pattern recognition. It, and many other historically important papers on information theory can be found in [23]. An excellent textbook at the level of this one is [5] and readers seeking a more abstract and formal treatment should consult [9]. The study of time complexity of algorithms began with [12], and space complexity [11, 19]. The multivolume [15, 16, 17] contains a description of computational complexity, the big oh and other asymptotic notations. Somewhat more accessible treatments can be found in [4] and [1].