Appendix A

Mathematical Background

A.1 Joint, Marginal and Conditional Probability

Let the n (discrete or continuous) random variables $y_1, \ldots, y_n$ have a *joint* probability $p(y_1, \ldots, y_n)$, or $p(\mathbf{y})$ for short.[1] Technically, one ought to distinguish between probabilities (for discrete variables) and probability densities for continuous variables. Throughout the book we commonly use the term "probability" to refer to both. Let us partition the variables in $\mathbf{y}$ into two groups, $\mathbf{y}_A$ and $\mathbf{y}_B$, where A and B are two disjoint sets whose union is the set $\{1, \ldots, n\}$, so that $p(\mathbf{y}) = p(\mathbf{y}_A, \mathbf{y}_B)$. Each group may contain one or more variables.

joint probability

The *marginal* probability of $\mathbf{y}_A$ is given by

marginal probability

$$p(\mathbf{y}_A) \;=\; \int p(\mathbf{y}_A, \mathbf{y}_B)\, d\mathbf{y}_B. \tag{A.1}$$

The integral is replaced by a sum if the variables are discrete valued. Notice that if the set A contains more than one variable, then the marginal probability is itself a joint probability—whether it is referred to as one or the other depends on the context. If the joint distribution is equal to the product of the marginals, then the variables are said to be *independent*, otherwise they are *dependent*.

independence

The *conditional* probability function is defined as

conditional probability

$$p(\mathbf{y}_A|\mathbf{y}_B) \;=\; \frac{p(\mathbf{y}_A, \mathbf{y}_B)}{p(\mathbf{y}_B)}, \tag{A.2}$$

defined for $p(\mathbf{y}_B) > 0$, as it is not meaningful to condition on an impossible event. If $\mathbf{y}_A$ and $\mathbf{y}_B$ are independent, then the marginal $p(\mathbf{y}_A)$ and the conditional $p(\mathbf{y}_A|\mathbf{y}_B)$ are equal.

[1] One can deal with more general cases where the density function does not exist by using the distribution function.

Bayes' rule

Using the definitions of both $p(\mathbf{y}_A|\mathbf{y}_B)$ and $p(\mathbf{y}_B|\mathbf{y}_A)$ we obtain *Bayes' theorem*

$$p(\mathbf{y}_A|\mathbf{y}_B) \;=\; \frac{p(\mathbf{y}_A)p(\mathbf{y}_B|\mathbf{y}_A)}{p(\mathbf{y}_B)}. \tag{A.3}$$

Since conditional distributions are themselves probabilities, one can use all of the above also when further conditioning on other variables. For example, in supervised learning, one often conditions on the inputs throughout, which would lead e.g. to a version of Bayes' rule with additional conditioning on X in all four probabilities in eq. (A.3); see eq. (2.5) for an example of this.

A.2 Gaussian Identities

Gaussian definition

The multivariate Gaussian (or Normal) distribution has a joint probability density given by

$$p(\mathbf{x}|\mathbf{m},\Sigma) \;= (2\pi)^{-D/2}|\Sigma|^{-1/2}\exp\left(-\tfrac{1}{2}(\mathbf{x}-\mathbf{m})^\top\Sigma^{-1}(\mathbf{x}-\mathbf{m})\right), \tag{A.4}$$

where $\mathbf{m}$ is the *mean* vector (of length D) and Σ is the (symmetric, positive definite) *covariance* matrix (of size $D \times D$). As a shorthand we write $\mathbf{x} \sim \mathcal{N}(\mathbf{m},\Sigma)$.

Let $\mathbf{x}$ and $\mathbf{y}$ be jointly Gaussian random vectors

$$\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} \;\sim\; \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}A & C\\ C^\top & B\end{bmatrix}\right) \;=\; \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\tilde{A} & \tilde{C}\\ \tilde{C}^\top & \tilde{B}\end{bmatrix}^{-1}\right), \tag{A.5}$$

conditioning and marginalizing

then the *marginal* distribution of $\mathbf{x}$ and the *conditional* distribution of $\mathbf{x}$ given $\mathbf{y}$ are

$$\begin{aligned}\mathbf{x} \;\sim\; \mathcal{N}(\boldsymbol{\mu}_x, A), \quad \text{and} \quad \mathbf{x}|\mathbf{y} \;&\sim\; \mathcal{N}\big(\boldsymbol{\mu}_x + CB^{-1}(\mathbf{y}-\boldsymbol{\mu}_y),\; A - CB^{-1}C^\top\big)\\ \text{or} \quad \mathbf{x}|\mathbf{y} \;&\sim\; \mathcal{N}\big(\boldsymbol{\mu}_x - \tilde{A}^{-1}\tilde{C}(\mathbf{y}-\boldsymbol{\mu}_y),\; \tilde{A}^{-1}\big).\end{aligned} \tag{A.6}$$

See, e.g. von Mises [1964, sec. 9.3], and eqs. (A.11 - A.13).

products

The product of two Gaussians gives another (un-normalized) Gaussian

$$\begin{aligned}\mathcal{N}(\mathbf{x}|\mathbf{a},A)\mathcal{N}(\mathbf{x}|\mathbf{b},B) \;&= Z^{-1}\mathcal{N}(\mathbf{x}|\mathbf{c},C)\\ \text{where} \quad \mathbf{c} \;&= C(A^{-1}\mathbf{a} + B^{-1}\mathbf{b}) \quad \text{and} \quad C \;=\; (A^{-1}+B^{-1})^{-1}.\end{aligned} \tag{A.7}$$

Notice that the resulting Gaussian has a precision (inverse variance) equal to the sum of the precisions and a mean equal to the convex sum of the means, weighted by the precisions. The normalizing constant looks itself like a Gaussian (in $\mathbf{a}$ or $\mathbf{b}$)

$$Z^{-1} \;=\; (2\pi)^{-D/2}|A+B|^{-1/2}\exp\left(-\tfrac{1}{2}(\mathbf{a}-\mathbf{b})^\top(A+B)^{-1}(\mathbf{a}-\mathbf{b})\right). \tag{A.8}$$

To prove eq. (A.7) simply write out the (lengthy) expressions by introducing eq. (A.4) and eq. (A.8) into eq. (A.7), and expand the terms inside the exp to

verify equality. Hint: it may be helpful to expand C using the matrix inversion lemma, eq. (A.9), $C = (A^{-1}+B^{-1})^{-1} = A-A(A+B)^{-1}A = B-B(A+B)^{-1}B$.

generating multivariate Gaussian samples

To generate samples $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, K)$ with arbitrary mean $\mathbf{m}$ and covariance matrix K using a scalar Gaussian generator (which is readily available in many programming environments) we proceed as follows: first, compute the Cholesky decomposition (also known as the "matrix square root") L of the positive definite symmetric covariance matrix $K = LL^\top$, where L is a lower triangular matrix, see section A.4. Then generate $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, I)$ by multiple separate calls to the scalar Gaussian generator. Compute $\mathbf{x} = \mathbf{m} + L\mathbf{u}$, which has the desired distribution with mean $\mathbf{m}$ and covariance $L\mathbb{E}[\mathbf{u}\mathbf{u}^\top]L^\top = LL^\top = K$ (by the independence of the elements of $\mathbf{u}$).

In practice it may be necessary to add a small multiple of the identity matrix ϵI to the covariance matrix for numerical reasons. This is because the eigenvalues of the matrix K can decay very rapidly (see section 4.3.1 for a closely related analytical result) and without this stabilization the Cholesky decomposition fails. The effect on the generated samples is to add additional independent noise of variance ϵ. From the context ϵ can usually be chosen to have inconsequential effects on the samples, while ensuring numerical stability.

A.3 Matrix Identities

matrix inversion lemma

The *matrix inversion lemma*, also known as the Woodbury, Sherman & Morrison formula (see e.g. Press et al. [1992, p. 75]) states that

$$(Z + UWV^\top)^{-1} \;=\; Z^{-1} - Z^{-1}U(W^{-1} + V^\top Z^{-1}U)^{-1}V^\top Z^{-1}, \qquad \text{(A.9)}$$

assuming the relevant inverses all exist. Here Z is $n\times n$, W is $m\times m$ and U and V are both of size $n\times m$; consequently if Z^{-1} is known, and a low rank (i.e. $m < n$) perturbation is made to Z as in left hand side of eq. (A.9), considerable speedup can be achieved. A similar equation exists for determinants

determinants

$$|Z + UWV^\top| \;=\; |Z|\,|W|\,|W^{-1} + V^\top Z^{-1}U|. \qquad \text{(A.10)}$$

inversion of a partitioned matrix

Let the invertible $n \times n$ matrix A and its inverse A^{-1} be partitioned into

$$A \;=\; \begin{pmatrix} P & Q \\ R & S \end{pmatrix}, \qquad A^{-1} \;=\; \begin{pmatrix} \tilde{P} & \tilde{Q} \\ \tilde{R} & \tilde{S} \end{pmatrix}, \qquad \text{(A.11)}$$

where P and $\tilde{P}$ are $n_1 \times n_1$ matrices and S and $\tilde{S}$ are $n_2 \times n_2$ matrices with $n = n_1 + n_2$. The submatrices of A^{-1} are given in Press et al. [1992, p. 77] as

$$\left.\begin{aligned} \tilde{P} &= P^{-1} + P^{-1}QMRP^{-1} \\ \tilde{Q} &= -P^{-1}QM \\ \tilde{R} &= -MRP^{-1} \\ \tilde{S} &= M \end{aligned}\right\} \text{where } M \;=\; (S - RP^{-1}Q)^{-1}, \qquad \text{(A.12)}$$

or equivalently

$$\left.\begin{array}{rcl} \tilde{P} &=& N \\ \tilde{Q} &=& -NQS^{-1} \\ \tilde{R} &=& -S^{-1}RN \\ \tilde{S} &=& S^{-1} + S^{-1}RNQS^{-1} \end{array}\right\} \text{where } N = (P - QS^{-1}R)^{-1}. \tag{A.13}$$

A.3.1 Matrix Derivatives

derivative of inverse

Derivatives of the elements of an inverse matrix:

$$\frac{\partial}{\partial\theta}K^{-1} = -K^{-1}\frac{\partial K}{\partial\theta}K^{-1}, \tag{A.14}$$

derivative of log determinant

where $\frac{\partial K}{\partial\theta}$ is a matrix of elementwise derivatives. For the log determinant of a positive definite symmetric matrix we have

$$\frac{\partial}{\partial\theta}\log|K| = \mathrm{tr}\left(K^{-1}\frac{\partial K}{\partial\theta}\right). \tag{A.15}$$

A.3.2 Matrix Norms

The Frobenius norm $\|A\|_F$ of a $n_1 \times n_2$ matrix A is defined as

$$\|A\|_F^2 = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}|a_{ij}|^2 = \mathrm{tr}(AA^\top), \tag{A.16}$$

[Golub and Van Loan, 1989, p. 56].

A.4 Cholesky Decomposition

The Cholesky decomposition of a symmetric, positive definite matrix A decomposes A into a product of a lower triangular matrix L and its transpose

$$LL^\top = A, \tag{A.17}$$

where L is called the Cholesky factor. The Cholesky decomposition is useful for solving linear systems with symmetric, positive definite coefficient matrix A.

solving linear systems

To solve $A\mathbf{x} = \mathbf{b}$ for $\mathbf{x}$, first solve the triangular system $L\mathbf{y} = \mathbf{b}$ by forward substitution and then the triangular system $L^\top\mathbf{x} = \mathbf{y}$ by back substitution. Using the backslash operator, we write the solution as $\mathbf{x} = L^\top\backslash(L\backslash\mathbf{b})$, where the notation $A\backslash\mathbf{b}$ is the vector $\mathbf{x}$ which solves $A\mathbf{x} = \mathbf{b}$.

computational cost

Both the forward and backward substitution steps require $n^2/2$ operations, when A is of size $n \times n$. The computation of the Cholesky factor L is considered numerically extremely stable and takes time $n^3/6$, so it is the method of choice when it can be applied.

Note also that the determinant of a positive definite symmetric matrix can be calculated efficiently by

determinant

$$|A| \;=\; \prod_{i=1}^{n} L_{ii}^2, \quad \text{or} \quad \log|A| \;=\; 2\sum_{i=1}^{n} \log L_{ii}, \tag{A.18}$$

where L is the Cholesky factor from A.

A.5 Entropy and Kullback-Leibler Divergence

The *entropy* $H[p(\mathbf{x})]$ of a distribution $p(\mathbf{x})$ is a non-negative measure of the amount of "uncertainty" in the distribution, and is defined as

entropy

$$H[p(\mathbf{x})] \;=\; -\int p(\mathbf{x}) \log p(\mathbf{x})\, d\mathbf{x}. \tag{A.19}$$

The integral is substituted by a sum for discrete variables. Entropy is measured in *bits* if the log is to the base 2 and in *nats* in the case of the natural log. The entropy of a Gaussian in D dimensions, measured in nats is

$$H[\mathcal{N}(\boldsymbol{\mu}, \Sigma)] \;=\; \tfrac{1}{2}\log|\Sigma| + \tfrac{D}{2}(\log 2\pi e). \tag{A.20}$$

The Kullback-Leibler (KL) divergence (or relative entropy) $\mathrm{KL}(p||q)$ between two distributions $p(\mathbf{x})$ and $q(\mathbf{x})$ is defined as

$$\mathrm{KL}(p||q) \;=\; \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})}\, d\mathbf{x}. \tag{A.21}$$

It is easy to show that $\mathrm{KL}(p||q) \geq 0$, with equality if $p = q$ (almost everywhere). For the case of two Bernoulli random variables p and q this reduces to

divergence of Bernoulli random variables

$$\mathrm{KL}_{\mathrm{Ber}}(p||q) \;=\; p \log \frac{p}{q} + (1-p) \log \frac{(1-p)}{(1-q)}, \tag{A.22}$$

where we use p and q both as the name and the parameter of the Bernoulli distributions. For two Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$ and $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ we have [Kullback, 1959, sec. 9.1]

divergence of Gaussians

$$\begin{aligned} \mathrm{KL}(\mathcal{N}_0||\mathcal{N}_1) \;&=\; \tfrac{1}{2}\log|\Sigma_1\Sigma_0^{-1}| + \\ &\quad \tfrac{1}{2}\operatorname{tr}\Sigma_1^{-1}\big((\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top + \Sigma_0 - \Sigma_1\big). \end{aligned} \tag{A.23}$$

Consider a general distribution $p(\mathbf{x})$ on $\mathbb{R}^D$ and a Gaussian distribution $q(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \Sigma)$. Then

minimizing $KL(p||q)$ divergence leads to moment matching

$$\begin{aligned} \mathrm{KL}(p||q) \;&=\; \int \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) p(\mathbf{x})\, d\mathbf{x} + \\ &\quad \tfrac{1}{2}\log|\Sigma| + \tfrac{D}{2}\log 2\pi + \int p(\mathbf{x}) \log p(\mathbf{x})\, d\mathbf{x}. \end{aligned} \tag{A.24}$$

Equation (A.24) can be minimized w.r.t. $\boldsymbol{\mu}$ and Σ by differentiating w.r.t. these parameters and setting the resulting expressions to zero. The optimal q is the one that matches the first and second moments of p.

The KL divergence can be viewed as the extra number of nats needed on average to code data generated from a source $p(\mathbf{x})$ under the distribution $q(\mathbf{x})$ as opposed to $p(\mathbf{x})$.

A.6 Limits

The limit of a rational quadratic is a squared exponential

$$\lim_{\alpha\to\infty} \big(1 + \frac{x^2}{2\alpha}\big)^{-\alpha} \;=\; \exp\big(-\frac{x^2}{2}\big). \tag{A.25}$$

A.7 Measure and Integration

Here we sketch some definitions concerning measure and integration; fuller treatments can be found e.g. in Doob [1994] and Bartle [1995].

Let Ω be the set of all possible outcomes of an experiment. For example, for a D-dimensional real-valued variable, $\Omega = \mathbb{R}^D$. Let $\mathcal{F}$ be a σ-field of subsets of Ω which contains all the events in whose occurrences we may be interested.[2] Then μ is a countably additive *measure* if it is real and non-negative and for all mutually disjoint sets $A_1,\ A_2, \ldots \in \mathcal{F}$ we have

$$\mu\big(\bigcup_{i=1}^{\infty} A_i\big) \;=\; \sum_{i=1}^{\infty} \mu(A_i). \tag{A.26}$$

finite measure
probability measure
Lebesgue measure

If $\mu(\Omega) < \infty$ then μ is called a *finite measure* and if $\mu(\Omega) = 1$ it is called a *probability measure*. The *Lebesgue measure* defines a uniform measure over subsets of Euclidean space. Here an appropriate σ-algebra is the Borel σ-algebra $\mathcal{B}^D$, where $\mathcal{B}$ is the σ-algebra generated by the open subsets of $\mathbb{R}$. For example on the line $\mathbb{R}$ the Lebesgue measure of the interval (a, b) is $b - a$.

We now restrict Ω to be $\mathbb{R}^D$ and wish to give meaning to integration of a function $f : \mathbb{R}^D \to \mathbb{R}$ with respect to a measure μ

$$\int f(\mathbf{x})\, d\mu(\mathbf{x}). \tag{A.27}$$

We assume that f is *measurable*, i.e. that for any Borel-measurable set $A \in \mathbb{R}$, $f^{-1}(A) \in \mathcal{B}^D$. There are two cases that will interest us (i) when μ is the Lebesgue measure and (ii) when μ is a probability measure. For the first case expression (A.27) reduces to the usual integral notation $\int f(\mathbf{x}) d\mathbf{x}$.

[2] The restriction to a σ-field of subsets is important technically to avoid paradoxes such as the Banach-Tarski paradox. Informally, we can think of the σ-field as restricting consideration to "reasonable" subsets.

For a probability measure μ on $\mathbf{x}$, the non-negative function $p(\mathbf{x})$ is called the *density* of the measure if for all $A \in \mathcal{B}^D$ we have

$$\mu(A) \;=\; \int_A p(\mathbf{x})\, d\mathbf{x}. \tag{A.28}$$

If such a density exists it is uniquely determined almost everywhere, i.e. except for sets with measure zero. Not all probability measures have densities—only distributions that assign zero probability to individual points in $\mathbf{x}$-space can have densities.[3] If $p(\mathbf{x})$ exists then we have

$$\int f(\mathbf{x})\, d\mu(\mathbf{x}) \;=\; \int f(\mathbf{x}) p(\mathbf{x})\, d\mathbf{x}. \tag{A.29}$$

If μ does not have a density expression (A.27) still has meaning by the standard construction of the Lebesgue integral.

For $\Omega = \mathbb{R}^D$ the probability measure μ can be related to the *distribution function* $F : \mathbb{R}^D \to [0,1]$ which is defined as $F(\mathbf{z}) = \mu(x_1 \le z_1, \ldots x_D \le z_D)$. The distribution function is more general than the density as it is always defined for a given probability measure. A simple example of a random variable which has a distribution function but no density is obtained by the following construction: a coin is tossed and with probability p it comes up heads; if it comes up heads x is chosen from $U(0,1)$ (the uniform distribution on $[0,1]$), otherwise (with probability $1-p$) x is set to $1/2$. This distribution has a "point mass" (or atom) at $x = 1/2$.

"point mass" example

A.7.1 L_p Spaces

Let μ be a measure on an input set $\mathcal{X}$. For some function $f : \mathcal{X} \to \mathbb{R}$ and $1 \le p < \infty$, we define

$$\|f\|_{L_p(\mathcal{X},\mu)} \;\triangleq\; \Big(\int |f(\mathbf{x})|^p \, d\mu(\mathbf{x}) \Big)^{1/p}, \tag{A.30}$$

if the integral exists. For $p = \infty$ we define

$$\|f\|_{L_\infty(\mathcal{X},\mu)} \;=\; \operatorname*{ess\,sup}_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x})|, \tag{A.31}$$

where ess sup denotes the essential supremum, i.e. the smallest number that upper bounds $|f(\mathbf{x})|$ almost everywhere. The function space $L_p(\mathcal{X},\mu)$ is defined for any p in for $1 \le p \le \infty$ as the space of functions for which $\|f\|_{L_p(\mathcal{X},\mu)} < \infty$.

A.8 Fourier Transforms

For sufficiently well-behaved functions on $\mathbb{R}^D$ we have

$$f(\mathbf{x}) \;=\; \int_{-\infty}^{\infty} \tilde{f}(\mathbf{s}) e^{2\pi i \mathbf{s}\cdot\mathbf{x}}\, d\mathbf{s}, \qquad \tilde{f}(\mathbf{s}) \;=\; \int_{-\infty}^{\infty} f(\mathbf{x}) e^{-2\pi i \mathbf{s}\cdot\mathbf{x}}\, d\mathbf{x}, \tag{A.32}$$

[3] A measure μ has a density if and only if it is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^D$, i.e. every set that has Lebesgue measure zero also has μ-measure zero.

where $\tilde{f}(\mathbf{s})$ is called the *Fourier transform* of $f(\mathbf{x})$, see e.g. Bracewell [1986]. We refer to the equation on the left as the *synthesis* equation, and the equation on the right as the *analysis* equation. There are other conventions for Fourier transforms, particularly those involving $\boldsymbol{\omega} = 2\pi\mathbf{s}$. However, this tends to destroy symmetry between the analysis and synthesis equations so we use the definitions given above.

Here we have defined Fourier transforms for $f(\mathbf{x})$ being a function on $\mathbb{R}^D$. For related transforms for periodic functions, functions defined on the integer lattice and on the regular N-polygon see section B.1.

A.9 Convexity

Below we state some definitions and properties of convex sets and functions taken from Boyd and Vandenberghe [2004].

convex sets

A set C is convex if the line segment between any two points in C lies in C, i.e. if for any $x_1,\ x_2 \in C$ and for any θ with $0 \leq \theta \leq 1$, we have

$$\theta x_1 + (1-\theta)x_2 \in C. \tag{A.33}$$

convex function

A function $f : \mathcal{X} \to \mathbb{R}$ is *convex* if its domain $\mathcal{X}$ is a convex set and if for all $x_1,\ x_2 \in \mathcal{X}$ and θ with $0 \leq \theta \leq 1$, we have:

$$f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta)f(x_2), \tag{A.34}$$

where $\mathcal{X}$ is a (possibly improper) subset of $\mathbb{R}^D$. f is *concave* if $-f$ is convex.

A function f is convex if and only if its domain $\mathcal{X}$ is a convex set and its Hessian is positive semidefinite for all $x \in \mathcal{X}$.