Appendix B

Mathematics

This appendix reviews mathematical concepts that are used in the main text.

B.1 Functions

A *function* defines a mapping from a set $\mathcal{X}$ (e.g., the set of real numbers) to another set $\mathcal{Y}$. An *injection* is a one-to-one function where *every* element in the first set maps to a unique position in the second set (but there may be elements of the second set that are not mapped to). A *surjection* is a function where every element in the second set receives a mapping from the first (but there may be elements of the first set that are not mapped). A *bijection* or *bijective mapping* is a function that is both injective and surjective. It provides a one-to-one correspondence between all members of the two sets. A *diffeomorphism* is a special case of a bijection where both the forward and reverse mapping are differentiable.

B.1.1 Lipschitz constant

A function f[z] is *Lipschitz continuous* if for all z_1, z_2:

$$||\text{f}[z_1] - \text{f}[z_2]|| \leq \beta||z_1 - z_2||, \tag{B.1}$$

where β is known as the Lipschitz constant and determines the maximum gradient of the function (i.e., how fast the function can change) with respect to the distance metric. If the Lipschitz constant is less than one, the function is a contraction mapping, and we can use Banach's theorem to find the inverse for any point (see figure 16.9).

Composing two functions with Lipschitz constants β_1 and β_2 creates a new Lipschitz continuous function with a constant that is less than or equal to $\beta_1\beta_2$. Adding two functions with Lipschitz constants β_1 and β_2 creates a new Lipschitz continuous function with a constant that is less than or equal to $\beta_1+\beta_2$. The Lipschitz constant of a linear transformation $\mathbf{f}[\mathbf{z}] = \mathbf{A}\mathbf{z}+\mathbf{b}$ with respect to a Euclidean distance measure is the maximum eigenvalue of $\mathbf{A}$.

B.1.2 Convexity

A function is *convex* if we can draw a straight line between any two points on the function, and this line always lies above the function. Similarly, a function is *concave* if a straight line between any two points always lies below the function. By definition, convex (concave) functions have at most one minimum (maximum).

A region of $\mathbb{R}^D$ is convex if we can draw a straight line between any two points on the boundary of the region without intersecting the boundary in another place. Gradient descent guarantees to find the global minimum of any function that is both convex and defined on a convex region.

B.1.3 Special functions

The following functions are used in the main text:

- The *exponential function* $y = \exp[x]$ (figure B.1a) maps a real variable $x \in \mathbb{R}$ to a non-negative number $y \in \mathcal{R}^+$ as $y = e^x$.
- The *logarithm* $x = \log[y]$ (figure B.1b) is the inverse of the exponential function and maps a non-negative number $y \in \mathcal{R}^+$ to a real variable $x \in \mathbb{R}$. Note that all logarithms in this book are natural (i.e., in base e).
- The *gamma function* $\Gamma[x]$ (figure B.1c) is defined as:

$$\Gamma[x] = \int_0^\infty t^{x-1} e^{-t} dt. \tag{B.2}$$

  This extends the factorial function to continuous values so that $\Gamma[x] = (x-1)!$ for $x \in \{1, 2, \ldots\}$.
- The *Dirac delta function* $\delta[\mathbf{z}]$ has a total area of one, all of which is at position $\mathbf{z} = \mathbf{0}$. A dataset with N elements can be thought of as a probability distribution consisting of a sum of N delta functions centered at each data point $\mathbf{x}_i$ and scaled by $1/N$. The delta function is usually drawn as an arrow (e.g., figure 5.12). The delta function has the key property that:

$$\int \mathrm{f}[\mathbf{x}]\delta[\mathbf{x} - \mathbf{x}_0]d\mathbf{x} = \mathrm{f}[\mathbf{x}_0]. \tag{B.3}$$

B.1.4 Stirling's formula

Stirling's formula (figure B.2) approximates the factorial function (and hence the Gamma function) using the formula:

$$x! \approx \sqrt{2\pi x}\left(\frac{x}{e}\right)^x. \tag{B.4}$$

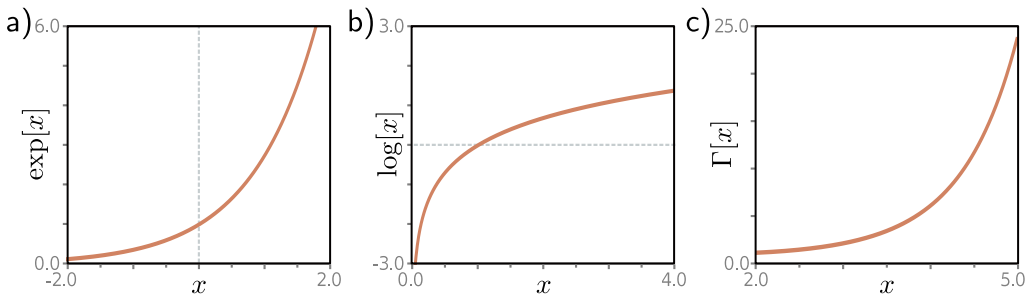

Figure B.1 Exponential, logarithm, and gamma functions. a) The exponential function maps a real number to a positive number. It is a convex function. b) The logarithm is the inverse of the exponential and maps a positive number to a real number. It is a concave function. c) The Gamma function is a continuous extension of the factorial function so that $\Gamma[x] = (x-1)!$ for $x \in \{1, 2, \ldots\}$.

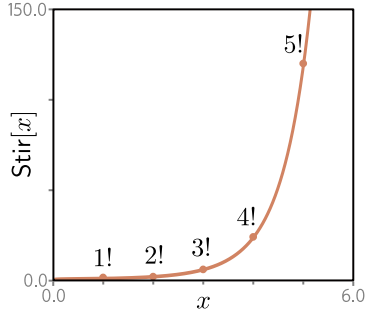

Figure B.2 Stirling's formula. The factorial function $x!$ can be approximated by Stirling's formula $\text{Stir}[x]$ which is defined for every real value.

B.2 Binomial coefficients

Binomial coefficients are written as $\binom{n}{k}$ and pronounced as "n choose k." They are positive integers that represent the number of ways of choosing an unordered subset of k items from a set of n items without replacement. Binomial coefficients can be computed using the simple formula:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}. \tag{B.5}$$

B.2.1 Autocorrelation

The autocorrelation $r[\tau]$ of a continuous function $\text{f}[z]$ is defined as:

$$r[\tau] = \int_{-\infty}^{\infty} \text{f}[t+\tau]\text{f}[t]dt, \tag{B.6}$$

where τ is the time lag. Sometimes, this is normalized by $r[0]$ so that the autocorrelation at time lag zero is one. The autocorrelation function is a measure of the correlation of the function with itself as a function of an offset (i.e., the time lag). If a function changes slowly and predictably, then the autocorrelation function will decrease slowly as the time lag increases from zero. If the function changes fast and unpredictably, then it will decrease quickly to zero.

B.3 Vector, matrices, and tensors

In machine learning, a vector $\mathbf{x} \in \mathbb{R}^D$ is a one-dimensional array of D numbers, which we will assume are organized in a column. Similarly, a matrix $\mathbf{Y} \in \mathbb{R}^{D_1 \times D_2}$ is a two-dimensional array of numbers with D_1 rows and D_2 columns. A tensor $\mathbf{z} \in \mathbb{R}^{D_1 \times D_2 \ldots \times D_N}$ is an N-dimensional array of numbers. Confusingly, all three of these quantities are stored in objects known as "tensors" in deep learning APIs such as PyTorch and TensorFlow.

B.3.1 Transpose

The transpose $\mathbf{A}^T \in \mathbb{R}^{D_2 \times D_1}$ of a matrix $\mathbf{A} \in \mathbb{R}^{D_1 \times D_2}$ is formed by reflecting it around the principal diagonal so that the k^{th} column becomes the k^{th} row and vice-versa. If we take the transpose of a matrix product $\mathbf{AB}$, then we take the transpose of the original matrices but reverse the order so that

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T. \tag{B.7}$$

The transpose of a column vector $\mathbf{a}$ is a row vector $\mathbf{a}^T$ and vice-versa.

B.3.2 Vector and matrix norms

For a vector $\mathbf{z}$, the ℓ_p norm is defined as:

$$||\mathbf{z}||_p = \left(\sum_{d=1}^{D} |z_d|^p\right)^{1/p}. \tag{B.8}$$

When $p = 2$, this returns the length of the vector, and this is known as the *Euclidean norm*. It is this case that is most commonly used in deep learning, and often the exponent p is omitted, and the Euclidean norm is just written as $||\mathbf{z}||$. When $p = \infty$, the operator returns the maximum absolute value in the vector.

Norms can be computed in a similar way for matrices. For example, the ℓ_2 norm of a matrix $\mathbf{Z}$ (known as the *Frobenius norm*) is calculated as:

$$||\mathbf{Z}||_F = \left(\sum_{i=1}^{I} \sum_{j=1}^{J} |z_{ij}|^2 \right)^{1/2}. \tag{B.9}$$

B.3.3 Product of matrices

The product $\mathbf{C} = \mathbf{AB}$ of two matrices $\mathbf{A} \in \mathbb{R}^{D_1 \times D_2}$ and $\mathbf{B} \in \mathbb{R}^{D_2 \times D_3}$ is a third matrix $\mathbf{C} \in \mathbb{R}^{D_1 \times D_3}$ where:

$$C_{ij} = \sum_{d=1}^{D_2} A_{id} B_{dj}. \tag{B.10}$$

B.3.4 Dot product of vectors

The dot product $\mathbf{a}^T\mathbf{b}$ of two vectors $\mathbf{a} \in \mathbb{R}^D$ and $\mathbf{b} \in \mathbb{R}^D$ is a scalar and is defined as:

$$\mathbf{a}^T\mathbf{b} = \mathbf{b}^T\mathbf{a} = \sum_{d=1}^{D} a_d b_d. \tag{B.11}$$

It can be shown that the dot product is proportional to the Euclidean norm of the first vector times the Euclidean norm of the second vector times the angle θ between them:

$$\mathbf{a}^T\mathbf{b} = ||\mathbf{a}||\,||\mathbf{b}||\cos[\theta]. \tag{B.12}$$

B.3.5 Inverse

A square matrix $\mathbf{A}$ may or may not have an inverse $\mathbf{A}^{-1}$ such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. If a matrix does not have an inverse, it is called *singular*. If we take the inverse of a matrix product $\mathbf{AB}$ then we can equivalently take the inverse of each matrix individually and reverse the order of multiplication.

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}. \tag{B.13}$$

In general, it takes $\mathcal{O}[D^3]$ operations to invert a $D \times D$ matrix. However, inversion is more efficient for special types of matrices, including diagonal, orthogonal, and triangular matrices (see section B.4).

B.3.6 Subspaces

Consider a matrix $\mathbf{A} \in \mathbb{R}^{D_1 \times D_2}$. If the number of columns D_2 of the matrix is fewer than the number of rows D_1 (i.e., the matrix is "portrait"), the product $\mathbf{Ax}$ cannot reach all

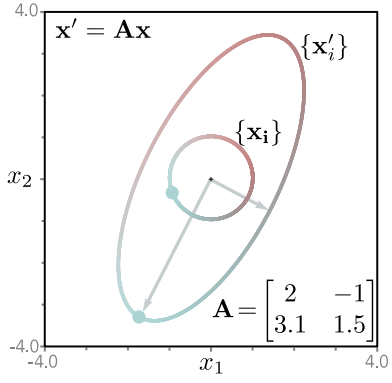

Figure B.3 Eigenvalues. When the points $\{\mathbf{x}_i\}$ on the unit circle are transformed to points $\{\mathbf{x}'_i\}$ by a linear transformation $\mathbf{x}'_i = \mathbf{A}\mathbf{x}_i$, they are mapped to an ellipse. For example, the light blue point on the unit circle is mapped to the light blue point on the ellipse. The length of the major (longest) axis of the ellipse (long gray arrow) is the magnitude of the first eigenvalue of the matrix, and the length of the minor (shortest) axis of the ellipse (short gray arrow) is the magnitude of the second eigenvalue.

possible positions in the D_1-dimensional output space. This product consists of the D_2 columns of $\mathbf{A}$ weighted by the D_2 elements of $\mathbf{x}$ and can only reach the *linear subspace* that is spanned by these columns. This is known as the *column space* of the matrix. Conversely, for a landscape matrix $\mathbf{A}$, the part of the input space that maps to zero (i.e., those $\mathbf{x}$ where $\mathbf{A}\mathbf{x} = \mathbf{0}$) is termed the *nullspace* of the matrix.

B.3.7 Eigenspectrum

If we multiply the set of 2D points on a unit circle by a 2×2 matrix $\mathbf{A}$, they map to an ellipse (figure B.3). The radii of the major and minor axes of this ellipse (i.e., the longest and shortest directions) correspond to the magnitude of the *eigenvalues* λ_1 and λ_2 of the matrix. The eigenvalues also have a sign, which relates to whether the matrix reflects the inputs about the origin. The same idea applies in higher dimensions. A $D-$dimensional spheroid is mapped by a $D \times D$ matrix $\mathbf{A}$ to a D-dimensional ellipsoid. The radii of the D principal axes of this ellipsoid determine the magnitude of the eigenvalues.

The *spectral norm* of a square matrix is the largest absolute eigenvalue. It captures the largest possible change in magnitude when the matrix is applied to a vector of unit length. As such, it tells us about the Lipschitz constant of the transformation. The set of eigenvalues is sometimes called the *eigenspectrum* and tells us about the magnitude of the scaling applied by the matrix across all directions. This information can be summarized using the *determinant* and *trace* of the matrix.

B.3.8 Determinant and trace

Every square matrix $\mathbf{A}$ has a scalar associated with it called the determinant and denoted by $|\mathbf{A}|$ or $\det[\mathbf{A}]$, which is the product of the eigenvalues. It is hence related to the average scaling applied by the matrix for different inputs. Matrices with small absolute determinants tend to decrease the norm of vectors upon multiplication. Matrices with large absolute determinants tend to increase the norm. If a matrix is *singular*, the determinant will be zero, and there will be at least one direction in space that is mapped

to the origin when the matrix is applied. Determinants of matrix expressions obey the following rules:

$$\begin{aligned} |\mathbf{A}^T| &= |\mathbf{A}| \\ |\mathbf{AB}| &= |\mathbf{A}||\mathbf{B}| \\ |\mathbf{A}^{-1}| &= 1/|\mathbf{A}|. \end{aligned} \tag{B.14}$$

The trace of a square matrix is the sum of the diagonal values (the matrix itself need not be diagonal) or the sum of the eigenvalues. Traces obey these rules:

$$\begin{aligned} \text{trace}[\mathbf{A}^\mathbf{T}] &= \text{trace}[\mathbf{A}] \\ \text{trace}[\mathbf{AB}] &= \text{trace}[\mathbf{BA}] \\ \text{trace}[\mathbf{A}+\mathbf{B}] &= \text{trace}[\mathbf{A}] + \text{trace}[\mathbf{B}] \\ \text{trace}[\mathbf{ABC}] &= \text{trace}[\mathbf{BCA}] = \text{trace}[\mathbf{CAB}], \end{aligned} \tag{B.15}$$

where in the last relation, the trace is invariant for cyclic permutations only, so in general, $\text{trace}[\mathbf{ABC}] \neq \text{trace}[\mathbf{BAC}]$.

B.4 Special types of matrix

Calculating the inverse of a square matrix $\mathbf{A} \in \mathbb{R}^{D\times D}$ has a complexity of $\mathcal{O}[D^3]$, as does the computation of the determinant. However, for some matrices with special properties, these computations can be more efficient.

B.4.1 Diagonal matrices

A *diagonal matrix* has zeros everywhere except on the principal diagonal. If these diagonal entries are all non-zero, the inverse is also a diagonal matrix, with each diagonal entry d_{ii} replaced by $1/d_{ii}$. The determinant is the product of the values on the diagonal. A special case of this is the *identity matrix*, which has ones on the diagonal. Consequently, its inverse is also the identity matrix, and its determinant is one.

B.4.2 Triangular matrices

A *lower triangular matrix* contains only non-zero values on the principal diagonal and the positions below this. An *upper* triangular matrix contains only non-zero values on the principal diagonal and the positions above this. In both cases, the matrix can be inverted in $\mathcal{O}[D^2]$ (see problem 16.4), and the determinant is just the product of the values on the diagonal.

B.4.3 Orthogonal matrices

Orthogonal matrices represent rotations and reflections around the origin, so in figure B.3, the circle would be mapped to another circle of unit radius but rotated and possibly reflected. Accordingly, the eigenvalues must all have magnitude one, and the determinant must be either one or minus one. The inverse of an orthogonal matrix is its transpose, so $\mathbf{A}^{-1} = \mathbf{A}^T$.

B.4.4 Permutation matrices

A permutation matrix has exactly one non-zero entry in each row and column, and all of these entries take the value one. It is a special case of an orthogonal matrix, so its inverse is its own transpose, and its determinant is always one. As the name suggests, it has the effect of permuting the entries of a vector. For example:

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} b \\ c \\ a \end{bmatrix}. \tag{B.16}$$

B.4.5 Linear algebra

Linear algebra is the mathematics of linear functions, which have the form:

$$\mathrm{f}[z_1, z_2, \ldots z_D] = \phi_1 z_1 + \phi_2 z_2 + \ldots \phi_D z_D, \tag{B.17}$$

where $\phi_1, \ldots, \phi_D$ are parameters that define the function. We often add a constant term ϕ_0 to the right-hand side. This is technically an *affine* function but is commonly referred to as linear in machine learning. We adopt this convention throughout.

B.4.6 Linear equations in matrix form

Consider a collection of linear functions:

$$\begin{aligned} y_1 &= \phi_{10} + \phi_{11} z_1 + \phi_{12} z_2 + \phi_{13} z_3 \\ y_2 &= \phi_{20} + \phi_{21} z_1 + \phi_{22} z_2 + \phi_{23} z_3 \\ y_3 &= \phi_{30} + \phi_{31} z_1 + \phi_{32} z_2 + \phi_{33} z_3. \end{aligned} \tag{B.18}$$

These can be written in matrix form as:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} \phi_{10} \\ \phi_{20} \\ \phi_{30} \end{bmatrix} + \begin{bmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{21} & \phi_{22} & \phi_{23} \\ \phi_{31} & \phi_{32} & \phi_{33} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}, \tag{B.19}$$

or as $\mathbf{y} = \boldsymbol{\phi}_0 + \boldsymbol{\Phi}\mathbf{z}$ for short, where $y_i = \phi_{i0} + \sum_{j=1}^{3} \phi_{ij} z_j$.

B.5 Matrix calculus

Most readers of this book will be accustomed to the idea that if we have a function $y = \text{f}[x]$, we can compute the derivative $\partial y/\partial x$, and this represents how y changes when we make a small change in x. This idea extends to functions $y = \text{f}[\mathbf{x}]$ mapping a vector $\mathbf{x}$ to a scalar y, functions $\mathbf{y} = \mathbf{f}[\mathbf{x}]$ mapping a vector $\mathbf{x}$ to a vector $\mathbf{y}$, functions $\mathbf{y} = \mathbf{f}[\mathbf{X}]$ mapping a matrix $\mathbf{X}$ to a vector $\mathbf{y}$, and so on. The rules of *matrix calculus* help us compute derivatives of these quantities. The derivatives take the following forms:

- For a function $y = \text{f}[\mathbf{x}]$ where $y \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^D$, the derivative $\partial y/\partial \mathbf{x}$ is also a D-dimensional vector, where the i^{th} element is computed as $\partial y/\partial x_i$.
- For a function $\mathbf{y} = \mathbf{f}[\mathbf{x}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{x} \in \mathbb{R}^{D_x}$, the derivative $\partial \mathbf{y}/\partial \mathbf{x}$ is a $D_x \times D_y$ matrix where element (i, j) contains the derivative $\partial y_j/\partial x_i$. This is known as a *Jacobian* and is sometimes written as $\nabla_{\mathbf{x}}\mathbf{y}$ in other documents.
- For a function $\mathbf{y} = \mathbf{f}[\mathbf{X}]$ where $\mathbf{y} \in \mathbb{R}^{D_y}$ and $\mathbf{X} \in \mathbb{R}^{D_1 \times D_2}$, the derivative $\partial \mathbf{y}/\partial \mathbf{X}$ is a 3D tensor containing the derivatives $\partial y_i/\partial x_{jk}$.

Often these matrix and vector derivatives have superficially similar forms to the scalar case. For example, we have:

$$y = ax \quad \longrightarrow \quad \frac{\partial y}{\partial x} = a, \tag{B.20}$$

and

$$\mathbf{y} = \mathbf{A}\mathbf{x} \quad \longrightarrow \quad \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}^T. \tag{B.21}$$