C Appendix C **Math Background**

I treat deep problems the same way I treat cold baths — get in and get out.
Friedrich Nietzsche

Given the finitude of this book, we cannot cover everything about everything; there are some prerequisites that will help you benefit from our exposition. We assume you've already taken courses on calculus, linear algebra, and differential equations; some familiarity with probability and statistics (e.g., that gained in a lab course) is also necessary. Here we provide the briefest of summaries on some essential topics, in order to refresh your memory; this is intended as reference (i.e., not pedagogical) material.

C.1 Taylor Series

We use a *Taylor series* to express a real function $f(x)$ about the point x_0 as an infinite sum:

$$f(x) = f(x_0)+(x-x_0)f'(x_0)+\frac{(x-x_0)^2}{2!}f''(x_0)+\frac{(x-x_0)^3}{3!}f'''(x_0)+\frac{(x-x_0)^4}{4!}f^{(4)}(x_0)+\cdots \tag{C.1}$$

If $x_0 = 0$, this is known as a *Maclaurin series*. Except for a few counter-examples in section 7.6, we will always assume that these derivatives exist. We typically express the Taylor series as a sum over a finite number of terms plus the *Lagrange remainder*:

$$\begin{aligned} f(x) = f(x_0) + (x-x_0)f'(x_0) + \frac{(x-x_0)^2}{2!}f''(x_0) + \frac{(x-x_0)^3}{3!}f'''(x_0) + \cdots \\ + \frac{(x-x_0)^{n-1}}{(n-1)!}f^{(n-1)}(x_0) + \frac{(x-x_0)^n}{n!}f^{(n)}(\xi) \end{aligned} \tag{C.2}$$

where ξ is some point between x_0 and x.

We will often encounter this series in different forms, e.g., expanding $f(x+h)$ around the point x: you should be able to apply Eq. (C.1) to such scenarios. To help you get oriented, let's see this Taylor series for the case where we only keep the first few terms:

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(x) + \frac{h^4}{24}f^{(4)}(\xi) \tag{C.3}$$

If we have no further information on $f^{(4)}(\xi)$, this can be compactly expressed by saying that the remainder is of fourth order, also known as $O(h^4)$. This is a way of listing only the most crucial dependence, dropping constants, prefactors, etc.

We will also encounter the case of a scalar function of many variables, i.e., $\phi(\mathbf{x})$, where

x bundles together the variables $x_0, x_1, \ldots, x_{n-1}$ but ϕ produces scalar values. We can then employ a multidimensional Taylor expansion, which takes the form:

$$\phi(\mathbf{x}+\mathbf{h}) = \phi(\mathbf{x}) + \sum_{j=0}^{n-1} \frac{\partial\phi}{\partial x_j} h_j + \frac{1}{2} \sum_{i,j=0}^{n-1} \frac{\partial\phi}{\partial x_i \partial x_j} h_i h_j + \cdots \tag{C.4}$$

where **h** bundles together the steps $h_0, h_1, \ldots, h_{n-1}$. This can be recast as:

$$\phi(\mathbf{x}+\mathbf{h}) = \phi(\mathbf{x}) + (\nabla\phi(\mathbf{x}))^T \mathbf{h} + \frac{1}{2}\mathbf{h}^T \mathbf{H}(\mathbf{x})\mathbf{h} + \cdots \tag{C.5}$$

where $\nabla\phi(\mathbf{x})$ is the *gradient vector* and $\mathbf{H}(\mathbf{x})$ is the *Hessian matrix*. This is more compact, but it assumes you know what the bold symbols mean, a topic we now turn to.

C.2 Matrix Terminology

Here we establish the matrix-related notation and terminology to be used in chapter 4. Crucially, in order to be consistent with Python's 0-indexing, we will employ notation that goes from 0 to $n-1$; this is different from most (all?) books on linear algebra; even so, having both the equations and the code employ the same notation seems to be an advantage, helping one avoid frustrating "off-by-one" errors.

Definitions A matrix is a rectangular array of numbers. We will be mainly dealing with *square matrices*, i.e., matrices of dimensions $n \times n$ (in fact, we will often refer to these simply as *matrices*). Here's a 3×3 example:

$$\mathbf{A} = \begin{pmatrix} A_{00} & A_{01} & A_{02} \\ A_{10} & A_{11} & A_{12} \\ A_{20} & A_{21} & A_{22} \end{pmatrix} \tag{C.6}$$

We denote the whole square matrix **A** (in bold) and the individual elements A_{ij} (where each of the indices goes from 0 to $n-1$). Note that the first index denotes the row and the second one the column. Since this is a 3×3 matrix, it has three rows and three columns.

You may encounter the notation $\left\{A_{ij}\right\}$, where a matrix element is surrounded by curly braces; here the braces mean that you should imagine i and j taking on all their possible values (and as a result, you get all possible matrix elements). In other words, this notation implies $\mathbf{A} = \left\{A_{ij}\right\}$.

We will also be using *column vectors*, which have a single column, e.g.:

$$\mathbf{x} = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \end{pmatrix} \tag{C.7}$$

Here we denote the whole column vector **x** (in bold) and the individual elements x_i. Viewed as a matrix, this has dimensions 3×1, i.e., three rows and one column. Similarly, one could define a *row vector*, which has a single row, for example:

$$\mathbf{y} = \begin{pmatrix} y_0 & y_1 & y_2 \end{pmatrix} \tag{C.8}$$

We denote the whole row vector **y** (in bold) and the individual elements y_i. This matrix has dimensions 1×3, i.e., one row and three columns.

Operations There are several mathematical operations one can carry out using matrices. For example, multiplication by a scalar can be expressed as either $\mathbf{B} = \kappa\mathbf{A}$ or $B_{ij} = \kappa A_{ij}$. The most interesting operations are *matrix-vector multiplication*:

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \qquad y_i = \sum_{j=0}^{n-1} A_{ij}x_j \tag{C.9}$$

and (two)-*matrix multiplication*:

$$\mathbf{C} = \mathbf{A}\mathbf{B}, \qquad C_{ij} = \sum_{k=0}^{n-1} A_{ik}B_{kj} \tag{C.10}$$

You could also take the *transpose* of a matrix:

$$\mathbf{B} = \mathbf{A}^T, \qquad B_{ij} = A_{ji} \tag{C.11}$$

A further definition: the *trace* of a square matrix **A** is the sum of the diagonal elements.

Finally, note that, if **x** is a 5×1 column vector, it is easier to display its transpose:

$$\mathbf{x}^T = \begin{pmatrix} x_0 & x_1 & x_2 & x_3 & x_4 \end{pmatrix} \tag{C.12}$$

since it fits on one line. Our lower-case bold symbols will always be column vectors, so we'll be using the transpose to save space on the page.

Special matrices There are several special matrices that you already know: diagonal matrices, the identity matrix, triangular matrices, symmetric matrices, real matrices, Hermitian matrices, and so on. Make sure you remember what these definitions mean; for example, symmetric means $\mathbf{A} = \mathbf{A}^T$ or $A_{ij} = A_{ji}$. In numerical linear algebra an important class consists of *sparse* matrices, for which most matrix elements are zero. Equally important are *tridiagonal* matrices, for which the only non-zero elements are on the main diagonal and on the two diagonals next to the main diagonal (more generally, *banded* matrices have non-zero elements on one or more diagonals).

Finally, a *diagonally dominant* matrix is one where each diagonal element is larger than or equal to (in absolute value) the sum of the magnitudes of all other elements on the same row. This might be easier to grasp by writing out an example:

$$\mathbf{A} = \begin{pmatrix} -3 & 2 & -7 \\ -9 & 1 & 6 \\ 1 & -5 & -2 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} -9 & 1 & 6 \\ 1 & -5 & -2 \\ -3 & 2 & -7 \end{pmatrix} \tag{C.13}$$

A is not diagonally dominant, but **B** is (though we simply re-arranged the rows).

Determinant For a square matrix **A**, one can evaluate a number known as the *determinant*, denoted by det(**A**) or |**A**|. It's easiest to start with the 2 × 2 case:

$$|\mathbf{A}| = \begin{vmatrix} A_{00} & A_{01} \\ A_{10} & A_{11} \end{vmatrix} = A_{00}A_{11} - A_{01}A_{10} \tag{C.14}$$

This is nothing other than a sum of products of matrix elements. For the $n \times n$ case:

$$|\mathbf{A}| = \sum_{i_0,i_1,\ldots,i_{n-1}=0}^{n-1} (-1)^k A_{0,i_0} A_{1,i_1} \ldots A_{(n-1),i_{n-1}} \tag{C.15}$$

This sum is over all the $n!$ permutations of degree n and k is the number of interchanges needed to put the i_j indices in the order $0, 1, 2, \ldots, n-1$. Qualitatively, we have a product of n elements with the appropriate sign and are dealing with $n!$ such products.

This is not the most efficient way of evaluating determinants. For a 10 × 10 matrix, we will need to sum 10! = 3 628 800 products, each of which involves nine multiplications (since you need nine multiplications to multiply 10 elements together). In total, this requires 9 × 3 628 800 = 32 659 200 multiplications and 3 628 799 additions/subtractions (since we were faced with 3 628 800 products to be added together). We would also need to keep track of the interchanges in order to make sure we are using the correct sign. In practice, one, instead, relies on the fact that the determinant of a triangular matrix is the product of the diagonal elements:

$$|\mathbf{A}| = \prod_{i=0}^{n-1} A_{ii} \tag{C.16}$$

In chapter 4 we see how to transform a matrix without changing the determinant; thus, we will be able to compute the determinant of a non-triangular matrix simply by evaluating the determinant of a "corresponding" triangular matrix.

Inverse Often, we can define the *inverse* of a matrix **A**, denoted by $\mathbf{A}^{-1}$, as follows:

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathcal{I} \tag{C.17}$$

where $\mathcal{I}$ is the identity matrix having the same dimensions as **A** (and $\mathbf{A}^{-1}$). Sometimes (rarely), we are not able to define the inverse: this happens when the determinant of the matrix is 0; we say the matrix is *singular*: such a matrix is made up of linearly dependent rows (or columns).

We are now in a position to introduce two more definitions of "special matrices". First, an *orthogonal matrix* has a transpose that is equal to the inverse: $\sum_{k=0}^{n-1} A_{ik}A_{jk} = \delta_{ij}$ or $\mathbf{A}^{-1} = \mathbf{A}^T$. Second, a *unitary matrix* has a Hermitian conjugate that is equal to the inverse: $\sum_{k=0}^{n-1} A_{ik}A^*_{jk} = \delta_{ij}$ or $\mathbf{A}^{-1} = \mathbf{A}^\dagger$. There are several other important cases (e.g., skew-symmetric: $\mathbf{A}^T = -\mathbf{A}$, skew-Hermitian: $\mathbf{A}^\dagger = -\mathbf{A}$), but what we have here is enough for our purposes.

C.3 Probability

In chapter 7 we discuss how to use random numbers in order to compute integrals. Here, we provide a brief summary of results from the theory of probability.

C.3.1 Discrete Random Variables

Consider a *discrete* random variable X: its possible values are x_i, each one appearing with the corresponding probability p_i^X. Observe that we are using an upper-case symbol for the random variable and a lower-case symbol for its possible values.

Mean and variance The *expectation* of this random variable (also known as the mean value or expected value) is simply:

$$\langle X \rangle = \sum_i p_i^X x_i \tag{C.18}$$

One can take the expected value of other quantities, for example the random variable X^2. This is called the second moment of X and is simply:

$$\langle X^2 \rangle = \sum_i p_i^X x_i^2 \tag{C.19}$$

This helps us calculate another useful quantity, known as the *variance*, var(X). The variance is the expectation of the random variable $(X - \langle X \rangle)^2$:

$$\text{var}(X) = \langle [X - \langle X \rangle]^2 \rangle \tag{C.20}$$

A simple calculation leads to an alternative expression for the variance:

$$\begin{aligned}\text{var}(X) &= \sum_i p_i^X (x_i - \langle X \rangle)^2 = \sum_i p_i^X x_i^2 - 2\sum_i p_i^X x_i \langle X \rangle + \sum_i p_i^X \langle X \rangle^2 \\ &= \langle X^2 \rangle - 2\langle X \rangle^2 + \langle X \rangle^2 = \langle X^2 \rangle - \langle X \rangle^2 \end{aligned} \tag{C.21}$$

In the first equality we applied the definition of an expected value. In the second equality we expanded out the square. In the third equality we identified a couple of expected values and used $\sum_i p_i^X = 1$. In the fourth equality we collected terms. This final result is very often the expression used to evaluate the variance. Another concept that is often used is that of the *standard deviation*; this is simply the square root of the variance, $\sqrt{\text{var}(X)}$.

Before concluding our discussion we note that, if X is a random variable, then $f(X)$ is also a random variable. That means that its expectation is:

$$\langle f(X) \rangle = \sum_i p_i^X f(x_i) \tag{C.22}$$

and, similarly, its variance is:

$$\text{var}[f(X)] = \langle [f(X) - \langle f(X) \rangle]^2 \rangle = \langle f^2(X) \rangle - \langle f(X) \rangle^2 \tag{C.23}$$

where the second step follows from a derivation analogous to that above.

Properties of the mean and variance We now turn to the problem of two random variables, X and Y. Specifically, we apply the definition of the expectaction to a linear combination of the two random variables:

$$\begin{aligned}\langle \lambda_1 X + \lambda_2 Y\rangle &= \sum_{i,j} p^{XY}_{ij}(\lambda_1 x_i + \lambda_2 y_j) = \lambda_1 \sum_{i,j} p^{XY}_{ij} x_i + \lambda_2 \sum_{i,j} p^{XY}_{ij} y_j \\ &= \lambda_1 \sum_i p^X_i x_i + \lambda_2 \sum_j p^Y_j y_j = \lambda_1 \langle X\rangle + \lambda_2 \langle Y\rangle \qquad \text{(C.24)}\end{aligned}$$

In the first equality we employed the *joint probabilities* p^{XY}_{ij}. In the third equality we evaluated sums like $\sum_j p^{XY}_{ij} = p^X_i$, leading to the *marginal* probabilities. Our result is known as the addition rule for expectations of random variables: the expectation of a linear combination of random variables is the same linear combination of the expectations of the random variables. This trivially generalizes to more random variables.

We now examine the variance of a linear combination of random variables:

$$\begin{aligned}\text{var}\,(\lambda_1 X + \lambda_2 Y) &= \langle(\lambda_1 X + \lambda_2 Y)^2\rangle - \langle \lambda_1 X + \lambda_2 Y\rangle^2 \\ &= \langle \lambda_1^2 X^2 + \lambda_2^2 Y^2 + 2\lambda_1\lambda_2 XY\rangle - (\lambda_1\langle X\rangle + \lambda_2\langle Y\rangle)^2 \\ &= \lambda_1^2\langle X^2\rangle + \lambda_2^2\langle Y^2\rangle + 2\lambda_1\lambda_2\langle XY\rangle - \lambda_1^2\langle X\rangle^2 - \lambda_2^2\langle Y\rangle^2 - 2\lambda_1\lambda_2\langle X\rangle\langle Y\rangle \\ &= \lambda_1^2\text{var}(X) + \lambda_2^2\text{var}(Y) + 2\lambda_1\lambda_2\,[\langle XY\rangle - \langle X\rangle\langle Y\rangle] \\ &= \lambda_1^2\text{var}(X) + \lambda_2^2\text{var}(Y) + 2\lambda_1\lambda_2\text{cov}(X,Y) \qquad \text{(C.25)}\end{aligned}$$

In the first line we applied the definition of the variance. In the second line we expanded out the square (in the first term) and used the addition rule that we established in the previous derivation (in the second term). In the third line we used the same addition rule again (in the first term) and expanded out the square (in the second term). In the fourth line we grouped terms (first and fourth, second and fifth, third and sixth, respectively). In the fifth line we introduced the *covariance*, $\text{cov}(X,Y) = \langle XY\rangle - \langle X\rangle\langle Y\rangle$, which measures the degree of independence of two random variables X and Y.

We now make our statement on the interpretation of the covariance more concrete. If X and Y are two independent random variables, then $p^{XY}_{ij} = p^X_i p^Y_j$, so the $\langle XY\rangle$ that appears on the right-hand side in the definition of the covariance can be calculated as follows:

$$\langle XY\rangle = \sum_{i,j} p^{XY}_{ij} x_i y_j = \sum_i p^X_i x_i \sum_j p^Y_j y_j = \langle X\rangle\langle Y\rangle \qquad \text{(C.26)}$$

implying that *the covariance vanishes for two independent random variables*.[1]

We now specialize the main result in our second derivation, Eq. (C.25), to the case of two independent random variables (and therefore vanishing covariance), to find:

$$\text{var}\,(\lambda_1 X + \lambda_2 Y) = \lambda_1^2\text{var}(X) + \lambda_2^2\text{var}(Y) \qquad \text{(C.27)}$$

In words, we find that for independent random variables the variance of a linear combination of the variables is a new linear combination of the variances of the random variables we started with: the coefficients on the right-hand side turn out to be squared. We will

[1] The reverse is not true: you can have two random variables that are not independent but give zero covariance.

refer to this as the addition rule for the variances of random variables. This, too, trivially generalizes to more random variables.

C.3.2 Continuous Random Variables

For the case of a *continuous* random variable X we are faced with a *probability density function*, $p(x)$, which plays the role p_i^X played in the discrete case. We typically assume that the probability density function is normalized, i.e.:

$$\int_{-\infty}^{+\infty} p(x)dx = 1 \tag{C.28}$$

holds. This time around, the definition of the *expectation* is:

$$\langle X\rangle = \int_{-\infty}^{+\infty} xp(x)dx \tag{C.29}$$

Similarly, we have for the *variance*:

$$\begin{aligned}\mathrm{var}(X) = \langle [X - \langle X\rangle]^2\rangle &= \int_{-\infty}^{+\infty} (x - \langle X\rangle)^2 p(x)dx \\ = \langle X^2\rangle - \langle X\rangle^2 &= \int_{-\infty}^{+\infty} x^2 p(x)dx - \left(\int_{-\infty}^{+\infty} xp(x)dx\right)^2\end{aligned} \tag{C.30}$$

where the second line follows from a derivation that is completely analogous to that in the discrete-variable case.

Just like in the previous subsection, if X is a random variable, then $f(X)$ is also a random variable; its expectation is:

$$\langle f(X)\rangle = \int_{-\infty}^{+\infty} f(x)p(x)dx \tag{C.31}$$

and, similarly, its variance is:

$$\mathrm{var}[f(X)] = \langle [f(X) - \langle f(X)\rangle]^2\rangle = \langle f^2(X)\rangle - \langle f(X)\rangle^2 \tag{C.32}$$

Note that both steps are identical to Eq. (C.23) but, of course, the meaning of the expectation is now different.

In complete analogy to the discrete-variable case, we can take linear combinations of random variables. For the expectation we find:

$$\langle \lambda_1 X + \lambda_2 Y\rangle = \lambda_1\langle X\rangle + \lambda_2\langle Y\rangle \tag{C.33}$$

and for the variance, for independent random variables, we get:

$$\mathrm{var}\,(\lambda_1 X + \lambda_2 Y) = \lambda_1^2 \mathrm{var}(X) + \lambda_2^2 \mathrm{var}(Y) \tag{C.34}$$

Observe that these two results are identical to Eq. (C.24) and to Eq. (C.27), respectively, but once again the expectation should now be interpreted as an integral instead of a sum. In other words, the addition rule for expectations or variances of random variables is the same, regardless of whether the variables are discrete or continuous.