Appendix A

Mathematics notation and review

This appendix gives brief coverage of the mathematical notation and concepts that you'll encounter in this book. In the space of a few pages it is of course impossible to do justice to topics such as integration and matrix algebra. Readers interested in strengthening their fundamentals in these areas are encouraged to consult XXX [calculus] and Healy (2000).

A.1 Sets ($\{\}, \cup, \cap, \emptyset$)

The notation $\{a, b, c\}$ should be read as "the set containing the elements a, b, and c". With sets, it's sometimes a convention that lower-case letters are used as names for elements, and upper-case letters as names for sets, though this is a weak convention (after all, sets can contain anything—even other sets!).

$A \cup B$ is read as "the union of A and B", and its value is the set containing exactly those elements that are present in A, in B, or in both.

$A \cap B$ is read as "the intersection of A and B", and its value is the set containing only those elements present in both A and B.

$\emptyset$, or equivalently $\{\}$, denotes the empty set—the set containing nothing. Note that $\{\emptyset\}$ isn't the empty set—it's the set containing only the empty set, and since it contains something, it isn't empty!

[introduce set complementation if necessary]

A.1.1 Countability of sets

[briefly describe]

A.2 Summation ($\sum$)

Many times we'll want to express a complex sum of systematically related parts, such as $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5}$ or $x_1 + x_2 + x_3 + x_4 + x_5$, more compactly. We use SUMMATION notation for this:

$$\sum_{i=1}^{5} \frac{1}{i} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} \qquad \sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

In these cases, i is sometimes called an INDEX VARIABLE, linking the RANGE of the sum (1 to 5 in both of these cases) to its contents. Sums can be nested:

$$\sum_{i=1}^{2}\sum_{j=1}^{2} x_{ij} = x_{11} + x_{12} + x_{21} + x_{22} \qquad \sum_{i=1}^{3}\sum_{j=1}^{i} x_{ij} = x_{11} + x_{21} + x_{22} + x_{31} + x_{32} + x_{33}$$

Sums can also be infinite:

$$\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \dots$$

Frequently, the range of the sum can be understood from context, and will be left out; or we want to be vague about the precise range of the sum. For example, suppose that there are n variables, x_1 through x_n. In order to say that the sum of all n variables is equal to 1, we might simply write

$$\sum_{i} x_i = 1$$

A.3 Product of a sequence ($\prod$)

Just as we often want to express a complex sum of systematically related parts, we often want to express a product of systematically related parts as well. We use PRODUCT notation to do this:

$$\prod_{i=1}^{5} \frac{1}{i} = 1 \times \frac{1}{2} \times \frac{1}{3} \times \frac{1}{4} \times \frac{1}{5} \qquad \prod_{i=1}^{5} x_i = x_1 x_2 x_3 x_4 x_5$$

Usage of product notation is completely analogous to summation notation as described in Section A.2.

A.4 "Cases" notation ($\{$)

Some types of equations, especially those describing probability functions, are often best expressed in the form of one or more conditional statements. As an example, consider a six-sided die that is weighted such that when it is rolled, 50% of the time the outcome is a six, with the other five outcomes all being equally likely (i.e. 10% each). If we define a discrete random variable X representing the outcome of a roll of this die, then the clearest way of specifying the probability mass function for X is by splitting up the real numbers into three groups, such that all numbers in a given group are equally probable: (a) 6 has probability 0.5; (b) 1, 2, 3, 4, and 5 each have probability 0.1; (c) all other numbers have probability zero. Groupings of this type are often expressed using "cases" notation in an equation, with each of the cases expressed on a different row:

$$P(X = x) = \begin{cases} 0.5 & x = 6 \\ 0.1 & x \in \{1, 2, 3, 4, 5\} \\ 0 & \text{otherwise} \end{cases}$$

A.5 Logarithms and exponents

The log in base b of a number x is expressed as $log_b x$; when no base is given, as in $logx$, the base should be assumed to be the mathematical constant e. The expression $exp[x]$ is equivalent to the expression e^x. Among other things, logarithms are useful in probability theory because they allow one to translate between sums and products: $\sum_i \log x_i = \log \prod_i x_i$. Derivatives of logarithmic and exponential functions are as follows:

$$\frac{d}{dx} \log_b x = \frac{1}{x \log b}$$

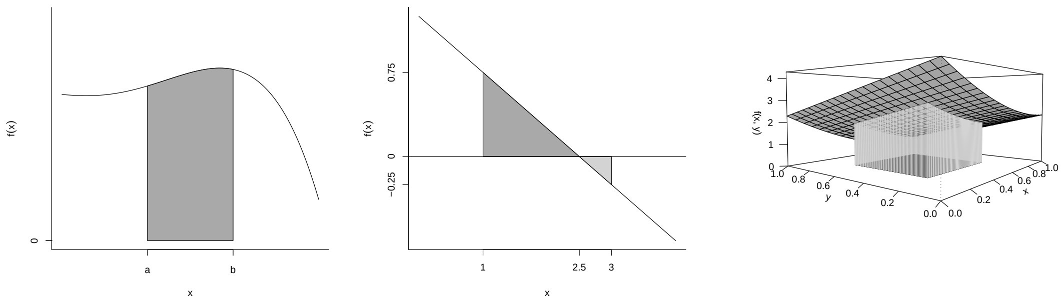
$$\frac{d}{dx} y^x = y^x \log y$$

A.6 Integration ($\int$)

Sums are always over countable (finite or countably infinite) sets. The analogue over a continuum is INTEGRATION. Correspondingly, you need to know a bit about integration in order to understand continuous random variables. In particular, a basic grasp of integration is essential to understanding how Bayesian statistical inference works.

One simple view of integration is as computing "area under the curve". In the case of integrating a function f over some range $[a, b]$ of a one-dimensional variable x in which $f(x) > 0$, this view is literally correct. Imagine plotting the curve $f(x)$ against x, extending straight lines from points a and b on the x-axis up to the curve, and then laying the plot down on a table. The area on the table enclosed on four sides by the curve, the x-axis, and the two additional straight lines is the integral

$$\int_a^b f(x)\,\mathrm{d}x$$



(a) (b) (c)

Figure A.1: Integration

This is depicted graphically in Figure A.1a.

The situation is perhaps slightly less intuitive, but really no more complicated, when $f(x)$ crosses the x-axis. In this case, area under the x-axis counts as "negative" area. An example is given in Figure A.1b; the function here is $f(x) = \frac{1}{2}(2.5 - x)$. Since the area of a triangle with height h and length l is $\frac{lh}{2}$, we can compute the integral in this case by subtracting the area of the smaller triangle from the larger triangle:

$$\int_1^3 f(x)\,\mathrm{d}x = \frac{1.5 \times 0.75}{2} - \frac{0.5 \times 0.25}{2} = 0.5$$

Integration also generalizes to multiple dimensions. For instance, the integral of a function f over an area in two dimensions x and y, where $f(x, y) > 0$, can be thought of as the volume enclosed by projecting the area's boundary from the x, y plane up to the $f(x, y)$ surface. A specific example is depicted in Figure A.1c, where the area in this case is the square bounded by 1/4 and 3/4 in both the x and y directions.

$$\int_{\frac{1}{4}}^{\frac{3}{4}} \int_{\frac{1}{4}}^{\frac{3}{4}} f(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

An integral can also be over the *entire* range of a variable or set of variables. For instance, one would write an integral over the entire range of x as $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x$. Finally, in this book and in the literature on probabilistic inference you will see the abbreviated notation $\int_\theta f(\theta)\,\mathrm{d}\theta$, where θ is typically an ensemble (collection) of variables. In this book, the proper interpretation of this notation is as the integral over the entire range of all variables in the ensemble θ.

A.6.1 Analytic integration tricks

Computing an integral ANALYTICALLY means finding an exact form for the value of the integral. There are entire books devoted to analytic integration, but for the contents of this book you'll get pretty far with just a few tricks.

1. **Multiplication by constants.** The integral of a function times a constant C is the product of the constant and the integral of the function:

$$\int_a^b Cf(x)\,\mathrm{d}x = C\int_a^b f(x)\,\mathrm{d}x$$

2. **Sum rule.** The integral of a sum is the sum of the integrals of the parts:

$$\int_a^b [f(x)+g(x)]\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x + \int_a^b g(x)\,\mathrm{d}x$$

3. **Expressing one integral as the difference between two other integrals:** For $c < a, b$,

$$\int_a^b f(x)\,\mathrm{d}x = \int_c^b f(x)\,\mathrm{d}x - \int_c^a f(x)\,\mathrm{d}x$$

This is an extremely important technique when asking whether the outcome of a continuous random variable falls within a range [a,b], because it allows you to answer this question in terms of cumulative distribution functions (Section 2.6); in these cases you'll choose $c = -\infty$.

4. **Polynomials.** For any $n \neq -1$:

$$\int_a^b x^n\,\mathrm{d}x = \frac{1}{n+1}(b^{n+1} - a^{n+1})$$

And the special case for $n = -1$ is:

$$\int_a^b x^{-1}\,\mathrm{d}x = \log b - \log a$$

Note that this generalization holds for $n = 0$, so that integration of a constant is easy:

$$\int_a^b C\,\mathrm{d}x = C(b-a)$$

5. **Normalizing constants.** If the function inside an integral looks the same as the probability density function for a known probability distribution, then its value is related to normalizing constant of the probability distribution. [Examples: normal distribution; beta distribution; others?] For example, consider the integral

$$\int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{18}\right] \mathrm{d}x$$

This may look hopelessly complicated, but by comparison with Equation 2.21 in Section 2.10 you will see that it looks just like the probability density function of a normally distributed random variable with mean $\mu = 0$ and variance $\sigma^2 = 9$, except that it doesn't have the normalizing constant $\frac{1}{\sqrt{2\pi\sigma^2}}$. In order to determine the value of this integral, we can start by noting that any probability density function integrates to 1:

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \mathrm{d}x = 1$$

Substituting in $\mu = 0, \sigma^2 = 9$ we get

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{18\pi}} \exp\left[-\frac{x^2}{18}\right] \mathrm{d}x = 1$$

By the rule of multiplication by constants we get

$$\frac{1}{\sqrt{18\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{18}\right] \mathrm{d}x = 1$$

or equivalently

$$\int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{18}\right] \mathrm{d}x = \sqrt{18\pi}$$

giving us the solution to the original problem.

A.6.2 Numeric integration

The alternative to analytic integration is NUMERIC integration, which means approximating the value of an integral by explicit numeric computation. There are many ways to do this—one common way is by breaking up the range of integration into many small pieces, approximating the size of each piece, and summing the approximate sizes. A graphical

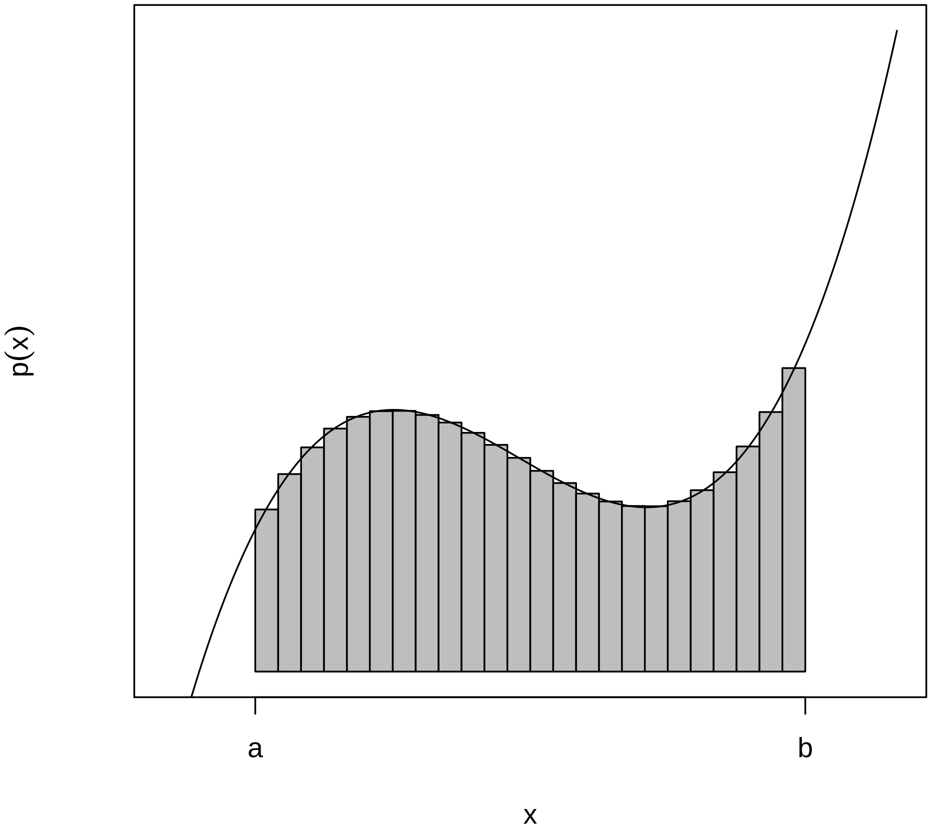


Figure A.2: Numeric integration

example of how this might be done is shown in Figure A.2, where each piece of the area under the curve is approximated as a rectangle whose height is the average of the distances from the x-axis to the curve at the left and right edges of the rectangle. There are many techniques for numeric integration, and we shall have occasional use for some of them in this book.

A.7 Precedence ($\prec$)

The $\prec$ operator is used occasionally in this book to denote LINEAR PRECEDENCE. In the syntax of English, for example, the information that "a verb phrase (VP) can consist of a verb (V) followed by a noun phrase (NP) object" is most often written as:

$$\text{V} \to \text{V NP}$$

This statement combines two pieces of information: (1) a VP can be comprised of a V and an NP; and (2) in the VP, the V should precede the NP. In a syntactic tradition stemming from Generalized Phrase Structure Grammar (Gazdar et al., 1985), these pieces of information can be separated:

$$(1)\text{V} \to \text{V, NP} \qquad\qquad (2)\text{V} \prec \text{NP}$$

where V,NP means "the unordered set of categories V and NP", and V $\prec$ NP reads as "V precedes NP".

A.8 Combinatorics ($\binom{n}{r}$)

The notation $\binom{n}{r}$ is read as "n choose r" and is defined as the number of possible ways of selecting r elements from a larger collection of n elements, allowing each element to be chosen a maximum of once and ignoring order of selection. The following equality holds generally:

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} \tag{A.1}$$

The solution to the closely related problem of creating m classes from n elements by selecting r_i for the i-th class and discarding the leftover elements is written as $\binom{n}{r_1 \dots r_m}$ and its value is

$$\binom{n}{r_1 \dots r_m} = \frac{n!}{r_1! \dots r_m!} \tag{A.2}$$

Terms of this form appear in this book in the binomial and multinomial probability mass functions, and as normalizing constant for the beta and Dirichlet distributions.

A.9 Basic matrix algebra

There are a number of situations in probabilistic modeling—many of which are covered in this book—where the computations needing to be performed can be simplified, both conceptually and notationally, by casting them in terms of MATRIX operations. A matrix $\boldsymbol{X}$ of dimensions $m \times n$ is a set of mn entries arranged rectangularly into m rows and n columns, with its entries indexed as x_{ij}:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{21} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m1} & \dots & x_{mn} \end{bmatrix}$$

For example, the matrix $\boldsymbol{A} = \begin{bmatrix} 3 & 4 & -1 \\ 0 & -2 & 2 \end{bmatrix}$ has values $a_{11} = 3$, $a_{12} = 4$, $a_{13} = -1$, $a_{21} = 0$, $a_{22} = -2$, and $a_{23} = 2$. For a matrix $\boldsymbol{X}$, the entry x_{ij} is often called the i, j-th entry of $\boldsymbol{X}$.

If a matrix has the same number of rows and columns, it is often called a SQUARE matrix. Square matrices are often divided into the DIAGONAL entries $\{x_{ii}\}$ and the OFF-DIAGONAL entries $\{x_{ij}\}$ where $i \neq j$. A matrix of dimension $m \times 1$—that is, a single-column matrix—is often called a VECTOR.

Symmetric matrices: a square matrix $\boldsymbol{A}$ is SYMMETRIC if $\boldsymbol{A}^T = \boldsymbol{A}$. For example, the matrix

$$\begin{bmatrix} 10 & -1 & 4 \\ -1 & 3 & 2 \\ 4 & 2 & 5 \end{bmatrix}$$

is symmetric. You will generally encounter symmetric matrices in this book as variance-covariance matrices (e.g., of the multivariate normal distribution, Section 3.5). Note that a symmetric $n \times n$ matrix has $\frac{n(n+1)}{2}$ "free" entries—one can choose the entries on and above the diagonal, but the entries below the diagonal are fully determined by the entries above it.

Diagonal and Identity matrices: For a square matrix $\boldsymbol{X}$, the entries x_{ii}—that is, when the column and row numbers are the same—are called the DIAGONAL entries. A square matrix whose non-diagonal entries are all zero is called a DIAGONAL MATRIX. A diagonal matrix of size $n \times n$ whose diagonal entries are all 1 is called the size-n IDENTITY matrix. Hence $\boldsymbol{A}$ below is a diagonal matrix, and $\boldsymbol{B}$ below is the size-3 identity matrix.

$$\boldsymbol{A} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The $n \times n$ identity matrix is sometimes notated as $\boldsymbol{I}_n$; when the dimension is clear from context, sometimes the simpler notation $\boldsymbol{I}$ is used.

Transposition: For any matrix $\boldsymbol{X}$ of dimension $m \times n$, the TRANSPOSE of $\boldsymbol{X}$, or $\boldsymbol{X}^{\mathrm{T}}$, is an $n \times m$-dimensional matrix such that the i, j-th entry of $\boldsymbol{X}^{\mathrm{T}}$ is the j, i-th entry of $\boldsymbol{X}$. For the matrix $\boldsymbol{A}$ above, for example, we have

$$\boldsymbol{A}^{\mathrm{T}} = \begin{bmatrix} 3 & 0 \\ 4 & -2 \\ -1 & 2 \end{bmatrix} \tag{A.3}$$

Addition: Matrices of like dimension can be added. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are both $m \times n$ matrices, then $\boldsymbol{X} + \boldsymbol{Y}$ is the $m \times n$ matrix whose i, j-th entry is $x_{ij} + y_{ij}$. For example,

$$\begin{bmatrix} 3 & 0 \\ 4 & -2 \\ -1 & 2 \end{bmatrix} + \begin{bmatrix} -1 & -1 \\ 0 & 2 \\ 5 & 5 \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ 4 & 0 \\ 4 & 7 \end{bmatrix} \tag{A.4}$$

Multiplication: If $\boldsymbol{X}$ is an $l \times m$ matrix and $\boldsymbol{Y}$ is an $m \times n$ matrix, then $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be multiplied together; the resulting matrix $\boldsymbol{XY}$ is an $l \times m$ matrix. If $\boldsymbol{Z} = \boldsymbol{XY}$, the i, j-th entry of $\boldsymbol{Z}$ is:

$$z_{ij} = \sum_{k=1}^{m} x_{ik} y_{kj}$$

For example, if $\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ -1 & 0 \\ 3 & 1 \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} 3 & 4 & -1 & 6 \\ 0 & -5 & 2 & -2 \end{bmatrix}$, we have

$$\begin{aligned} \boldsymbol{AB} &= \begin{bmatrix} 1 \times 3 + 2 \times 0 & 1 \times 4 + 2 \times (-5) & 1 \times (-1) + 2 \times 2 & 1 \times 6 + 2 \times (-2) \\ (-1) \times 3 + 0 \times 0 & (-1) \times 4 + 0 \times (-5) & (-1) \times (-1) + 0 \times 2 & (-1) \times 6 + 0 \times (-2) \\ 3 \times 3 + 1 \times 0 & 3 \times 4 + 1 \times (-5) & 3 \times (-1) + 1 \times 2 & 3 \times 6 + 1 \times (-2) \end{bmatrix} \\ &= \begin{bmatrix} 3 & -6 & 3 & 2 \\ -3 & -4 & 1 & -6 \\ 9 & 7 & -1 & 16 \end{bmatrix} \end{aligned}$$

Unlike multiplication of scalars, matrix multiplication is **not** commutative—that is, it is not generally the case that $\boldsymbol{XY} = \boldsymbol{YX}$. In fact, being able to form the matrix product $\boldsymbol{XY}$ does not even guarantee that we can do the multiplication in the opposite order and form the matrix product $\boldsymbol{YX}$; the dimensions may not be right. (Such is the case for matrices A and B in our example.)

Determinants. For a square matrix $\boldsymbol{X}$, the DETERMINANT $|\boldsymbol{X}|$ is a measure of the matrix's "size". In this book, determinants appear in coverage of the multivariate normal distribution (Section 3.5); the normalizing constant of the multivariate normal density includes the determinant of the covariance matrix. (The univariate normal density, introduced in Section 2.10, is a special case; there, it is simply the variance of the distribution that appears in the normalizing constant.) For small matrices, there are simple techniques for calculating determinants: as an example, the determinant of a 2×2 matrix $\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is $|\boldsymbol{A}| = ad - bc$. For larger matrices, computing determinants requires more general and complex techniques, which can be found in books on linear algebra such as Healy (2000).

Matrix Inversion. The INVERSE or RECIPROCAL of an $n \times n$ square matrix $\boldsymbol{X}$, denoted $\boldsymbol{X}^{-1}$, is the $n \times n$ matrix such that $\boldsymbol{X}\boldsymbol{X}^{-1} = \boldsymbol{I}_n$. As with scalars, the inverse of the inverse of a matrix $\boldsymbol{X}$ is simply $\boldsymbol{X}$. However, not all matrices have inverses (just like the scalar 0 has no inverse).

For example, the following pair of matrices are inverses of each other:

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \qquad \boldsymbol{A}^{-1} = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

A.9.1 Algebraic properties of matrix operations

Associativity, Commutativity, and Distributivity

Consider matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$. Matrix multiplication is associative ($\boldsymbol{A}(\boldsymbol{B}\boldsymbol{C}) = (\boldsymbol{A}\boldsymbol{B})\boldsymbol{C}$) and distributive over addition ($\boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C}) = (\boldsymbol{A}+\boldsymbol{B})\boldsymbol{C}$), but not commutative: even if the multiplication is possible in both orderings (that is, if $\boldsymbol{B}$ and $\boldsymbol{A}$ are both square matrices with the same dimensions), in general $\boldsymbol{A}\boldsymbol{B} \neq \boldsymbol{B}\boldsymbol{A}$.

Transposition, inversion and determinants of matrix products.

- The transpose of a matrix product is the product of each matrix's transpose, in reverse order: $(\boldsymbol{A}\boldsymbol{B})^T = \boldsymbol{B}^T\boldsymbol{A}^T$.
- Likewise, the inverse of a matrix product is the product of each matrix's inverse, in reverse order: $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.
- The determinant of a matrix product is the product of the determinants:

$$|\boldsymbol{A}\boldsymbol{B}| = |\boldsymbol{A}|\,|\boldsymbol{B}|$$

- Because of this, the determinant of the inverse of a matrix is the reciprocal of the matrix's determinant:

$$\left|\boldsymbol{A}^{-1}\right| = \frac{1}{|\boldsymbol{A}|}$$

A.10 Miscellaneous notation

- $\propto$: You'll often see $f(x) \propto g(x)$ for some functions f and g of x. This is to be read as "$f(x)$ is proportional to $g(x)$", or "$f(x)$ is equal to $g(x)$ to within some constant". Typically it's used when $f(x)$ is intended to be a probability, and $g(x)$ is a function that obeys the first two axioms of probability theory, but is improper. This situation obtains quite often when, for example, conducting Bayesian inference.