COMS20011 Mathematical preliminaries

Laurence Aitchison

March 2022

Here, I'm going to introduce all the mathematical preliminaries and notation that I'm going to use, with a view to their later use in data science. You may well have seen some stuff here before, but everything here is examinable, so please do at least skim through. And please do the exercises: they're great preparation for later lectures.

Some terminology I'll use:

- "Unit" exercise: test a small part of the material. Ranges from easier-than to about the same as a typical exam question. (Remember though that exam questions themselves have a range of difficulties).
- "Integration" exercise: test your ability to combine different bits of lecture material to solve questions. Ranges from about the same to quite a bit harder than a typical exam question.
- "FYI" (for your interest). Non-examinable note, generally around a technical issue that might occur to more mathematical students isn't necessary for data science.

1 Calculus

In data science, we often end up model fitting: e.g. we find the straight line that is closest to the data. For just a single straight line, you can make a pretty good guess based on a plot. But if you're working with lots of data, you can't just guess. It turns out that you can do *much* better than just guessing, but to do so, you need calculus!

1.1 Polynomials

The derivative of x^p, where p is some power,

$$\frac{\partial x^p}{\partial x} = p x^{p-1}. \tag{1}$$

Perhaps the most important example (which we're going to encounter many times) is the derivative of a quadratic,

$$\frac{\partial x^2}{\partial x} = 2x. \tag{2}$$

But the formula also applies for higher powers,

$$\frac{\partial x^5}{\partial x} = 5x^4. \tag{3}$$

And for negative powers,

$$\frac{\partial x^{-3}}{\partial x} = -3x^{-4}. \tag{4}$$

And for fractional powers,

$$\frac{\partial \sqrt{x}}{\partial x} = \frac{\partial x^{1/2}}{\partial x} = \frac{1}{2}x^{-1/2} = \frac{1}{2\sqrt{x}}. \tag{5}$$

And for powers of zero,

$$\frac{\partial 1}{\partial x} = \frac{\partial x^0}{\partial x} = 0x^{-1} = 0. \tag{6}$$

(note that $x^0 = 1$, a constant, so the gradient has to be zero). And for powers of one,

$$\frac{\partial x}{\partial x} = \frac{\partial x^1}{\partial x} = 1x^0 = 1. \tag{7}$$

(note that $x^1 = x$, which has a slope of 1). We can apply the rule to each term in a polynomial,

$$\frac{\partial}{\partial x}[3x^4 + 2x^{-1/2} + x^{-2}] = 3\frac{\partial x^4}{\partial x} + 2\frac{\partial x^{-1/2}}{\partial x} + \frac{\partial x^{-2}}{\partial x} \tag{8}$$

Looking at each term separately,

$$\frac{\partial x^4}{\partial x} = 4x^3 \tag{9}$$

$$\frac{\partial x^{-1/2}}{\partial x} = -\tfrac{1}{2}x^{-1.5} \tag{10}$$

$$\frac{\partial x^{-2}}{\partial x} = -2x^{-3} \tag{11}$$

Putting everything back together,

$$\frac{\partial}{\partial x}[3x^4 + 2x^{-1/2} + x^{-2}] = 3(4x^3) + 2(-\tfrac{1}{2}x^{1.5}) - 2x^{-3} \tag{12}$$

$$= 12x^3 - x^{1.5} - 2x^{-3}. \tag{13}$$

1.2 Chain rule

To differentiate more complex expressions, we need the chain rule. For instance, we might have,

$$\frac{\partial y}{\partial x} \qquad \text{where} \qquad y = (x+1)^3 \tag{14}$$

We could expand the brackets, but we don't want to because that would be alot of terms. Instead, we rewrite y in terms of u,

$$u = x + 1 \qquad\qquad y = u^3 \tag{15}$$

Then, we use the chain rule,

$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial u}\frac{\partial u}{\partial x}. \tag{16}$$

And each of these derivatives is much easier,

$$\frac{\partial y}{\partial u} = \frac{\partial u^3}{\partial u} = 3u^2 \tag{17}$$

$$\frac{\partial u}{\partial x} = \frac{\partial x+1}{\partial x} = \frac{\partial x}{\partial x} + \frac{\partial 1}{\partial x} = 1 + 0 = 1 \tag{18}$$

Subsituting these derivatives into the chain rule, we get our answer,

$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial u}\frac{\partial u}{\partial x} = 3u^2 \times 1 \tag{19}$$

And finally substituting $u = x + 1$,

$$\frac{\partial y}{\partial x} = 3(x+1)^2. \tag{20}$$

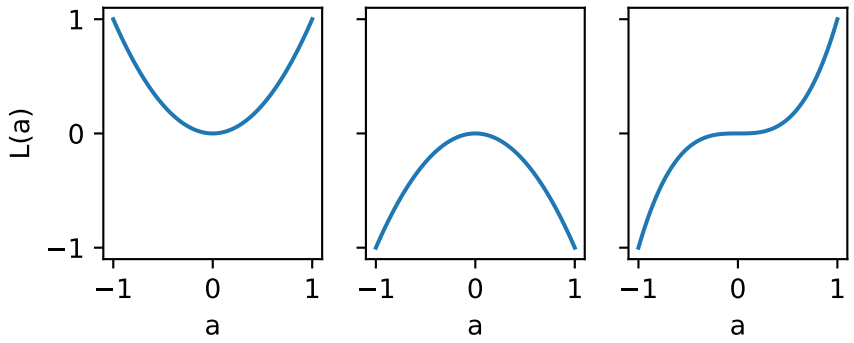
FYI: Despite its simplicity in this example, it turns out that the chain rule is the basis of backprop, and backprop is the basis of all modern AI, from ChatGPT to Stable Diffusion to just about everything else. This is because the chain rule allows you to "chain" together derivatives in a complex, multi-step pipeline. But that's for future courses...

1.3 Using calculus to do optimization

At the top, we mentioned that we were going to use calculus to find e.g. the best fitting model (e.g. straight line) to some data. The starting point is to find a function that measures how good our model is at fitting the data. Typically, our model has some tunable parameters (e.g. the slope of a straight line, which we're going to call a). Then, we have a "loss" function $\mathcal{L}(a)$, which takes our tunable parameters as input, and tells us how well the corresponding model fits the data; small values indicate a good fit, and bad values indicate a bad fit.

Typically, the loss function is defined in terms of a *distance*, e.g. the distance between our predictions and the data (we'll see this in more depth later).

For the moment, our question is how to find the best model, i.e. the value of a with the smallest loss, $\mathcal{L}(a)$. Well, this is an optimization problem, and it turns out we can use calculus to solve optimization problems.



Specifically, the minimum of the loss is usually at a location where the gradient is zero (left). Of course, we have to be careful with this: the gradient can also be zero at a maximum (middle), or at neither a maximum or minmum (right). (Though issues caused by this are very rare, and won't occur in the simple cases in this course.)

In fact, most of the functions we're going to try to minimize are as simple as quadratics! For instance, the loss might look like,

$$\mathcal{L}(a) = a^2 + 5a - 3 \tag{21}$$

To minimize $\mathcal{L}(a)$, we find the place where the slope is zero,

$$0 = \frac{\partial \mathcal{L}(a)}{\partial a} \tag{22}$$

$$0 = \frac{\partial}{\partial a}[a^2 + 5a - 3] \tag{23}$$

$$0 = \frac{\partial a^2}{\partial a} + \frac{\partial 5a}{\partial a} - \frac{\partial 3}{\partial a} \tag{24}$$

$$0 = 2a + 5 \tag{25}$$

Then, we can solve for a,

$$2a = -5, \tag{26}$$

$$a = -\tfrac{5}{2}. \tag{27}$$

Now, this looks kind-of simplistic. Of course, the real calculation is more complicated, largely because it involves summing over datapoints (which we will see next). But finding the best model/straight line really does end up involving minimizing a quadratic!

2 Sums and products

Typically, we're going to be working N datapoints, where the ith datapoint is x_i. So the first datapoint is x_1, the second datapoint is x_2, and the last datapoint is x_N. We're often going to need to sum or multiply across many different datapoints, similarly to a for loop.

To sum over datapoints, we use the summation notation. You can and should think of this as just a for-loop. For instance, the sum of i^2 for $i = 1$ to $i = 3$ can be written,

$$\sum_{i=1}^{3} i^2 = 1^2 + 2^2 + 3^2. \tag{28}$$

If we have datapoints, x_i, then we could sum over datapoints using,

$$\sum_{i=1}^{N} x_i = x_1 + x_2 + \cdots + x_N, \tag{29}$$

which indicates that we should sum x_i for $i = 1$ to $i = N$.

To take a product over datapoints, we use product (or capital pi) notation. This is just like the sum, except that we multiply each term, rather than adding them. For instance, the product of i for $i = 1$ to $i = 3$ can be written,

$$\prod_{i=1}^{3} i = 1 \times 2 \times 3. \tag{30}$$

We could also take the product over datapoints, x_i, using,

$$\prod_{i=1}^{N} x_i = x_1 \times x_2 \times \cdots \times x_N, \tag{31}$$

which indicates that we should take the product of x_i for $i = 1$ to $i = N$.

I will generally try to be explicit about the upper and lower limits. But sometimes, especially in other people's material, you will see abbreviated notation, missing off the upper and lower limits. If the limits are left off, then there should be some "natural" values for them to take on. For instance, if we know that there are N datapoints ranging from x_1 to x_N, then "natural" lower limit is $i = 1$ and the "natural" upper limit is $i = N$.

$$\sum_{i} x_i = \sum_{i=1}^{N} x_i \qquad\qquad \prod_{i} x_i = \prod_{i=1}^{N} x_i \tag{32}$$

Of course this is context dependent. The natural limits here are $i = 1$ to $i = N$ only because this we knew we had N datapoints.

3 Logarithms

Here, log is *always* the natural logarithm, i.e.

$$x = e^{\log x} \qquad\qquad x = \log(e^x). \tag{33}$$

(and this is also true in numerical programming e.g. Python). For our purposes, the logarithm is super-useful because it converts products into sums,

$$\log(b \times c) = \log b + \log c. \tag{34}$$

This extends to powers,

$$\log(x^y) = \log \underbrace{(x \times x \times \cdots \times x)}_{y \text{ times}} \tag{35}$$

$$= \underbrace{\log x + \log x + \cdots + \log x}_{y \text{ times}} \tag{36}$$

$$= y \log x. \tag{37}$$

But most importantly, it extends to products and summations over many elements,

$$\log(\prod_{i=1}^{N} x_i) = \sum_{i=1}^{N} \log x_i \tag{38}$$

Using logs to switch products to sums is going to be important for two reasons:

- its much easier to do e.g. calculus on sums than on products (products in calculus are a nightmare: I didn't even tell you the product rule in this document).
- If you have a big product, the result might lie outside the range of float32/float64. The minimum value of float32 is 1.2×10^{-38}. If we have $x_i = 0.1$ and $N = 38$, then $\prod_{i=1}^{N} x_i = 10^{-38}$, and we're only just inside the range of float32's.

4 Vectors, matrices and index notation

4.1 Notation/types

To be super-clear about everything's type, we always ensure scalars/vectors/matrices are distinguishable. In particular, we always write:

- Scalars as non-bold (e.g. a or A).
- Vectors as bold and lowercase (e.g. $\mathbf{a}$). (Single-underline when handwritten.)
- Matrices as bold and uppercase (e.g. $\mathbf{A}$). (Double-underline when handwritten.)

The components of the vector $\mathbf{a}$ can be written,

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{pmatrix} \tag{39}$$

where a_i is a single component of the vector. Note that elements of vectors/matrices, here a_i, are always scalars, so they are always non-bold.

We write matrices as bold upper-case letters, such as $\mathbf{A}$. A single component of $\mathbf{A}$ is written as A_{ij}, where i indexes the row and j indexes the column. We write the shape as $N \times M$, which means it has N rows and M columns.

- i indexes rows and we have N rows, so i runs from 1 to N.
- j indexes columns and we have M columns, so j runs from 1 to M.

Remember that we write N first in $N \times M$, and i is the first index in A_{ij}, so it is i that runs from 1 to N.

Dropping the colours, an $N \times M$ matrix $\mathbf{A}$, can be written,

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1M} \\ a_{21} & a_{22} & \dots & a_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \dots & a_{NM}. \end{pmatrix}, \tag{40}$$

Note how the first index maxes out in the last row at N and the second index maxes out in the last column at M.

4.2 Vectors as matrices with one row/column

Its going to be useful when we come to transposes and matrix products to think of a vector as a matrix with either one row or one column.

A matrix with only one column is called a column vector. A column vector $\mathbf{a}$ with length N, can be written,

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{pmatrix}, \tag{41}$$

and it can be understood as an $N \times 1$ matrix.

You can also have a row vector, which is a matrix $1 \times N$ with only one row.

$$\mathbf{a}^T = \begin{pmatrix} a_1 & a_2 & \dots & a_N \end{pmatrix}. \tag{42}$$

You can get a row vector by transposing a column vector (see below).

4.3 Transposes

The transpose "mirrors" the matrix or vector along the diagonal. For instance, transpose converts the $N \times 1$ column vector $\mathbf{a}$ into a $1 \times N$ row vector, $\mathbf{a}^T$,

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{pmatrix} \qquad \mathbf{a}^T = \begin{pmatrix} a_1 & a_2 & \dots & a_N \end{pmatrix} \tag{43}$$

Likewise, it converts the 2×3 matrix $\mathbf{B}$, into the 3×2 matrix $\mathbf{B}^T$,

$$\mathbf{B} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \qquad \mathbf{B}^T = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} \tag{44}$$

4.4 Matrix-matrix product

The matrix product is just a very short notation for writing a sum and a product. Before delving into the exact definition, its worth thinking about the "types" (i.e. matrix sizes). For the product of $\mathbf{A}$ and $\mathbf{B}$,

$$\underbrace{\mathbf{C}}_{M \times P} = \underbrace{\mathbf{A}}_{M \times N} \underbrace{\mathbf{B}}_{N \times P}. \tag{45}$$

In particular:

- the two inner dimensions, N, must be the same size and they're going to disappear (these are the dimensions we're going to sum over).
- $\mathbf{C}$ and $\mathbf{A}$ have the same number of rows, M (first).
- $\mathbf{C}$ and $\mathbf{B}$ have the same number of columns, P (second).

Now lets delve in to the exact definition as a sum and product,

$$C_{ik} = \sum_{j=1}^{N} A_{ij} B_{jk}. \tag{46}$$

We've highlighted the indices in the same colors as above. You can see that the structure of the indices matches that of the sizes above. In particular:

- we sum over the inner two indices, j (i.e. the second index of $\mathbf{A}$ and the first index of $\mathbf{B}$.
- i is the first index of $\mathbf{A}$ and $\mathbf{C}$ (rows).
- k is the second index of $\mathbf{B}$ and $\mathbf{C}$ (columns).

4.4.1 Matrix multiplication by hand

To actually do the matrix multiplication, you can use the "two finger method". You move your left finger left-to-right along a row on the first matrix, and your right finger top-to-bottom down a column on the second matrix. You then multiply the first two elements you see, and add them to your running total. You choose the row and column based on the element you're trying to calculate. If you're trying to calculate the element in the second row and the third column, then you'd run your left finger along the second row and you'd run your right finger down the third column.

$$\begin{pmatrix} r_{11} & \\ & \end{pmatrix} = \begin{pmatrix} \rightarrow & \rightarrow \\ & \end{pmatrix} \begin{pmatrix} \downarrow & \\ \downarrow & \end{pmatrix} \tag{47}$$

$$\begin{pmatrix} & r_{12} \\ & \end{pmatrix} = \begin{pmatrix} \rightarrow & \rightarrow \\ & \end{pmatrix} \begin{pmatrix} & \downarrow \\ & \downarrow \end{pmatrix} \tag{48}$$

$$\begin{pmatrix} & \\ r_{21} & \end{pmatrix} = \begin{pmatrix} \rightarrow & \rightarrow \\ & \end{pmatrix} \begin{pmatrix} \downarrow & \\ \downarrow & \end{pmatrix} \tag{49}$$

$$\begin{pmatrix} & \\ & r_{22} \end{pmatrix} = \begin{pmatrix} \rightarrow & \rightarrow \\ & \end{pmatrix} \begin{pmatrix} & \downarrow \\ & \downarrow \end{pmatrix} \tag{50}$$

4.4.2 Matrix multiplication example

You can verify the first line by using the "two-finger method" described above, or by using the explicit formula for matrix multiplication.

$$\begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 5\times 1 + 6\times 3 & 5\times 2 + 6\times 4 \\ 7\times 1 + 8\times 3 & 7\times 2 + 8\times 4 \end{pmatrix} \tag{51}$$

$$= \begin{pmatrix} 5 + 18 & 10 + 24 \\ 7 + 24 & 14 + 32 \end{pmatrix} \tag{52}$$

$$= \begin{pmatrix} 23 & 34 \\ 31 & 46 \end{pmatrix} \tag{53}$$

4.5 Matrix-vector product

A matrix-vector product is the same as the matrix-matrix product, but we take $P = 1$. That gives sizes,

$$\underbrace{\mathbf{c}}_{M\times 1} = \underbrace{\mathbf{A}}_{M\times N}\underbrace{\mathbf{b}}_{N\times 1} \tag{54}$$

where both $\mathbf{c}$ and $\mathbf{b}$ are column-vectors. The exact form for the summation is usually written a bit differently, omitting k because there's only one possible value for k,

$$c_i = \sum_{j=1}^{N} A_{ij} b_j. \tag{55}$$

4.6 Vector-matrix product

A vector-matrix product is the same as the matrix-matrix product, but we take $M = 1$. That gives sizes,

$$\underbrace{\mathbf{c}^T}_{1\times P} = \underbrace{\mathbf{a}^T}_{1\times N}\underbrace{\mathbf{B}}_{N\times P} \tag{56}$$

where both $\mathbf{c}^T$ and $\mathbf{a}^T$ are row-vectors. The exact form for the summation is usually written a bit differently, omitting i because there's only one possible value for i,

$$c_k = \sum_{j=1}^{N} b_j A_{jk} \tag{57}$$

4.7 Vector "inner" product

A vector-vector product is the same as the matrix-matrix product, but we take $M = 1$ and $P = 1$. That gives sizes,

$$\underbrace{c}_{1\times 1} = \underbrace{\mathbf{a}^T}_{1\times N}\underbrace{\mathbf{b}}_{N\times 1} \tag{58}$$

where $\mathbf{a}^T$ is a row-vector, $\mathbf{b}$ is a column-vector, and c is a scalar (or a 1×1 matrix). We write the summation omitting i and k,

$$c = \sum_{j=1}^{N} a_j b_j. \tag{59}$$

4.8 Vector "outer" product

A vector-vector product is the same as the matrix-matrix product, but we take $M = 1$ and $P = 1$. That gives sizes,

$$\underbrace{\mathbf{C}}_{M\times P} = \underbrace{\mathbf{a}}_{M\times 1}\underbrace{\mathbf{b}^T}_{1\times P} \tag{60}$$

where $\mathbf{a}$ is a column-vector, $\mathbf{b}$ is a row-vector, and $\mathbf{C}$ is a matrix. We write the summation omitting i and k,

$$C_{i,k} = a_i b_k. \tag{61}$$

4.9 Identity matrix

The identity matrix has ones along the diagonal, and zero elsewhere,

$$\mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{62}$$

If we multiply the identity matrix by any vector / matrix, we just get back the same thing,

$$\mathbf{IC} = \mathbf{C} \tag{63}$$
$$\mathbf{CI} = \mathbf{C} \tag{64}$$
$$\mathbf{a}^T\mathbf{I} = \mathbf{a}^T \tag{65}$$
$$\mathbf{Ia} = \mathbf{Ia} \tag{66}$$

4.10 Matrix inverse

The inverse, $\mathbf{A}^{-1}$, of a matrix, $\mathbf{A}$ is the matrix such that,

$$\mathbf{I} = \mathbf{A}^{-1}\mathbf{A} \tag{67}$$

In the 2×2 case,

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad \mathbf{A}^{-1} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \tag{68}$$

We can prove that this holds by first substituting the value of $\mathbf{A}$ and $\mathbf{A}^{-1}$,

$$\mathbf{A}^{-1}\mathbf{A} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{69}$$

Then we use the expression for matrix multiplication,

$$\mathbf{A}^{-1}\mathbf{A} = \frac{1}{ad - bc}\begin{pmatrix} ad - bc & db - bd \\ -ca + ac & ad - bc \end{pmatrix} \tag{70}$$
$$\mathbf{A}^{-1}\mathbf{A} = \frac{1}{ad - bc}\begin{pmatrix} ad - bc & 0 \\ 0 & ad - bc \end{pmatrix} \tag{71}$$
$$\mathbf{A}^{-1}\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{I} \tag{72}$$

In the "real-world", we use the computer to compute our matrix inverses.

5 Kronecker delta

The Kronecker delta, δ_{ij} appears when we start differentiating sums / vectors /matrices. Its a bit like the indices of an identity matrix (except that we never write I_{ij},

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{73}$$

.

5.1 The Kronecker delta often appears when we're taking gradients of vectors

Consider a vector,

$$\mathbf{a}^T = \begin{pmatrix} a_1 & a_2 & a_3 \end{pmatrix} \tag{74}$$

One thing we might want to do (usually as part of a larger calculation) is compute the gradient of $\mathbf{a}$ wrt one of the components, a_1,

$$\frac{\partial \mathbf{a}^T}{\partial a_2} = \begin{pmatrix} \frac{\partial a_1}{\partial a_2} & \frac{\partial a_2}{\partial a_2} & \frac{\partial a_3}{\partial a_2} \end{pmatrix} \tag{75}$$

The gradient is zero when the variables don't match, as e.g. changing a_2 doesn't cause a_1 to change at all,

$$\frac{\partial \mathbf{a}^T}{\partial a_2} = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \tag{76}$$

Now, we can do the same thing a bit more abstractly,

$$\mathbf{a}^T = \begin{pmatrix} a_1 & a_2 & \dots & a_N \end{pmatrix} \tag{77}$$

$$\frac{\partial \mathbf{a}^T}{\partial a_j} = \begin{pmatrix} \frac{\partial a_1}{\partial a_j} & \frac{\partial a_2}{\partial a_j} & \dots & \frac{\partial a_N}{\partial a_j} \end{pmatrix} \tag{78}$$

To represent that these gradients are 1 only when the top index matches the bottom index, we use the Kronecker delta,

$$\frac{\partial \mathbf{a}^T}{\partial a_j} = \begin{pmatrix} \delta_{1j} & \delta_{2j} & \dots & \delta_{Nj} \end{pmatrix} \tag{79}$$

And if we select out the ith element of the vector,

$$\frac{\partial a_i}{\partial a_j} = \delta_{ij}. \tag{80}$$

5.2 The Kronecker delta picks out an element of a sum

You should be happy to see a Kronecker delta turn up, because it typically makes things alot simpler! In particular, we often have Kronecker deltas in sums, and the Kronecker delta "picks out" one element of the sum.

$$\sum_{j=1}^{3} \delta_{2j} a_j = \delta_{21} x_1 + \delta_{22} x_2 + \delta_{23} x_3 \tag{81}$$

$$= 0x_1 + 1x_2 + 0x_3 \tag{82}$$

$$= x_2. \tag{83}$$

This notion of "picking out one element" is perhaps easier to see in the more general case,

$$\sum_j \delta_{ij} a_j = a_i, \tag{84}$$

which happens because δ_{ij} is zero for all j except when $j = i$.

6 Exercises

6.1 "Unit" exercises

Exercise 1. *Calculate:*

$$\frac{\partial}{\partial x}[4x^{2.5} + x^{1/2} - 6x^{-1/2}] \tag{85}$$

Exercise 2. *Use the chain rule to calculate:*

$$\frac{\partial}{\partial x}[(x+1)^3] \tag{86}$$

Exercise 3. *Calculate the same thing, without using the chain rule, by explicitly expanding the brackets and applying* $\frac{\partial x^p}{\partial x} = px^{p-1}$*,*

$$\frac{\partial}{\partial x}[(x+1)^3] \tag{87}$$

Check they give the same answer!

Exercise 4. *Find the minimum of,*

$$\mathcal{L}(a) = 4a^2 + 2a + 1. \tag{88}$$

Exercise 5. *Calculate,*

$$\sum_{i=0}^{5} i^2. \tag{89}$$

Exercise 6. *Calculate,*

$$\prod_{i=0}^{5} i. \tag{90}$$

Exercise 7. *Simplify,*

$$\sum_{i=1}^{N} b \tag{91}$$

where b *does not depend on* i*.*

Exercise 8. *Simplify,*

$$\prod_{i=1}^{N} b \tag{92}$$

where b *does not depend on* i*.*

Exercise 9. *Compute the matrix product,*

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} \tag{93}$$

Exercise 10. *Compute the matrix inverse,*

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}^{-1} \tag{94}$$

6.2 "Integration" exercises

Exercise 11. *Use the matrix inverse to solve the following expression for x_1 and x_2,*

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 6 \end{pmatrix} \tag{95}$$

Exercise 12. *We have N datapoints, x_i. We seek to find a location, b, which minimizes the sum of squared distances between each datapoint, x_i and b. Specifically, our loss function is,*

$$\mathcal{L}(b) = \sum_{i=1}^{N} (x_i - b)^2. \tag{96}$$

Find the value of b that minimizes this loss function, by solving for the value of b at which the gradient is zero, $0 = \frac{\partial}{\partial b}\mathcal{L}(b)$. Given an interpretation of this value of b.

Exercise 13. *The absolute value is defined as,*

$$|x| = \begin{cases} x & \text{if } x > 0 \\ -x & \text{otherwise.} \end{cases} \tag{97}$$

Sketch a plot of $y = |x|$, and compute,

$$\frac{\partial |x|}{\partial x} \tag{98}$$

(Compute the gradient separately for $x > 0$ and $x \leq 0$.

Exercise 14. *(This is quite a bit harder than the typical exam question)! We have N datapoints, x_i. We seek to find a location, b, which minimizes the sum of distances (not squared distances) between each datapoint, x_i and b. Specifically, our loss function is,*

$$\mathcal{L}(b) = \sum_{i=1}^{N} |x_i - b|, \tag{99}$$

Part 1: Compute the gradient of this following loss function, wrt b. Write your answer in terms of the sign *function (see answer to the prevous question for details).*

Part 2: Rewrite the gradient in words in terms of the number of datapoints above and below b.

Part 3: Give an interpretation of the optimal value of b (i.e. the value of b for which the gradient is zero).

Exercise 15. *Find the value of*

$$\frac{\partial y_i}{\partial w_k} \tag{100}$$

in terms of the fixed $\mathbf{X}$*, where* $\mathbf{y}$ *is given by,*

$$\mathbf{y} = \mathbf{X}\mathbf{w} \tag{101}$$

7 Answers

7.1 "Unit" Answers

Answer 1. *Calculate:*

$$\frac{\partial}{\partial x}[4x^{2.5} + x^{1/2} - 6x^{-1/2}] = 4\frac{\partial x^{2.5}}{\partial x} + \frac{\partial x^{1/2}}{\partial x} - 6\frac{\partial x^{-1/2}}{\partial x} \tag{102}$$

$$= 4(2.5x^{1.5}) + \tfrac{1}{2}x^{-1/2} - 6(-\tfrac{1}{2}x^{-1.5}) \tag{103}$$

$$= 10x^{1.5} + \tfrac{1}{2}x^{-1/2} + 3x^{-1.5}) \tag{104}$$

Answer 2. *Calculate the same thing, without using the chain rule, by explicitly expanding the brackets and applying* $\frac{\partial x^p}{\partial x} = px^{p-1}$*,*

$$\frac{\partial}{\partial x}[(x+1)^3] \tag{105}$$

set

$$u = (x+1) \tag{106}$$

$$y = (x+1)^3 = u^3 \tag{107}$$

Thus we can apply the chain rule,

$$\frac{\partial}{\partial x}[(x+1)^3] = \frac{\partial y}{\partial x} \tag{108}$$

$$= \frac{\partial u}{\partial x}\frac{\partial y}{\partial u} \tag{109}$$

$$= \frac{\partial x+1}{\partial x}\frac{\partial u^3}{\partial u} \tag{110}$$

$$= 1 \times 3u^2 \tag{111}$$

$$= 3(x+1)^2 \tag{112}$$

Answer 3. *Calculate the same thing, without using the chain rule, but explicitly expanding the brackets and applying* $\frac{\partial x^p}{\partial x} = px^{p-1}$*,*

$$\frac{\partial}{\partial x}[(x+1)^3] = \frac{\partial}{\partial x}[(x+1)(x+1)(x+1)] \tag{113}$$

$$= \frac{\partial}{\partial x}[(x^2+2x+1)(x+1)] \tag{114}$$

$$= \frac{\partial}{\partial x}[(x^3+2x^2+x) + (x^2+2x+1)] \tag{115}$$

$$= \frac{\partial}{\partial x}[(x^3+3x^2+3x+1)] \tag{116}$$

$$= 3x^2 + 6x + 3 \tag{117}$$

$$= 3(x^2+2x+1) \tag{118}$$

$$= 3(x+1)^2 \tag{119}$$

Answer 4. *Find the minimum of,*

$$\mathcal{L}(a) = 4a^2 + 2a + 1. \tag{120}$$

Solve for the value of a where the gradient is zero,

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}(a)}{\partial a} && (121)\\ &= \frac{\partial}{\partial a}[4a^2 + 2a + 1] && (122)\\ &= 4\frac{\partial}{\partial a}[a^2] + 2\frac{\partial a}{\partial a} + \frac{\partial 1}{\partial a} && (123)\\ &= 4(2a) + 2 && (124)\\ &= 8a + 2. && (125) \end{aligned}$$

Now, we can solve for a,

$$\begin{aligned} 8a &= -2 && (126)\\ a &= -\tfrac{1}{2}. && (127) \end{aligned}$$

Answer 5. *Calculate,*

$$\begin{aligned} \sum_{i=0}^{5} i^2 &= 0^2 + 1^2 + 2^2 + 3^2 + 4^2 + 5^2 && (128)\\ &= 0 + 1 + 4 + 9 + 16 + 25 && (129)\\ &= 14 + 16 + 25 && (130)\\ &= 30 + 25 && (131)\\ &= 55 && (132) \end{aligned}$$

Answer 6. *Calculate,*

$$\begin{aligned} \prod_{i=0}^{5} i &= 0 \times 1 \times 2 \times 3 \times 4 \times 5 && (133)\\ &= 2 \times 3 \times 4 \times 5 && (134)\\ &= 6 \times 20 && (135)\\ &= 120 && (136) \end{aligned}$$

Answer 7.

$$\sum_{i=1}^{N} b = \underbrace{b + b + \cdots + b}_{N \ times} = Nb \tag{137}$$

Answer 8.

$$\prod_{i=1}^{N} b = \underbrace{b \times b \times \cdots \times b}_{N \text{ times}} = b^N \tag{138}$$

Answer 9.

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} 1 \times 5 + 2 \times 7 & 1 \times 6 + 2 \times 8 \\ 3 \times 5 + 4 \times 7 & 3 \times 6 + 4 \times 8 \end{pmatrix} \tag{139}$$

$$= \begin{pmatrix} 5 + 14 & 6 + 16 \\ 15 + 28 & 18 + 32 \end{pmatrix} \tag{140}$$

$$= \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix} \tag{141}$$

Answer 10. *Matrix inverse:*

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}^{-1} = \frac{1}{1 \times 4 - 2 \times 3} \begin{pmatrix} 4 & -2 \\ -3 & 1 \end{pmatrix} \tag{142}$$

$$= \frac{1}{4 - 6} \begin{pmatrix} 4 & -2 \\ -3 & 1 \end{pmatrix} \tag{143}$$

$$= -\frac{1}{2} \begin{pmatrix} 4 & -2 \\ -3 & 1 \end{pmatrix} \tag{144}$$

$$= \frac{1}{2} \begin{pmatrix} -4 & 2 \\ 3 & -1 \end{pmatrix} \tag{145}$$

7.2 "Integration" Answers

Answer 11.

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 6 \end{pmatrix} \tag{146}$$

Multiply on both sides by the inverse of the matrix,

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 5 \\ 6 \end{pmatrix} \tag{147}$$

A matrix times matrix-inverse is the identity,

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 5 \\ 6 \end{pmatrix} \tag{148}$$

We can substitute for the value of the matrix inverse from the previous question,

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} -4 & 2 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} 5 \\ 6 \end{pmatrix} \tag{149}$$

Then compute the matrix-vector product,

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} (-4)\times 5 + 2\times 6 \\ 3\times 5 + (-1)\times 6 \end{pmatrix} \tag{150}$$

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} -20 + 12 \\ 15 - 6 \end{pmatrix} \tag{151}$$

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} -8 \\ 9 \end{pmatrix}. \tag{152}$$

We can check this value for x_1 and x_2 is correct by substituting it back in,

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}\frac{1}{2}\begin{pmatrix} -8 \\ 9 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1\times(-8) + 2\times 9 \\ 3\times(-8) + 4\times 9 \end{pmatrix} \tag{153}$$

$$= \frac{1}{2}\begin{pmatrix} -8 + 18 \\ -24 + 36 \end{pmatrix} \tag{154}$$

$$= \frac{1}{2}\begin{pmatrix} 10 \\ 12 \end{pmatrix} \tag{155}$$

$$= \begin{pmatrix} 5 \\ 6 \end{pmatrix} \tag{156}$$

So our result was correct!

Answer 12. *Solve for the value of b at which the gradient of the loss is zero,*

$$0 = \frac{\partial \mathcal{L}(b)}{\partial b} \tag{157}$$

$$0 = \frac{\partial}{\partial b}[\sum_{i=1}^{N}(x_i - b)^2] \tag{158}$$

$$0 = \sum_{i=1}^{N}\frac{\partial}{\partial b}[(x_i - b)^2] \tag{159}$$

You can chose to expand the brackets, or to use the chain rule. We're going to expand the brackets,

$$0 = \sum_{i=1}^{N}\frac{\partial}{\partial b}[x_i^2 - 2bx_i + b^2] \tag{160}$$

$$0 = \sum_{i=1}^{N}(\frac{\partial}{\partial b}x_i^2 - \frac{\partial}{\partial b}2bx_i + \frac{\partial}{\partial b}b^2) \tag{161}$$

$$0 = \sum_{i=1}^{N}(-2x_i + 2b). \tag{162}$$

Divide both sides by 2,

$$0 = \sum_{i=1}^{N}(b - x_i). \tag{163}$$

Push the sum inside the bracket,

$$0 = \sum_{i=1}^{N} x_i + \sum_{i=1}^{N} b \tag{164}$$

$\sum_{i=1}^{N} b = Nb$, *as* b *doesn't depend on* i,

$$0 = Nb - \sum_{i=1}^{N} x_i \tag{165}$$

Now, we can solve for b *by adding* $\sum_{i=1}^{N} x_i$ *to both sides,*

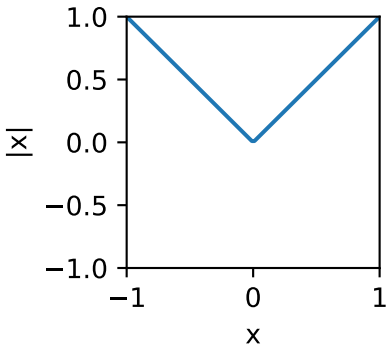
$$Nb = \sum_{i=1}^{N} x_i \tag{166}$$

and dividing both sides by N,

$$b = \tfrac{1}{N} \sum_{i=1}^{N} x_i \tag{167}$$

Therefore, b *is the mean of the data!*

Answer 13. *The sketch plot looks like:*



We can differentiate each "case" separately.

$$\frac{\partial |x|}{\partial x} = \begin{cases} \frac{\partial x}{\partial x} & \text{if } x > 0 \\ \frac{\partial -x}{\partial x} & \text{otherwise.} \end{cases} \tag{168}$$

$$= \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{otherwise.} \end{cases} \tag{169}$$

$$= \text{sign}(x) \tag{170}$$

The sign function is 1 *when the input is positive, and is* -1 *otherwise.*

FYI, there is no well-defined gradient at the "kink" (i.e. $x = 0$*), but that isn't really relevant in practice, so we aren't going to worry about it here.*

Answer 14. *Part 1 (gradient of* $\mathcal{L}(b)$ *in terms of* sign*),*

$$\mathcal{L}(b) = \sum_{i=1}^{N} \text{sign}(x_i - b). \tag{171}$$

$$\frac{\partial \mathcal{L}(b)}{\partial b} = \sum_{i=1}^{N} \frac{\partial |x_i - b|}{\partial b} \tag{172}$$

use the chain rule with $u_i = x_i - b$*,*

$$\frac{\partial \mathcal{L}(b)}{\partial b} = \sum_{i=1}^{N} \frac{\partial |u_i|}{\partial u_i} \frac{\partial u_i}{\partial b} \tag{173}$$

$$\frac{\partial \mathcal{L}(b)}{\partial b} = \sum_{i=1}^{N} \text{sign}(x_i - b)(-1) \tag{174}$$

$$\frac{\partial \mathcal{L}(b)}{\partial b} = -\sum_{i=1}^{N} \text{sign}(x_i - b). \tag{175}$$

Part 2 (interpretation in terms of the number of datapoints above and below a threshold):

$$\text{sign}(x_i - b) = \begin{cases} 1 & \text{if } x_i \geq b \\ -1 & \text{otherwise.} \end{cases} \tag{176}$$

$$-\text{sign}(x_i - b) = \begin{cases} -1 & \text{if } x_i \geq b \\ 1 & \text{otherwise.} \end{cases} \tag{177}$$

Thus, the gradient of the loss in effect counts the number of datapoints above and below b*,*

$$\frac{\partial \mathcal{L}(b)}{\partial b} = (\textit{Number of datapoints below } b) - (\textit{Number of datapoints above } b). \tag{178}$$

Part 3: the optimal value of b *is at,*

$$0 = \frac{\partial \mathcal{L}(b)}{\partial b} = (\textit{Number of datapoints below } b) - (\textit{Number of datapoints above } b). \tag{179}$$

i.e. this is a value of b *for which,*

$$(\textit{Number of datapoints below } b) = (\textit{Number of datapoints above } b) \tag{180}$$

And this value for b is the median.

FYI: There are some subtleties here about how exactly how we define the median. If there an odd number of datapoints, then the median really is the "middle" datapoint. If there an even number of datapoints, then there isn't a single "middle" datapoint: there's two. Most computer implementations of the median would return halfway between the two middle datapoints. But we have $0 = \frac{\partial \mathcal{L}(b)}{\partial b}$ *for any b between those two datapoints.*

Answer 15. *We can write out* y_i *index notation,*

$$y_i = \sum_{j=1}^{N} X_{ij} w_j \tag{181}$$

Then subsitute this expression for y_i *into the gradient we're trying to compute,*

$$\frac{\partial y_i}{\partial w_k} = \frac{\partial}{\partial w_k} [\sum_{j=1}^{N} X_{ij} w_j] \tag{182}$$

As X_{ij} *is constant, we can push the derivative inside the sum,*

$$\frac{\partial y_i}{\partial w_k} = \sum_{j=1}^{N} X_{ij} \frac{\partial w_j}{\partial w_k} \tag{183}$$

The gradient is one when $j = k$*, and zero otherwise, which matches the definition of the Kronecker delta,*

$$\frac{\partial y_i}{\partial w_k} = \sum_{j=1}^{N} X_{ij} \delta_{jk}. \tag{184}$$

Remember that the Kronecker delta is 1 *when* $j = k$ *and zero otherwise, so the Kronecker delta picks out the* $j = k$ *element of the loop,*

$$\frac{\partial y_i}{\partial w_k} = X_{ik}. \tag{185}$$