Contents

1 Data Manipulation

2 Data Preprocessing

3 Linear Algebra

1. Prove that the transpose of the transpose of a matrix is the matrix itself:

$$(A^T)^T = A$$

   - The transpose per definition satisfies $(A^T)_{ij} = A_{ji}$ for all i, j. As such $((A^T)^T)_{ij} = (A^T)_{ji} = A_{ij}$, which shows that $(A^T)^T = A$.

2. Given two matrices A and B, show that sum and transposition commute:

$$(A + B)^T = A^T + B^T$$

   - Let i, j be arbitrary valid indiced then

$$((A + B)^T)_{ij} = (A + B)_{ji} = A_{ji} + B_{ji} = (A^T)_{ij} + (B^T)_{ij}$$

     holds, this proves that $(A + B)^T = A^T + B^T$.

3. Given any square matrix A, is $A + A^T$ always symmetric? Can you prove the result by using only the result of the previous two exercises?

   - A square matrix B is symmetric if and only if $B^T = B^T$. In view of the previous two exercises the symmetry of $A + A^T$ immediately follows:

$$(A + A^T)^T \overset{2}{=} A^T + (A^T)^T \overset{1}{=} A^T + A = A + A^T.$$

     where in the last step we have used that matrix addition is commutative.

4. We defined the tensor X of shape $(2, 3, 4)$ in this section. What is the output of `len`(X)? Write your answer without implementing any code, then check your answer using code.

    - I assume that internally the tensor X is represented as $((\mathbb{R}^4)^3)^2$, which would then give the length as 2.

5. For a tensor `X` of arbitrary shape, does `len`(X) always correspond to the length of a certain axis of `X`? What is that axis?

6. Run `A / A.sum(axis=1)` and see what happens. Can you analyze the reason?

7. When traveling between two points in downtown Manhattan, what is the distance that you need to cover in terms of the coordinates, i.e., in terms of avenues and streets? Can you travel diagonally?

    - This problem wants me to state the manhattan distance:

$$d((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \sum_{i=1}^{n} |x_i - y_i|.$$

8. Consider a tensor with shape $(2, 3, 4)$. What are the shapes of the summation outputs along axis 0, 1, and 2?

9. Feed a tensor with 3 or more axes to the `linalg.norm` function and observe its output. What does this function compute for tensors of arbitrary shape?

10. Define three large matrices, say A, B, and C, for instance initialized with Gaussian random variables. You want to compute the product ABC. Is there any difference in memory footprint and speed, depending on whether you compute AB or BC first? Why?

    - Let $A \in \mathbb{R}^{r \times s}, B \in \mathbb{R}^{s \times t}$ and $C \in \mathbb{R}^{t \times u}$. We are assuming the naive matrix multiplication algorithm is used, then to calculate $A \cdot B \in \mathbb{R}^{r \times t}$ we need $O(rst)$ steps and similiarly for $B \cdot C \in \mathbb{R}^{s \times u}$ we need $O(stu)$ steps. To then calculate $(A \cdot B) \cdot C$ we need $O(rtu)$ steps and to calculate $A \cdot (B \cdot C)$ we need $O(rsu)$. As such in total we need $O(rst+rtu) = O(rt(u+s))$ to calculate $(A \cdot B) \cdot C$ and $O(stu+rsu) = O(su(t+r))$.

11. Define three large matrices, say A, B, and C. Is there any difference in speed depending on whether you compute AB or BA first? Why? What changes if you initialize B without cloning memory? Why?

12. Define three matrices, say A, B, and C. Constitute a tensor with 3 axes by stacking A, B, and C. What is the dimensionality? Slice out the second coordinate of the third axis to recover B. Check that your answer is correct.

4 Calculus

1. So far we took the rules for derivatives for granted. Using the definition and limits prove the properties for (i) $f(x) = c$, (ii) x^n, (iii) e^x and (iv) $\log x$.

   - We note that $f_1(x) = c$ satisfies $f_1(x) - f_1(x_0) = 0$ for all $x, x_0 \in \mathbb{R}$, which allows us to calculate:

   $$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{x \to x_0} \frac{c - c}{x - x_0} = 0.$$

   - Note that $f_2(x) = x^n$ is a polynomial and $F(x, y) = x^n - y^n$ is a polynomial in two variables which satisfies $F(x, x) = 0$ as such we can factor out $x - y$ from $F(x, y)$, it can inductively be shown that $F(x, y) = (x - y) \sum_{i=0}^{n-1} x^i y^{n-1-i}$. With this we can now calculate

   $$\begin{aligned}
   \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} &= \lim_{x \to x_0} \frac{x^n - x_0^n}{x - x_0} \\
   &= \lim_{x \to x_0} \frac{(x - x_0) \left(\sum_{i=0}^{n-1} x^i x_0^{n-1-i} \right)}{x - x_0} \\
   &= \lim_{x \to x_0} \sum_{i=0}^{n-1} x^i x_0^{n-1-i} \\
   &= \sum_{i=0}^{n-1} x_0^{n-1} = \left(\sum_{i=0}^{n-1} 1 \right) x_0^{n-1} = n x_0^{n-1}.
   \end{aligned}$$

   - This problem depends on which representation of the exponential is used. In the book they defined it via its differential equation, i.e. $f_3(x) = e^x$ is the unique solution of $y' = y$ with initial value $y(0) = 1$, but then there is nothing to show here. Instead we'll use the power series representation

   $$\sum_{n=0}^{\infty} \frac{x^n}{n!}$$

   of $f_3(x)$. It is easily verifiable that this power series is absolutely convergent, which allows us to show that

   $$\begin{aligned}
   f_3(x) - f_3(x_0) &= \sum_{n=0}^{\infty} \frac{x^n}{n!} - \sum_{n=0}^{\infty} \frac{x_0^n}{n!} \\
   &= \sum_{n=0}^{\infty} \frac{x^n - x_0^n}{n!} = (x - x_0) + \sum_{n=2}^{\infty} \frac{x^n - x_0^n}{n!}
   \end{aligned}$$

holds. We can use the calculation for $f_2(x)$ so write this as

$$\begin{aligned} f_3(x) - f_3(x_0) &= (x - x_0) + (x - x_0) \sum_{n=2}^{\infty} \frac{\sum_{j=0}^{n-1} x^j x_0^{n-1-j}}{n!} \\ &= (x - x_0) \left(1 + \frac{x + x_0}{2} + \dots \right) \end{aligned}$$

In total this proves that

$$f_3'(x_0) = \lim_{x \to x_0} \left(1 + \frac{x + x_0}{2} + \dots \right) = \sum_{j=0}^{\infty} \frac{x_0^j}{j!} = f_3(x_0)$$

- Consider $f_4(x) = \log(x)$ and note that $f_3 \circ f_4 = \mathrm{id}$. Using the chain rule then yields

  $$1 = f_3' \circ f_4 \cdot f_4'$$

  and because $f_3' = f_3$ this reduces to

  $$f_4' = \frac{1}{f_3 \circ f_4}$$

  so that

  $$f_4'(x) = \frac{1}{x}.$$

2. In the same vein, prove the product, sum, and quotient rule from first principles.

   - Using a similiar construction to the triangle inequality we can write

     $$\begin{aligned} (fg)(x) - (fg)(y) &= f(x)g(x) - f(y)g(y) \\ &= f(x)g(x) - f(x)g(y) + f(x)g(y) - f(y)g(y) \\ &= f(x)(g(x) - g(y)) + (f(x) - f(y))g(y), \end{aligned}$$

     from which we can conclude

     $$\begin{aligned} \lim_{x \to y} \frac{fg(x) - fg(y)}{x - y} &= \lim_{x \to y} \frac{f(x)(g(x) - g(y)) + (f(x) - f(y))g(y)}{x - y} \\ &= \lim_{x \to y} f(x) \frac{g(x) - g(y)}{x - y} + \frac{f(x) - f(y)}{x - y} g(y) \\ &= f(y)g'(y) + f'(y)g(y). \end{aligned}$$

     This can also be expressed as

     $$(fg)' = f'g + fg',$$

     which is what was to be shown.

   - The sum formula immediately follows from the linearity of lim.

- The quotient formula immediately follows from the product rule by noting that $\frac{f}{g} = f \cdot \frac{1}{g}$ as well as $\left(\frac{1}{g(x)}\right)' = -\frac{g'(x)}{g^2(x)}$. To see this apply the product rule on $1 = g(x)\frac{1}{g(x)}$.

3. Prove that the constant multiple rule follows as a special case of the product rule.

   - $(cf(x))' = c'f(x) + cf'(x)$, we have already shown that $c' = 0$ so that $(cf)' = cf'$.

4. Calculate the derivative of $f(x) = x^x$.

   - We assume that $x > 0$ then $f(x) = x^x = e^{x\log(x))}$. Let $u(x) = x\log(x)$ then $u'(x) = \log(x) + x \cdot \frac{1}{x} = \log(x) + 1$. This allows us to write $f(x) = e^{u(x)}$, the derivative of f is then, by using the chain rule and the fact that $\frac{d}{dx}e^x = e^x$, given by

$$f'(x) = u'(x)e^{u(x)} = (1 + \log(x))e^{x\log(x)} = (1 + \log(x))x^x.$$

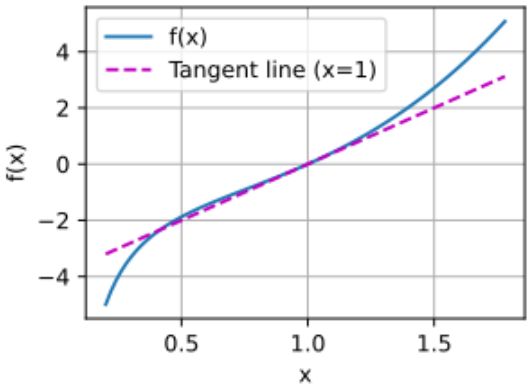
5. What does it mean that $f'(x) = 0$ for some x? Give an example of a function f and a location x for which this might hold.

   - Depending on the neighboring values (or if it's C^2 on the sign of the second derivative), that x is either a minimizer ($x = 0$ for $f(x) = x^2$), a maximizer ($x = 0$ for $f(x) = -x^2$) or a saddle point ($x = 0$ for $f(x) = x^3$).

6. Plot the function $y = f(x) = x^3 - \frac{1}{x}$ and plot its tangent line at $x = 1$.

   - The derivative of f is given by $f'(x) = 3x^2 + \frac{1}{x^2}$, evaluated at $x = 1$ this yields a slope of $f'(1) = 3 + 1 = 4$, so the tangent line is given by $y(x) = 4x + b$ where b is such thatt $y(1) = 4 + b = f(1) = 0$ so that $b = -4$. In total this means the tangent line at $x = 1$ is given by

$$y(x) = 4x - 4.$$



7. Find the gradient of the function $f(x) = ||x||_2$? What happens for $x = 0$?

   - Note that if $x = (x_1, \ldots, x_n)$ then $g(x) := f(x)^2 = ||x||_2^2 = \sum_{i=1}^n x_i^2$. It follows that for $x \neq 0$ the corresponding partial derivates are simply given by
   $$\frac{\partial}{\partial x_j} f(x)^2 = \frac{\partial}{\partial x_j} \sum_{i=1}^n x_i^2 = 2x_i.$$
   This shows that $\nabla g(x) = 2x$. Furthermore $\frac{\partial}{\partial x_j} g(x) = \frac{\partial}{\partial x_j} f(x)^2 = 2\frac{\partial f}{\partial x_j} \cdot f(x)$. With that we can calculate
   $$2x = 2f(x)\nabla f \iff \nabla f = \frac{x}{f(x)} = x/||x||.$$
   Note that this construction breaks down at $x = 0$ as we would divide by 0. But this is not surprising because $x \mapsto \sqrt{x}$ is not differentiable at 0 and $f(x) = \sqrt{g(x)}$.

8. Can you write out the chain rule for the case where $u = f(x, y, z)$ and $x = x(a, b), y = y(a, b)$ and $z = z(a, b)$?

   - We can write $u(a, b) = f(x(a, b), y(a, b), z(a, b))$ and with that
   $$\frac{\partial}{\partial a} u(a, b) = \frac{\partial f}{\partial x}(\ldots) \cdot \frac{\partial x}{\partial a} + \frac{\partial f}{\partial y}(\ldots) \cdot \frac{\partial y}{\partial a} + \frac{\partial f}{\partial z}(\ldots) \cdot \frac{\partial z}{\partial a},$$
   which can be expressed as
   $$(\nabla f)(x(a, b), y(a, b), z(a, b)) \cdot \nabla_a(x, y, z).$$
   Similiarly once can show taht
   $$(\nabla f)(x(a, b), y(a, b), z(a, b)) \cdot \nabla_b(x, y, z).$$
   With that it follows that
   $$\nabla u(a, b) = (\nabla f, \nabla f)^t \cdot (D(x, y, z)).$$

9. Given a function $f(x)$ that is invertible, compute the derivative of its inverse $f^{-1}(x)$. Here we have that $f^{-1}(f(x)) = x$ and conversely $f(f^{-1}(y)) = y$. Hint: Use these properties in your derivation.

   - We can calculate
   $$1 = \frac{\partial}{\partial x} x = \frac{\partial}{\partial x} f^{-1}(f(x)) = (f^{-1})'(f(x)) f'(x),$$
   which can be rewritten as
   $$(f^{-1})'(f(x)) = \frac{1}{f'(x)}.$$
   If we write $y = f(x)$ this becomes
   $$(f^{-1})'(y) = \frac{1}{f'(x)}.$$

5 Automatic Differentiation

1. Why is the second derivative much more expensive to compute than the first derivative?

   The derivative can be approximated by sampling two points, the second derivative needs more points. If we consider the second derivative as the derivative of the derivative then we need to sample two points at each of the original sampling points, which more than doubles the workload. Note that the second derivative is the average rate of change of the rate of change of the two sample points, symbolically:

   $$f'(x_0) \sim f(x_0 + \varepsilon) - f(x_0 - \varepsilon)$$

   and

   $$f'(x_0 + \varepsilon) = f(x_0 + 2\varepsilon) - f(x_0), \quad f'(x_0 - \varepsilon) = f(x_0) - f(x_0 - 2\varepsilon).$$

   From this it can be seen that the second derivative can be expressed as

   $$\begin{aligned} f''(x_0) &= f'(x_0 + \varepsilon) - f'(x_0 - \varepsilon) \\ &= f(x_0 + 2\varepsilon) - f(x_0) - (f(x_0) - f(x_0 - 2\varepsilon)) \\ &= f(x_0 + 2\varepsilon) - 2f(x_0) + f(x_0 - 2\varepsilon)). \end{aligned}$$

   There are much better choice for sampling points to get faster numerical convergence, but this is beyond the scope of what we're doing here.

2. After running the function for backpropagation, immediately run it again and see what happens. Why?

3. In the control flow example where we calculate the derivative of d with respect to a, what would happen if we changed the variable a to a random vector or a matrix? At this point, the result of the calculation f(a) is no longer a scalar. What happens to the result? How do we analyze this?

4. Let $f(x) = \sin(x)$. Plot the graph of f and its derivative f'. Do not exploit the fact that $f'(x) = \cos(x)$ but rather use automatic differentiation to get the result.

5. Let $f(x) = ((\log x^2) \cdot \sin x) + x^{-1}$. Write out a dependency graph tracing results from x to $f(x)$.

6. Use the chain rule to compute the derivative f' of the aforementioned function, placing each term on the dependency graph that you constructed previously.

7. Given the graph and the intermediate derivative results, you have a number of options when computing the gradient. Evaluate the result once starting from x to f and once from f tracing back to x. The path from x to f is commonly known as *forward differentiation*, whereas the path from f to x is known as *backward differentiation*.

8. When might you want to use forward differentiation and when backward differentiation? Hint: consider the amount of intermediate data needed, the ability to parallelize steps, and the size of matrices and vectors involved.

6 Probability and Statistics

This might be a bit too technical for the course, but to remind myself of the background. Let $(\Omega, \mathcal{A}, P)$ be a probability space, this is a measure space (i.e. Ω is some set, $\mathcal{A}$ is a sigma algebra on Ω and P is a measure, i.e. a σ-additive function $P : \mathcal{A} \to [0, +\infty)$ which satisfies $P(\emptyset) = 0$), where the measure is a probability function, i.e. $P(\Omega) = 1$. Recall that this means that probability theory is only interested in the theory of finite measures.

A random variable is a measureable function $X : \Omega \to \mathbb{R}$. We call the elements of the sigma algebra an event (and the sigma-algebra itself the event space) and the elements of Ω a sample (and Ω itself a sample space). The expression

$$P\left(\lim_{n\to\infty} X_n = \alpha\right) = p$$

is simply shorthand for

$$P\left(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = \alpha\}\right) = p.$$

1. Give an example where observing more data can reduce the amount of uncertainty about the outcome to an arbitrarily low level.

   - Consider flipping a weighted coin, whose probability to land on heads is given by $p_H \in (0, 1)$. Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of random variables where $X_i \in \{0, 1\}$ and is 1 if the ith flip landed on heads. Counting the outcome of the first n flips yields the random variable

     $$\tilde{X}_n := \sum_i^n X_i.$$

     Note that

     $$P\left(\lim_{n\to\infty} \frac{\tilde{X}_n}{n} \neq p_H\right) = 0,$$

     or if we consider the running average $Y_n := \frac{1}{n}\tilde{X}_n$ this can be expressed as

     $$P\left(\lim_{n\to\infty} Y_n = p_H\right) = 1.$$

2. Give an example where observing more data will only reduce the amount of uncertainty up to a point and then no further. Explain why this is the case and where you expect this point to occur.

3. We empirically demonstrated convergence to the mean for the toss of a coin. Calculate the variance of the estimate of the probability that we see a head after drawing n samples.

   (a) How does the variance scale with the number of observations?

   (b) Use Chebyshev's inequality to bound the deviation from the expectation.

   (c) How does it relate to the central limit theorem?

4. Assume that we draw samples from a probability distribution with zero mean and unit variance. Compute the averages

$$z_m = m^{-1} \sum_{i=1}^{m} x_i.$$

   Can we apply Chebychev's inequality for every z_m independently? Why not?

5. Given two events with probability $P(\mathcal{A})$ and $P(\mathcal{B})$ compute upper and lower bounds on $P(\mathcal{A} \cup \mathcal{B})$ and $P(\mathcal{A} \cap \mathcal{B})$. Hint: graph the situation using a Venn diagram.

   - Using a Venn diagram we can immediately see that

     $$\max\{P(A), P(B)\} \leq P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B),$$

     and

     $$P(A) + P(B) - 1 \leq P(A \cap B) \leq \min\{P(A), P(B)\}$$

     holds.

6. Assume that we have a sequence of random variables, say A, B, and C, where B only depends on A, and C only depends on B, can you simplify the joint probability $P(A, B, C)$? Hint: this is a Markov chain.

   - Note that $P(A, B)$ is just shorthand for $P(A \cap B)$.

     $$\begin{aligned} P(A, B, C) &= P(A|B, C)P(B, C) \\ &= P(A|B)P(B, C) \\ &= P(A|B)P(B|C)P(C) \end{aligned}$$

     where we have used (2.6.1) from the book, as well as because A and C are independent it follows that $P(A|B, C) = P(A|B)$. This can be

expressed more generally, if we have a collection of random variables X_i such that

$$P(X_{i+1}|X_i, X_{i-1}, \ldots, X_1) = P(X_{i+1}|X_i),$$

i.e. X_{i+1} only "remembers" the previous variable then

$$P(X_n, \ldots, X_1) = P(X_n|X_{n-1})P(X_{n-2}|X_{n-3}) \cdot \cdots \cdot P(X_1|X_0)P(X_0).$$

7. In Section 2.6.5, assume that the outcomes of the two tests are not independent. In particular assume that either test on its own has a false positive rate of 10% and a false negative rate of 1%. That is, assume that

$$P(D = 1|H = 0) = 0.1,$$

and that

$$P(D_1, D_2|H = 1) = 0.02.$$

   (a) Work out the joint probability table for D_1 and D_2 given $H = 0$ based on the information you have so far.

   •

   (b) Derive the probability of the patient being positive ($H = 1$) after one test returns positive. You can assume the same baseline probability $P(H = 1) = 0.0015$ as before.

   (c) Derive the probability of the patient being positive ($H = 1$) and both tests return positive.

8. Assume that you are an asset manager for an investment bank and you have a choice of stocks s_i to invest in. Your portfolio needs to add up to 1 with weights α_i for each stock. The stocks have an average return $\mu = E_{s\sim P}[s]$ and covariance $\Sigma = \text{Cov}_{s\sim P}[s]$.

   (a) Compute the expected return for a given portfolio α.

   (b) If you wanted to maximize the return of the portfolio, how should you choose your investment?

   (c) Compute the variance of the portfolio.

   (d) Formulate an optimization problem of maximizing the return while keeping the variance constrained to an upper bound. This is the Nobel-Prize winning Markovitz portfolio (Mangram, 2013). To solve it you will need a quadratic programming solver, something way beyond the scope of this book.

7 Documentation