70015 Mathematics for Machine Learning: Exercises

Mark van der Wilk, Yingzhen Li*
{m.vdwilk,yingzhen.li}@imperial.ac.uk

October 21, 2022

Contents

1 Notation

1.1 Sets

Throughout this course, we will be using some standard mathematical notation which may be unfamiliar to some. It's ultimately not that special or even crucial to the overall argument, but it is compact (which is practical), and it helps somewhat with practising with expressing things mathematically. Wikipedia has good definitions on these things too.

*Many thanks to teaching assistants Carles Balsells Rodas, and Alex Spies for their solutions and improvements to the document.

- Notation referring to sets of numbers, e.g. the natural numbers $\mathbb{N} = \{0, 1, 2, \dots\}$, integers $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$, or real numbers $\mathbb{R}$.
- Vectors are sets containing n of some type of object, like reals. We denote the set of all such sets using a superscript notation. For example, all n-dimensional vectors becomes $\mathbb{R}^n$.
- With $x \in \mathcal{S}$ we denote that x is an element of the set $\mathcal{S}$. This allows us to specify that a variable comes from a particular set (or, has a particular type), e.g. $x \in \mathbb{R}^D$.
- We sometimes use "set builder" notation. We did this informally above when defining $\mathbb{N}$! Usually this works by specifying elements with some property, e.g. $\mathbf{S} = \{2n \,|\, n \in \mathbb{N}\}$, which means "all the elements 2n such that n is a natural number". This creates the set of all even positive whole numbers.
- We denote the union of two sets (the set with all elements that are in either set or both) as $A \cup B$. With set-builder notation this is $A \cup B = \{x \,|\, x \in A \vee x \in B\}$, where $\vee$ means "or".
- We denote the intersection of two sets (the set of all elements that are in both stes) as $A \cap B = \{x \,|\, x \in A \wedge x \in B\}$.
- For intervals of real numbers, we use brackets, $[,]$, to denote the elements in the set which are "greater than or equal to" and "less than or equal to" an element, respectively. We use parentheses, $(,)$ to denote a strict lower bound or upper bound on the set, respectively. E.g. $[1, 5)$ is equivalent to $1 \leq x < 5, x \in \mathbb{R}$.
- We use the symbol $\neg$ to denote the complement of a set. Given a set containing all elements under consideration Ω, $\neg A$ contains all elements of Ω that are not in A, i.e. $\neg A = \{x \in \Omega | x \notin A\}$. We can also denote this as $\neg A = \Omega \backslash A$.

1.2 Probabilities

In this course we will use the notation for probabilities that is common in machine learning. The main advantage is that this notation is shorter, although it does leave certain things implicit. We include this to reduce confusion.

Consider a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ with sample space Ω (all possible outcomes of a random procedure), event space $\mathcal{E}$ (the set of all sets of outcomes that we assign a probability to), and probability function $\mathbb{P} : \mathcal{E} \to [0, 1]$ (a function that assigns a probability to an event), with a random variable $X : \Omega \to \mathbb{R}^D$.

- With $\mathbb{P}(E)$ we denote the probability of an event $E \in \mathcal{E}$, where E is a set of outcomes.
- Following the usual convention, we use the same notation when considering random variables, e.g. $\mathbb{P}(X < 2)$ is short for $\mathbb{P}(\{s \in \Omega : X(s) < 2\})$ (see §6.1 in `50008` *Probability & Statistics*).
- We usually work directly with random variables, and specify all properties using a probability mass function (pmf) or probability density function (pdf). For a specific outcome of the random variable α, we write:

$$\mathbb{P}(X = \alpha) = p_X(\alpha) \qquad \text{for a pmf } p_X(\cdot)\,, \tag{1}$$

$$\mathbb{P}(X \in [a, b]) = \int_a^b p_X(\alpha) d\alpha \qquad \text{for a pdf } p_X(\cdot) \text{ with } \alpha \in \mathbb{R}\,, \tag{2}$$

$$\mathbb{P}(X \in A) = \int_A p_X(\alpha) d\alpha \qquad \text{for a pdf } p_X(\cdot) \text{ with } \alpha \in \mathbb{R}^D\,. \tag{3}$$

- Sometimes we may write vectors in boldface, i.e. $\mathbf{x} \in \mathbb{R}^D$. We won't always though, so keep track of how we define variables!
- We generally denote outcomes of random variables without referring explicitly to the random variable itself. For example, when we refer to an outcome $\mathbf{x}$, we implicitly know there is a random variable that can take this value. We usually denote this as the capital, for example here X.
- Sometimes we abuse notation, and drop the random variable when denoting distributions when the argument of the function identifies it, e.g. $p(\mathbf{x}) = p_X(\mathbf{x})$.

- If we want to be explicit about the random variable that we are evaluating the density/mass of, I will write e.g. $p_{X,Y}(\mathbf{x}, \mathbf{y}) = p_{X|Y}(\mathbf{x}|\mathbf{y})p_Y(\mathbf{y})$.
- Expectations can be denoted in two ways:

$$\mathbb{E}_X[f(X)] \qquad \text{to emphasise that X is random, if it is clear what its distribution is}\,, \tag{4}$$
$$\mathbb{E}_{p(\mathbf{x})}[f(\mathbf{x})] \qquad \text{to emphasise that we will be integrating over the distribution } p(\mathbf{x})\,. \tag{5}$$

  In both cases this corresponds to the integral $\int p(\mathbf{x})f(\mathbf{x})\mathrm{d}\mathbf{x}$.
- Often, densities and pmfs can be discussed in exactly the same way, if we think of the density of a discrete RV as a sum of delta functions. I.e. $p(\mathbf{x}) = \sum_o \delta(\mathbf{x}-\mathbf{x}_o)p_o$, where $\{\mathbf{x}_o\}$ is the set of discrete possible outcomes that X can take, and p_o are their corresponding probabilities. This allows us to write an expectation as an integral, regardless of whether the RV is continuous or discrete, because for discrete RVs we get:

$$\mathbb{E}_{p(\mathbf{x})}[f(\mathbf{x})] = \int p(\mathbf{x})f(\mathbf{x})\mathrm{d}\mathbf{x} = \int \sum_o \delta(\mathbf{x}-\mathbf{x}_o)p_o f(\mathbf{x})\mathrm{d}\mathbf{x} = \sum_o f(\mathbf{x}_o)p_o\,. \tag{6}$$

  (A delta function has the property that $\int_A \delta(\mathbf{x})\mathrm{d}\mathbf{x}$ is 1 if $0 \in A$, and 0 otherwise. Linearity of integrals still holds. It can often be seen as the limit of a Gaussian distribution with zero variance.)

2 Formula Sheet

- Gaussian probability density function (pdf) with input $\mathbf{x} \in \mathbb{R}^D$, denoted as $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{7}$$

- For a joint Gaussian density

$$p\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}; \begin{bmatrix}\mathbf{m_x}\\ \mathbf{m_y}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma_{xx}} & \boldsymbol{\Sigma_{xy}}\\ \boldsymbol{\Sigma_{yx}} & \boldsymbol{\Sigma_{yy}}\end{bmatrix}\right), \tag{8}$$

  we have the conditional density

$$p(\mathbf{x}\,|\,\mathbf{y}) = \mathcal{N}\big(\mathbf{x};\quad \mathbf{m_x} + \boldsymbol{\Sigma_{xy}}\boldsymbol{\Sigma_{yy}}^{-1}(\mathbf{y}-\mathbf{m_y}),\quad \boldsymbol{\Sigma_{xx}} - \boldsymbol{\Sigma_{xy}}\boldsymbol{\Sigma_{yy}}^{-1}\boldsymbol{\Sigma_{yx}}\big)\,. \tag{9}$$

3 Warm-up Exercises

To start, here are some exercises which test knowledge which is assumed in the course.

3.1 Probability Theory

We assume that you are familiar with probability theory up to the Computing 2nd year `50008` *Probability & Statistics* course. Here are some questions to serve as a refresher. Students who are not familiar with this background should refer to the notes of `50008` *Probability & Statistics* or relevant chapters of **[mml]**. **We recommend you look at these questions when/before the course starts**. If you need a refresher, or if you do not know the notation, refer to the `50008` *Probability & Statistics* notes, or discuss with a TA.

Question 1 (Set Theory and Probability)**.** Using the three axioms of probability show that

a. Write down the sample space of a dice. In your notation, use the set A to denote the event of a 3 or 4 occurring. What is the complement of A, denoted $\neg A$?

b. For a problem about lengths, we have a sample space $\Omega = [0, 1]$. For $A = (0.3, 0.4]$, what is $\neg A$?

c. $\mathbb{P}(\neg A) = 1 - \mathbb{P}(A)$

d. $\mathbb{P}(\varnothing) = 0$, where $\varnothing$ is the empty set

e. $0 \leq \mathbb{P}(A) \leq 1$

f. $A \subseteq B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$

Hint: Consider the following definition. $B \backslash A = \{x \in B : x \notin A\}$

g. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

h. (*) if $\{A_i\}_{i=1}^{\infty} \subseteq \Omega$ and $A_i \subseteq A_{i+1} \forall i$ then:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i \to \infty} \mathbb{P}(A_i)$$

Hint: Use axiom 3. *: The emphasis of this course isn't on these kinds of details, even though this should be doable with 1st-year calculus.

i. For two mutually exclusive events A, B, what is $\mathbb{P}(A \cup B)$?

See **mml** for a general overview, and §4, §§5.1-5.4 of `50008` *Probability & Statistics* for more details.

Question 2 (Independent events)**. Independent events don't come up as much as independent random variables, so it's ok to just follow this answer, rather than spending lots of time on it.** When tossing two coins (where we care about the order), we have a sample space $\Omega = \{HH, HT, TH, TT\}$.

a. What outcomes are contained in the event that corresponds to the the first coin being heads? We denote the event E_{1H}, and others similarly.

b. If you assume that all outcomes have equal probability, show that E_{1H} and E_{2T} are independent.

c. If you assume that E_{1H} and E_{2H} are independent and 0.5 each, show that all outomes must have equal probability.

See §5.3.3 in `50008` *Probability & Statistics*.

Question 3 (Random Variables)**.** Consider throwing two fair dice.

a. What is the sample space for all outcomes that you can get from throwing two dice? We specify the probability of each outcome to be the same.

b. Define two random variables A, B which map the outcome to the face value on each die respectively. Find the probability mass function for A from the probability on outcomes. The answer will work from the definition of a random variable, but you will probably intuitively get the right answer as well.

c. Show that A and B are independent.

d. Define the random variable $C = A + B$. Derive the probability mass function of C.

See §6 of `50008` *Probability & Statistics*.

Question 4 (Continuous Random Variables)**.** Consider the random variable X with a probability density $p(x) = C \cdot x$ when $x \in [0, 1]$ and 0 elsewhere.

a. Calculate C.

b. Calculate $\mathbb{P}(0.3 \leq X \leq 0.75)$.

c. Calculate $\mathbb{P}(X \in [0.3, 0.75] \cup [0.8, 0.9])$.

d. Calculate $\mathbb{E}_X[X]$, $\mathbb{E}_X\left[X^2\right]$, $\mathbb{V}_X[X]$.

Check your answers by performing numerical integration, e.g. in Python.
See §6.3, §7 of `50008` *Probability & Statistics* or **mml**.

Question 5 (Joint Discrete Random Variables)**.** Consider two random variables A, C, where A is the outcome of one die, and C gives the sum of A and the sum of another die B.

a. From intuition, write a table of $\mathbb{P}(C = c|A = a)$, which we use to denote the probability of C taking the value c, if we know that A has taken the value a.

b. Write a table of $\mathbb{P}(C = c, A = a)$. To help you think it through, consider a tree of outcomes that can occur. This helps illustrate independence between outcomes, which helps you figure out when you can multiply probabilities.

c. From the values in the table $\mathbb{P}(C = c, A = a)$ find $\mathbb{P}(2 \leq C \leq 4)$ and $\mathbb{P}(2 \leq C \leq 4, 2 \leq A \leq 4)$.

We will cover conditional probability more later, but for now just think it through.

Question 6 (Multivariate Integration)**.** Consider two continuous random variables X, Y with joint density $p(x, y) = C \cdot (x^2 + xy)$ when $x \in [0, 1]$ and $y \in [0, 1]$, and 0 elsewhere.

a. Find C.

b. Find $\mathbb{P}(0.3 \leq X \leq 0.5)$.

c. Find $\mathbb{P}(X < Y)$. Perform the integration twice in both orders, once integrating over x first, once by integrating over y first.

d. **Bonus:** Convince yourself that you know how to do this for $p(x, y, z) = C \cdot (x^2 + xyz)$ as well.

Check your answers by performing numerical integration, e.g. in Python.

Question 7 (Statistics Terminology)**.** Recall the following statistical terminology.

a. What is a statistic?

b. What is an estimator?

c. What is a consistent estimator?

d. What is a sample?

3.2 Linear Algebra

Question 8 (Dot product)**.** Compute $\mathbf{x}^\top \mathbf{y}$ where $\mathbf{x} = (1, -2, 5, -1)^\top$ and $\mathbf{y} = (0, 4, -3, 7)^\top$.

Question 9 (Matrix product)**.** Compute $\mathbf{y} = A\mathbf{x}$ as well as the ℓ_2 norm of $\mathbf{x}$ and $\mathbf{y}$, where

$$A = \begin{pmatrix} -1 & 4 & 7 & 2 \\ 3 & -2 & -1 & 0 \\ 5 & 3 & 0 & -1 \end{pmatrix}, \quad \mathbf{x} = (-3, 2, 1, 3)^\top.$$

Question 10 (Basis)**.** Which of the following set of vectors are basis for $\mathbb{R}^2$?

1. $\{(1, 1), (1, 0)\}$
2. $\{(2, 4), (3, -1)\}$
3. $\{(1, -1), (0, 2), (2, 1)\}$
4. $\{(2, -1), (-2, 1)\}$
5. $\{(0, 3)\}$

Question 11 (Span of vectors)**.** Which of the following points are within the span of $\{(-1, 0, 2), (3, 1, 0)\}$?

1. $(0, 1, 1)$
2. $(1, 1, 4)$
3. $(2, 1, 1)$
4. $(-3, 4, 2)$
5. $(0, 0, 0)$

Question 12 (Rotation matrix in $\mathbb{R}^2$)**.** What is the 2×2 matrix that rotates all the non-zero vectors in $\mathbb{R}^2$ by $45°$ counter-clockwise?

Question 13 (Linear equations)**.** Given the following system of linear equations:

$$\begin{aligned} x + 2y &= 2 \\ 3x + 2y + 4z &= 5 \\ -2x + y - 2z &= -1 \end{aligned}$$

Answer the following questions:

a Writing this system in a matrix form $A\mathbf{x} = \mathbf{b}$ with $\mathbf{x} = (x, y, z)^\top$. What are A and $\mathbf{b}$?

b Solve this system, or show that the solution does not exist.

c What is the rank of A?

Question 14 (Eigen decomposition)**.** Consider a matrix $A \in \mathbb{R}^{d \times d}$ and assume it has an eigen decomposition of $A = Q\Lambda Q^{-1}$ where $\Lambda = \text{diag}(\lambda_1, ..., \lambda_d)$. When A is symmetric we also have $Q^{-1} = Q^\top$. Answer the following questions:

a. If A is symmetric, show that $\mathbf{x}^\top A\mathbf{x} \geq 0$ for any $\mathbf{x} \in \mathbb{R}^{d \times 1}$ if and only if $\lambda_i \geq 0$ for all $i = 1, ..., d$.

b. Show that $Tr(A) = \sum_{i=1}^{d} \lambda_i$ where $Tr(A)$ is the trace of A.

c. Show that $det(A) = \prod_{i=1}^{d} \lambda_i$ where $det(A)$ is the determinant of A.

d. Why an entry λ_i in the diagonal matrix Λ is one of the solutions for the equation $A\boldsymbol{q} = \lambda\boldsymbol{q}$, $\boldsymbol{q} \neq \mathbf{0}$?

4 Lecture 1: Probability, Vectors, Differentiation

Question 15 (Vector notation)**.** We define the probability density on the vector $\mathbf{x} \in \mathbb{R}^3$ with all elements $0 \leq x_k \leq 1$ as

$$p(\mathbf{x}) = \frac{1}{C}(x_1^2 + x_1x_2 + x_2^2 + 2x_2x_3)\,. \tag{10}$$

Put this into notation that only uses $\mathbf{x}$ as a single whole vector.

Question 16 (Noise conditional independence)**.** Consider the probability of the data in linear regression, for a fixed setting of the parameters $\boldsymbol{\theta}$ and given inputs $\mathrm{X} \in \mathbb{R}^{N \times D}$ where $\mathrm{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$:

$$p(\mathbf{y}|\boldsymbol{\theta}, \mathrm{X}) = \mathcal{N}\big(\mathbf{y}; \boldsymbol{\theta}^\intercal\mathbf{x}, \sigma^2\mathrm{I}\big) \tag{11}$$

Show that all y_ns are independent, for a fixed setting of the parameters $\boldsymbol{\theta}$ and given inputs X.

Question 17 (Maximum likelihood revision)**.** For a Gaussian distribution with mean μ and variance σ^2.

a. Derive the probability distribution for N iid draws.

b. Derive the maximum likelihood estimator for the mean μ.

Question 18 (Maximum likelihood and minimum loss)**.** Show that the solution to the Maximum Likelihood estimator for linear regression is the same as the minimum squared loss estimator.

Question 19 (MML 5.1-5.3)**.** This is revision. Compute the derivatives for w.r.t. x for

a. $f(x) = \log(x^4)\sin(x^3)$

b. $f(x) = (1 + \exp(-x))^{-1}$

c. $f(x) = \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)$

5 Lecture 2: Vector Differentiation

Question 20 (Circle)**.** Consider a vector function $\mathbf{x}(t) = \begin{bmatrix}\cos t & \sin t\end{bmatrix}^\intercal$.

a. Draw the set of points that this function passes through.

b. To build intuition, draw the velocity vector at a few points by considering the direction that the point moves in.

c. Find the derivative $\mathrm{d}\mathbf{x}/\mathrm{d}t$. Draw this vector for some point t.

Question 21 (Index notation)**.** Turn the following matrix-vector expressions into index notation:

a. $\mathrm{ABC}\mathbf{x}$

b. $\mathrm{Tr(A)}$

c. $\mathrm{Tr(AB)}$

d. $\mathbf{y}^\intercal \mathrm{A}^\intercal \mathbf{x}$

Turn the following index expressions back to matrix-vector notation:

a. $\sum_{ijk} A_{ij} B_{jk} C_{ki}$

b. $b_i + \sum_j A_{ij} b_j$

c. $x_i x_j$

d. $\sum_j \delta_{ij} a_j$

Question 22 (Index notation proofs)**.** Using index notation, show that

1. $\mathbf{x}^\intercal \mathrm{A}\mathbf{y} = \mathbf{y}^\intercal \mathrm{A}\mathbf{x}$ if A is symmetric, i.e. $\mathrm{A} = \mathrm{A}^\intercal$.
2. $\mathbf{x}^\intercal \mathbf{y} = \mathrm{Tr}(\mathbf{x}^\intercal \mathbf{y}) = \mathrm{Tr}(\mathbf{y}^\intercal \mathbf{x})$, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$.
3. $\mathrm{Tr(ABC)} = \mathrm{Tr(CAB)}$.

Question 23 (MML 5.5-5.6)**.** First find the dimensions, then the Jacobian. It's probably easiest here to use index notation.

a. $f(\mathbf{x}) = \sin(x_1)\cos(x_2)$, find $\mathrm{d}f/\mathrm{d}\mathbf{x}$.

b. $f(\mathbf{x}) = \mathbf{x}^\intercal \mathbf{y}$, find $\mathrm{d}f/\mathrm{d}\mathbf{x}$.

c. $f(\mathbf{x}) = \mathbf{x}\mathbf{x}^\intercal$, find $\mathrm{d}f/\mathrm{d}\mathbf{x}$.

d. $f(\mathbf{t}) = \sin(\log(\mathbf{t}^\intercal \mathbf{t}))$, find $\mathrm{d}f/\mathrm{d}\mathbf{t}$.

e. $f(\mathrm{X}) = \mathrm{Tr(AXB)}$ for $\mathrm{A} \in \mathbb{R}^{D\times E}$, $\mathrm{X} \in \mathbb{R}^{E\times F}$, $\mathrm{B} \in \mathbb{R}^{F\times D}$, find $\mathrm{d}f/\mathrm{dX}$.

Question 24 (MML 5.7-5.8: Chain rule)**.** Comupte the derivatives $\mathrm{d}f/\mathrm{d}\mathbf{x}$ of the following functions.

- First, write out the chain rule for the given decomposition.
- Give the shapes of intermediate results, and make clear which dimension(s) will be summed over.
- Provide expressions for the derivatives, and describe your steps in detail. Providing an expression means specifying everything up to the point where you could implement it.
- Give the results in vector notation if you can.

a. $f(z) = \log(1+z)$, $\quad z = \mathbf{x}^\intercal \mathbf{x}$, $\quad \mathbf{x} \in \mathbb{R}^D$.

b. $f(\mathbf{z}) = \sin(\mathbf{z})$, $\quad \mathbf{z} = \mathrm{A}\mathbf{x} + \mathbf{b}$, $\quad \mathrm{A} \in \mathbb{R}^{E\times D}$. What sizes are $\mathbf{x}$ and $\mathbf{b}$?

c. $f(z) = \exp(-\frac{1}{2}z)$, $\quad z = \mathbf{y}^\intercal \mathrm{S}^{-1}\mathbf{y}$, $\quad \mathbf{y} = \mathbf{x} - \boldsymbol{\mu}$.

d. $f(\mathrm{A}) = \mathrm{Tr(A)}$, $\quad \mathrm{A} = \mathbf{x}\mathbf{x}^\intercal + \sigma^2\mathrm{I}$.

e. $\mathbf{f}(\mathbf{z}) = \tanh(\mathbf{z})$, $\quad \mathbf{z} = \mathrm{A}\mathbf{x} + \mathbf{b}$, $\quad \mathrm{A} \in \mathbb{R}^{M\times N}$.

f. $f(\mathrm{A}) = \mathbf{x}^\intercal \mathrm{A}\mathbf{x}$, $\quad \mathrm{A} = \mathbf{x}\mathbf{x}^\intercal$.

Remember: Generally, scalar functions are applied elementwise to vectors/matrices.

Question 25 (Hessian of Linear Regression)**.** For the stationary point of linear regression, find the Hessian, and prove that it is positive definite, perhaps by making some assumptions. Discuss your assumptions.

6 Lecture 3: Automatic Differentiation

Question 26 (Product rule)**.** Consider the function $f(a, b) = a \cdot b$, where $a = a(x), b = b(x)$, i.e. unspecified functions of x.

- Show that by following forward mode autodiff, you effectively calculate the product rule.
- Show that if $a(x) = x, b(x) = x$, which means that the overall function $f(x) = x^2$, the gradient that is computed will be $2x$.

(Note from MvdW (autumn 2022): I somewhat messily described this on the board. The question is included here to provide a clearer explanation.)

Question 27 (Multivariate Autodiff)**.** This is a rather big question that should test your understanding of all material in the first three lectures. Consider the overall function $f(\boldsymbol{\ell}, \mathrm{X})$ consisting of the parts:

$$f = \mathbf{y}^\intercal (\mathrm{K}_1 + \mathrm{K}_2)^{-1} \mathbf{y}\,, \tag{12}$$

$$\mathrm{K}_a = \exp(\Lambda_a)\,, \tag{13}$$

$$\Lambda_a = -\frac{\mathrm{D}_a}{2\ell_a^2}\,, \tag{14}$$

$$\mathrm{D}_a = (\mathrm{X}[:, \texttt{None}, a] - \mathrm{X}[\texttt{None}, :, a])^2\,, \tag{15}$$

where we use `numpy` broadcasting notation in the final equation.

a. Given $\boldsymbol{\ell} \in \mathbb{R}^2$ and $\mathrm{X} \in \mathbb{R}^{N \times 2}$, find the shape of all intermediate computations.

b. Draw the computational graph for $f(\boldsymbol{\ell}, \mathrm{X})$.

c. For forward and reverse mode differentiation, state which intermediate derivatives are computed at each step, and their computational and memory costs.

7 Lecture 4: Probabilistic Modelling Principles

Question 28 (Training translation models)**.** Imagine you want to train a neural network $T_{\boldsymbol{\theta}}(\cdot)$ to translate French words to English words. Assume you are given a dataset $\mathcal{D} = \{(f_n, e_n)\}_{n=1}^N$ where f_n is a French word and e_n is an English word. Suppose the vocabulary of French and English is $\mathcal{F}$ and $\mathcal{E}$, respectively.

a. Assuming a probabilistic model $p(e|T_{\boldsymbol{\theta}}(f))$, which distribution would you choose for this model?

b. Continuing a), what is the corresponding MLE objective?

Question 29 (Clustering)**.** We consider a clustering task where given a dataset $\mathcal{D} = \{x_1, ..., x_N\}$, we would like to group them into K clusters. The model we will use here is a Gaussian mixture model:

$$\text{GMM:} \quad p(x|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \sigma^2), \quad \boldsymbol{\theta} = \{\pi_k, \mu_k, \}_{k=1}^K, \sigma^2.$$

a. What is the MLE objective for this clustering task?

b. Derive the gradient of the MLE objective w.r.t. μ_k. What is the fixed-point equation for finding the optimal $\{\mu_k\}$ parameters?

Question 30 (Geometric interpretation of linear regression)**.** Consider the following linear regression model:

$$y = \boldsymbol{\theta}^\top \phi(\mathbf{x}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

For a given dataset $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$, Writing $\boldsymbol{\Phi} = (\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), ..., \phi(\mathbf{x}_N))^\top$ and $\mathbf{y} = (y_1, ..., y_N)^\top$, we have the optimal solution satisfies $\boldsymbol{\theta}^* = (\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \mathbf{y}$. Show that by using the optimal parameter $\boldsymbol{\theta}^*$, the prediction $\hat{\mathbf{y}} = (\hat{y}_1, ..., \hat{y}_N), \hat{y}_n = (\boldsymbol{\theta}^*)^\top \phi(\mathbf{x}_n)$ is the projection of $\mathbf{y}$ onto the sub-space spanned by the columns of Φ.

(Hint: consider singular value decomposition.)

8 Lecture 5: Gradient Descent Convergence

Question 31 (Rayleigh quotient)**.** The *Rayleigh quotient* is defined for a symmetric matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ and a non-zero vector $\mathbf{x} \in \mathbb{R}^{d\times 1}$:

$$R(\mathbf{A}, \mathbf{x}) = \frac{\mathbf{x}^\top \mathbf{A}\mathbf{x}}{||\mathbf{x}||_2^2}, \quad ||\mathbf{x}||_2^2 = \mathbf{x}^\top \mathbf{x}.$$

Show that $R(\mathbf{A}, \mathbf{x}) \in [\lambda_{min}(\mathbf{A}), \lambda_{max}(\mathbf{A})]$.

This result immediately indicates that $\lambda_{min}(\mathbf{A})||\mathbf{x}||_2^2 \leq \mathbf{x}^\top \mathbf{A}\mathbf{x} \leq \lambda_{max}(\mathbf{A})||\mathbf{x}||_2^2$, which is used to prove gradient descent convergence.

Question 32 (Gradient descent with pre-conditioning)**.** Consider the following update rule named *pre-conditioned gradient descent*:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_t \mathbf{P}_t^{-1} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t).$$

Here $\mathbf{P}_t$ is called *pre-conditioner* at time step t. We consider linear regression as an example, and assume constant learning rate and pre-conditioner, i.e., $\gamma_t = \gamma$ and $\mathbf{P}_t = \mathbf{P}$ for all t. Show that with an appropriate choice of the pre-conditioner $\mathbf{P}$, we can achieve a robust selection of the learning rate γ, i.e., if the selected γ works for an initialisation $\boldsymbol{\theta}_0$, it will also work for all other initialisations.

Hints: you can follow the below steps to solve the question:

1. Work out the pre-conditioned gradient descent update in linear regression, and derive $\boldsymbol{\theta}_t$ as a function of $\boldsymbol{\theta}_0$, γ, $\mathbf{P}$ and the dataset $(\mathbf{X}, \mathbf{y})$;
2. For a given $\mathbf{P}$, work out the learning rates γ_{min} and γ_{max} such that pre-conditioned gradient descent converges when $\gamma < \gamma_{min}$, or diverges when $\gamma \geq \gamma_{max}$;
3. Select $\mathbf{P}$ such that $\gamma_{min} = \gamma_{max}$, therefore there exist no interval (like $[\gamma_{min}, \gamma_{max})$) such that convergence depends on initialisation when γ falls into such interval.

Question 33 (Momentum gradient descent)**.** Consider the following update rule named *momemtum gradient descent*, with constant learning rate γ and momentum step-size α:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t) + \alpha \Delta\boldsymbol{\theta}_t,$$

$$\Delta\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t, \quad \Delta\boldsymbol{\theta}_0 = \mathbf{0}.$$

Show that solving linear regression using momemtum gradient descent, if converges, converges to $\boldsymbol{\theta}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$.

Hint: follow the below steps and practice your linear algebra skills :)

1. Write down the update equations for the parameters $\boldsymbol{\theta}_t$ and the momentum $\Delta\boldsymbol{\theta}_t$;
2. Collect both terms as a long vector $(\boldsymbol{\theta}_t^\top, \Delta\boldsymbol{\theta}_t^\top)^\top$, and merge the two linear update equations in step 1 into one "joint" linear equation using block matrices;
3. Apply the analysis techniques for gradient descent convergence for linear regression to show the converged solution (if converges).

9 Lecture N: Multivariate Probability

Question 34 (Vector independence)**.** While you can probabily figure this one out already, we will discuss this in more detail later. Consider the density on $\mathbf{x} \in \mathbb{R}^4$ with all elements $0 \leq x_k \leq 1$ as

$$p(\mathbf{x}) = \mathbf{x}^\intercal \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix} \mathbf{x}\,. \tag{16}$$

a. Rewrite the density in terms of $\tilde{\mathbf{x}} = [x_1, x_3, x_2, x_4]^\intercal$. Note that you can do this by a substitution $\mathbf{x} = \mathrm{P}\tilde{\mathbf{x}}$, where P is a permutation matrix. You will see that you just need to swap the relevant rows and columns of the matrix. However, make sure that you understand the mathematical steps that really show this.

b. Divide up $\mathbf{x}$ into two sub vectors $\mathbf{y} = [x_2, x_4]^\intercal$ and $\mathbf{z} = [x_1, x_3]^\intercal$. Show that $\mathbf{y} \perp\!\!\!\perp \mathbf{z}$, i.e. that they are independent.

10 Warm-up Exercises Answers

10.1 Warm-up Exercises

Question 1 – Set Theory and Probability

a. We can choose any representation denoting the events, e.g. using abstract symbols $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$. Alternatively, we can represent each of the outcomes as a number $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Following the latter notation, $A = \{3, 4\}$, and $A = \{1, 2, 5, 6\}$.

b. Length problem with sample space $\Omega = [0, 1]$.

$\neg A = [0, 0.3] \cup (0.4, 1]$

c. $P(\neg A) = 1 - P(A)$

Since $\neg A$ and A are mutually exclusive: $A \cup \neg A = \Omega$ and $A \cap \neg A = \varnothing$.

By combining axiom 2 and 3: $P(A) + P(\neg A) = P(A \cup \neg A) = P(\Omega) = 1$

Thus: $P(\neg A) = 1 - P(A)$

d. $P(\varnothing) = 0$, where $\varnothing$ is the empty set

Given the sample space, Ω, its complementary is the empty set $\varnothing$.

We use property (c) and axiom 2: $P(\varnothing) = 1 - P(\Omega) = 1 - 1 = 0$.

e. $0 \leq P(A) \leq 1$

We use property (c) and axiom 1.

Consider an event A, where $P(A) \geq 0$ and $P(\neg A) \geq 0$ by axiom 1.

Then, $P(\neg A) = 1 - P(A) \geq 0 \implies 1 \geq P(A)$.

By joining both inequalities, $0 \leq P(A) \leq 1$.

f. $A \subseteq B \implies P(A) \leq P(B)$

Hint: Consider the following definition. $B \backslash A = \{x \in B : x \notin A\}$

Assume $A \subseteq B$ and construct B as the union of two disjoint sets: $B = B \backslash A \cup A$.

Then, $B \backslash A \cap A = \varnothing$ by definition of $B \backslash A$. By axiom 1, we have $P(B \backslash A) \geq 0$.

Use axiom 3: $P(B) = P(B \backslash A) + P(A) \geq P(A) \implies P(A) \leq P(B)$.

g. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Define the union $(A \cup B)$ in terms of two disjoint sets. $(A \cup B) = A \cup B \backslash A$, where $A \cap B \backslash A = \varnothing$.

Use axiom 3: $P(A \cup B) = P(A) + P(B \backslash A)$.

To compute $P(B \backslash A)$, we define B in terms of A, and the union of two disjoint sets: $B = (B \cap A) \cup (B \backslash A)$, where $(B \cap A) \cap (B \backslash A) = \varnothing$ by definition.

Use axiom 3 again: $P(B) = P(B \cap A) + P(B \backslash A) \implies P(B \backslash A) = P(B) - P(B \cap A)$.

Finally: $P(A \cup B) = P(A) + P(B \backslash A) = P(A) + P(B) - P(B \cap A)$.

h. (*) if $\{A_i\}_{i=1}^{\infty} \subseteq \Omega$ and $A_{i-1} \subseteq A_i \quad \forall i > 0$ then:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i \to \infty} P(A_i)$$

Hint: Use axiom 3.

Let us define the following: $A := \bigcup_{i=1}^{\infty} A_i$. We would like to write A in terms of disjoint sets to use axiom 3.

$$A_{i-1} \subseteq A_i \quad \forall i > 0 \implies A = \bigcup_{i=1}^{\infty} A_i \backslash A_{i-1} \tag{17}$$

where the expression holds if we have $A_0 = \varnothing$. We regard 17 as starting with A_1 and adding the new information from $A_2, A_3, \dots$ (e.g $A_2 \backslash A_1, A_3 \backslash A_2, \dots$).

$$P(A) = P\left(\bigcup_{i=1}^{\infty} A_i \backslash A_{i-1}\right) = \sum_{i=1}^{\infty} P(A_i \backslash A_{i-1}) \qquad \text{(by axiom 3)} \qquad (18)$$

$$P(A) = \sum_{i=1}^{\infty} P(A_i \backslash A_{i-1}) = \lim_{n \to \infty} \sum_{i=1}^{n} P(A_i \backslash A_{i-1}) \qquad \text{(the infinite summation is a limit)} \qquad (19)$$

From (f), we have $P(A_i) = P(A_i \backslash A_{i-1}) + P(A_{i-1}) \implies P(A_i \backslash A_{i-1}) = P(A_i) - P(A_{i-1})$. Then,

$$P(A) = \lim_{n \to \infty} \sum_{i=1}^{n} P(A_i) - P(A_{i-1}) = \lim_{n \to \infty} \left(\sum_{i=1}^{n} P(A_i) - \sum_{i=1}^{n-1} P(A_i) \right) = \lim_{n \to \infty} P(A_n) \qquad (20)$$

where we used $P(A_0) = P(\varnothing) = 0$ from (d).

In summary:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P(A) = \lim_{i \to \infty} P(A_i) \qquad (21)$$

Question 7 – Statistics Terminology

a. A statistic is a function that is computed from data. For example, take a data set $X = \{x_1, x_2, x_3, \dots\}$ where we compute the empirical mean $\bar{X} = \frac{1}{|X|} \sum_n x_n$.

b. An estimator is a function of data that tries to estimate an unknown quantity. Estimators are statistics. Some statistics are also estimators. For example, if we have some data set from that is sampled from some unknown density $p(x)$, then its mean is unknown, and $\bar{X}$ is an estimator of it.

c. A consistent estimator finds the correct value of the unknown quantity if the dataset grows to infinity. We will prove that $\bar{X}$ is a consistent estimate of $\int p(x) x \mathrm{d}x$ later on in the course.

d. A sample from a random variable is an outcome of the random experiment it represents. For example, you can have a random variable representing the outcome of a coin toss. A sample from it would be heads or tails. We sampled a random variable independently many times, then the outcomes would occur with the frequency specified by the probability distribution of the random variable. Thinking about sampling outcomes from a random variable is often a helpful conceptual technique to think about randomness.

10.2 Linear Algebra

Question 8 $\mathbf{x}^\top \mathbf{y} = 1 \times 0 + (-2) \times 4 + 5 \times (-3) + (-1) \times 7 = 0 + (-8) + (-15) + (-7) = -30$.

Question 9 $\mathbf{y} = (24, -14, -12)^\top$, $||\mathbf{x}||_2 = \sqrt{23}$, $||\mathbf{y}||_2 = \sqrt{916}$.
Note that by definition the ℓ_2 norm of a vector is $||\mathbf{x}||_2 = \sqrt{\mathbf{x}^\top \mathbf{x}}$.

Question 10 1, 2.
A set of vectors $\{\mathbf{b}_1, ..., \mathbf{b}_K\}$ with $\mathbf{b}_k \in \mathbb{R}^d$ can form a basis of $\mathbb{R}^d$ iff $K = d$ the vectors are linearly independent to each other.

Question 11 2, 5.
A point $\mathbf{x} \in \mathbb{R}^d$ is in $span(\{\mathbf{b}_1, ..., \mathbf{b}_K\})$ with $\mathbf{b}_k \in \mathbb{R}^d$ iff we can find $a_1, ..., a_K \in \mathbb{R}$ such that $\mathbf{x} = \sum_{k=1}^{K} a_k \mathbf{b}_k$.

Question 12 The rotation matrix is

$$\begin{pmatrix} \cos\frac{\pi}{4} & -\sin\frac{\pi}{4} \\ \sin\frac{\pi}{4} & \cos\frac{\pi}{4} \end{pmatrix}.$$

Question 13 a) The matrix A and vector $\mathbf{b}$ are

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 3 & 2 & 4 \\ -2 & 1 & -2 \end{pmatrix}, \quad \mathbf{b} = (2, 5, 1)^\top.$$

b) The inverse of A is

$$A^{-1} = \begin{pmatrix} 2/3 & -1/3 & -2/3 \\ 1/6 & 1/6 & 1/3 \\ 7/12 & 5/12 & 1/3 \end{pmatrix}.$$

Therefore we have $\mathbf{x} = A^{-1}\mathbf{b} = (-1, 3/2, 43/12)^\top$.

c) $\text{rank}(A) = 3$: as A is invertible, it must have full rank.

Question 14 a) When A is symmetric, then $A = Q\Lambda Q^\top$, and $\mathbf{x}^\top A\mathbf{x} = \mathbf{x}^\top Q\Lambda Q^\top \mathbf{x} = (Q^\top \mathbf{x})^\top \Lambda (Q^\top \mathbf{x})$. As Q is an orthonormal matrix, we have $\mathbf{x} \to Q^\top \mathbf{x}$ a one-to-one mapping. Therefore we have

$$\mathbf{x}^\top A\mathbf{x} = \boldsymbol{z}^\top \Lambda \boldsymbol{z} = \sum_{i=1}^{d} \lambda_i z_i^2, \quad \boldsymbol{z} = (z_1, ..., z_d)^\top = Q^\top \mathbf{x}.$$

Therefore $\mathbf{x}^\top A\mathbf{x} \geq 0 \Leftrightarrow \sum_{i=1}^{d} \lambda_i z_i^2 \geq 0$. This is true for any $\mathbf{x} \in \mathbb{R}^{d\times 1}$ if and only if $\lambda_i \geq 0$ for all $i = 1, ..., d$.

b) We use the permuation invariance property of matrix trace to show the result:

$$Tr(A) = Tr(Q\Lambda Q^{-1}) = Tr(Q^{-1}Q\Lambda) = Tr(\Lambda) = \sum_{i=1}^{d} \lambda_i.$$

c) We use the product rule of matrix determinant to show the result:

$$det(A) = det(Q\Lambda Q^{-1}) = det(Q)det(\Lambda)det(Q^{-1}) = det(Q)det(\Lambda)det(Q)^{-1} = det(\Lambda) = \prod_{i=1}^{d} \lambda_i.$$

d) Let us assume the statement is false, i.e., there exists a solution $\lambda^* \neq \lambda_i, \forall i = 1, ..., d$ for the equation $A\boldsymbol{q} = \lambda \boldsymbol{q}, \boldsymbol{q} \neq 0$. Then we can rewrite the equation as

$$A\boldsymbol{q} = \lambda^* \boldsymbol{q} \quad \Rightarrow \quad (A - \lambda^* I)\boldsymbol{q} = \mathbf{0} \quad \Rightarrow \quad Q(\Lambda - \lambda^* I)Q^{-1}\boldsymbol{q} = 0.$$

By definition, the column vectors of Q forms a basis of $\mathbb{R}^d$. Notice that the diagonal entries of $\Lambda - \lambda^* I$ are non-zero as we assume $\lambda^* \neq \lambda_i$. This indicates a contradiction to the assumption of $\boldsymbol{q} \neq 0$:

$$Q(\Lambda - \lambda^* I)Q^{-1}\boldsymbol{q} = 0 \quad \Rightarrow \quad Q^{-1}\boldsymbol{q} = \mathbf{0} \quad \Rightarrow \quad \boldsymbol{q} = \mathbf{0}.$$

11 Answers Lecture 2: Vector Differentiation

Question 20 – Circle Answer discussed in lectures.

Question 25 – Hessian of Linear Regression The objective function and gradient w.r.t. $\boldsymbol{\theta}$ (see lectures) for Linear Regression is

$$L(\boldsymbol{\theta}) = |\mathbf{y} - \Phi(X)\boldsymbol{\theta}|^2, \qquad \frac{\mathrm{d}L}{\mathrm{d}\boldsymbol{\theta}} = 2(\Phi(X)\boldsymbol{\theta} - \mathbf{y})^\intercal \Phi(X). \tag{22}$$

We begin by finding the Hessian, i.e. the matrix containing all second partial derivatives. We need to do this in index notation, as the vector conventions of our vector chain rule break down. So we first write

the derivative in index notation, and then we take the derivative again, after which we return to vector notation:

$$\frac{\partial}{\partial\theta_j}\left(\frac{\partial L}{\partial\theta_i}\right) = \frac{\partial}{\partial\theta_j}\left(2\sum_k\left(\sum_m \Phi_{km}\theta_m - y_k\right)\Phi_{ki}\right) = \frac{\partial}{\partial\theta_j}\left(2\sum_k\left(\sum_m \Phi_{km}\theta_m - y_k\right)\Phi_{ki}\right) \tag{23}$$

$$= 2\sum_{km}\Phi_{km}\delta_{mj}\Phi_{ki} = 2\sum_k \Phi_{kj}\Phi_{ki}\,, \tag{24}$$

$$\implies \mathrm{H}_{\boldsymbol{\theta}}(L) = 2\Phi(X)^\intercal\Phi(X)\,. \tag{25}$$

The Hessian doesn't depend on the parameter $\boldsymbol{\theta}$, so if we prove that the matrix is positive definite, then the local where $\frac{\mathrm{d}L}{\mathrm{d}\theta} = 0$ (see lecture slides) will be a minimum. For a matrix to be PD, we need $\mathbf{v}^\intercal\mathrm{H}\mathbf{v} > 0$ for all $\mathbf{v}$. We substitute our Hessian into H to prove this

$$\mathbf{v}^\intercal\mathrm{H}\mathbf{v} = 2\mathbf{v}^\intercal\Phi(X)^\intercal\Phi(X)\mathbf{v} \tag{26}$$

$$= \mathbf{w}^\intercal\mathbf{w} = \sum_i w_i^2\,, \qquad \text{with } \mathbf{v} = \Phi(X)\mathbf{v}\,. \tag{27}$$

This already shows that $\mathbf{v}^\intercal\mathrm{H}\mathbf{v} \geq 0$, with equality if there exists a $\mathbf{v}$ such that $\Phi(X)\mathbf{v} = 0$. So now we need to prove that *there cannot be* a $\mathbf{v}$ for which $\Phi(X)\mathbf{v} = 0$. If $\operatorname{rank}\Phi(X) \geq M$, then this will not happen, by the rank-nullity theorem [**mml**].

At this point, we need to assume this is the case. For full marks though, you should state the implications on the problem at hand, rather than in abstract maths. One *necessary* implication of this is that $N \geq M$. This is only a necessary condition, rather than a sufficient one, since even of $N \geq M$, $\Phi(X)$ can still have many linearly dependent rows. This will at least happen if you observe repeated input points. However, to prove more than this, you need more information about $\Phi(X)$.[1]

So to summarise, we could prove that **if** $\operatorname{rank},\Phi(X) \geq M$**, which at least needs** $N \geq M$**, then Linear Regression has a single minimum solution**.

If we are coding up a linear regression problem, and we want to check numerically for a *specific* regression problem whether there is a unique solution, we can compute the eigenvalues of $\Phi(X)^\intercal\Phi(X)$, and see if they are all positive. This implies a PD Hessian because

$$\mathbf{v}^\intercal\mathrm{H}\mathbf{v} = \mathbf{v}^\intercal\mathrm{Q}\Lambda\mathrm{Q}^{-1}\mathbf{v} \qquad \text{(eigenvalue decomposition)} \tag{28}$$

$$= \mathbf{v}^\intercal\mathrm{Q}\Lambda\mathrm{Q}^\intercal\mathbf{v} \qquad (\mathrm{H} = \mathrm{H}^\intercal\text{, so } \mathrm{Q}\Lambda\mathrm{Q}^{-1} = (\mathrm{Q}\Lambda\mathrm{Q}^{-1})^\intercal\text{, so } \mathrm{Q}^{-1} = Q^\intercal) \tag{29}$$

$$= \mathbf{z}^\intercal\Lambda\mathbf{z}\,, \tag{30}$$

which is only > 0 if all the elements in the diagonal matrix Λ are positive.

If any of the linear algebra was unfamiliar, I recommend looking at chapter 2 in **mml**, particularly §2.3, §2.6, and §2.7, or the 1st year linear algebra course.

12 Answers Lecture 3: Automatic Differentiation

Question 26 – Product rule We begin by drawing the computational graph (fig. 1). We now find the primal trace and the forward tangent trace:

$$v_0 = x \qquad \frac{\partial v_0}{\partial x} = 1 \tag{31}$$

$$v_1 = a(x) \qquad \frac{\partial v_1}{\partial x} = \frac{\partial v_1}{\partial v_0}\frac{\partial v_0}{\partial x} = \frac{\partial a(x)}{\partial x} \tag{32}$$

$$v_2 = b(x) \qquad \frac{\partial v_2}{\partial x} = \frac{\partial v_2}{\partial v_0}\frac{\partial v_0}{\partial x} = \frac{\partial b(x)}{\partial x} \tag{33}$$

$$v_3 = v_1 \cdot v_2 \qquad \frac{\partial v_3}{\partial x} = \sum_{j\in\text{inputs}(3)}\frac{\partial v_3}{\partial v_j}\frac{\partial v_j}{\partial x} = v_2\frac{\partial a(x)}{\partial x} + v_1\frac{\partial b(x)}{\partial x}\,. \tag{34}$$

This means that for any x, forward mode autodiff calculates the derivative to be:

$$\frac{\mathrm{d}f}{\mathrm{d}x} = b(x)\frac{\mathrm{d}a(x)}{\mathrm{d}x} + a(x)\frac{\mathrm{d}b(x)}{\mathrm{d}x}\,. \tag{35}$$

[1] A case that is harder think about is if you observe points that make the feature vectors $\boldsymbol{\phi}(\mathbf{x}_n)$ linearly dependent. One example is if you have a 2D input with $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}^\intercal$, and all your input points lie on a line.

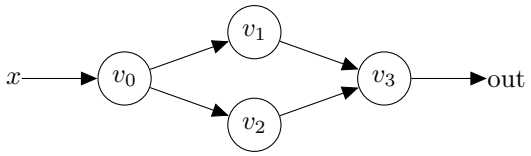

Figure 1: Computational graph for Question 26 – Product rule.

Which is the product rule.

If we substitute $a(x) = x, b(x) = x$, then we obtain $\mathrm{d}f/\mathrm{d}x = 2x$, as expected.

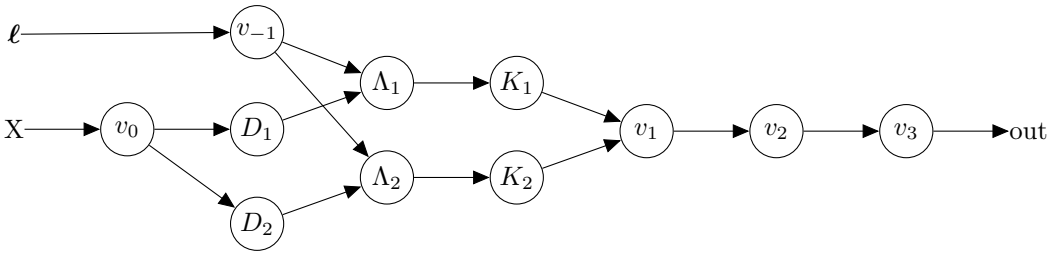

Figure 2: Computational graph for Question 27 – Multivariate Autodiff, where we define $v_1 = \mathrm{K}_1 + \mathrm{K}_2$, $v_2 = v_1^{-1}$, and $v_3 = \mathbf{y}^\intercal v_1 \mathbf{y}$.

Question 27 – Multivariate Autodiff

a.

$$D_a : \mathbb{R}^{N\times 2} \to \mathbb{R}^{N\times N} \tag{36}$$
$$\Lambda_a : \mathbb{R}^{N\times N} \times \mathbb{R}^2 \to \mathbb{R}^{N\times N} \tag{37}$$
$$K_a : \mathbb{R}^{N\times N} \to \mathbb{R}^{N\times N} \tag{38}$$
$$f : \mathbb{R}^{N\times N} \times \mathbb{R}^{N\times N} \to \mathbb{R} \tag{39}$$

b. See fig. 2. We use names evident from the question for some nodes, but give new names to some additional intermediate notes.

c. Let's first consider **forward mode** for the derivatives w.r.t. X.

$$v_{-1} = \boldsymbol{\ell} \qquad \dot{v}_{-1,iab} = \frac{\partial [v_{-1}]_i}{\partial X_{ab}} = 0 \qquad v_{-1} \in \mathbb{R}^{2\times(N\times 2)}, O(N) \tag{40}$$

$$v_0 = X \qquad \dot{v}_0 = \frac{\partial [v_0]_{ij}}{\partial X_{ab}} = \delta_{ia}\delta_{jb} \qquad \dot{v}_0 \in \mathbb{R}^{(N\times 2)\times(N\times 2)}, O(N^2) \tag{41}$$

$$D_z = \dots \qquad \frac{\partial [D_z]_{nm}}{\partial v_{0ij}} = \frac{\partial}{\partial v_{0ij}}(v_{0nz} - v_{0mz})^2 \qquad \dot{D}_z \in \mathbb{R}^{(N\times N)\times(N\times 2)}$$
$$= 2(v_{0nz} - v_{0mz})(\delta_{ni}\delta_{zj} - \delta_{mi}\delta_{zj}) \tag{42}$$
$$\dot{D}_{znmab} = \left[\frac{\partial D_z}{\partial v_0}\dot{v}_0\right]_{nmab} \qquad O(N) \text{ for sum, so total } O(N^4). \tag{43}$$
$$= 2(v_{0nz} - v_{0mz})(\dot{v}_{0nzab} - \dot{v}_{0mzab}) \qquad \text{Structure allows } O(N^3). \tag{44}$$

$$\Lambda_z = -\frac{D_z}{2v_{-1z}^2} \qquad \frac{\partial \Lambda_{zij}}{\partial D_{znm}} = -\frac{D_z}{2v_{-1z}^2}\delta_{in}\delta_{jm} \qquad \dot{\Lambda}_z \in \mathbb{R}^{(N\times N)\times(N\times 2)} \tag{45}$$
$$\frac{\partial \Lambda_{zij}}{\partial v_{-1,k}} = \frac{D_{zij}}{v_{-1z}^3}\delta_{zk} \tag{46}$$
$$\dot{\Lambda}_{zijab} = \left[\frac{\partial \Lambda_z}{\partial D_z}\dot{D}_z + \frac{\partial \Lambda_z}{\partial v_{-1}}\dot{v}_{-1}\right]_{ijab} \qquad O(N^2) \text{ for sum, so total } O(N^5). \tag{47}$$

d. To be continued...

13 Answers Lecture 4: Probabilistic Modelling Principles

Question 28

a. Choose a categorical distribution. Let $T_{\boldsymbol{\theta}}(\cdot) : \mathcal{F} \to \mathbb{R}^{|\mathcal{E}|}$ maps a French word f to a real-value vector of length $|\mathcal{E}|$. Then define

$$p(e|T_{\boldsymbol{\theta}}(f)) = \text{Categorical}(softmax(T_{\boldsymbol{\theta}}(f))).$$

b. The MLE objective is (if we write e_n using one-hot encoding: $e_n = (0, ..., 0, 1, 0, ..., 0)$)

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{n=1}^{N} \log p(e_n|T_{\boldsymbol{\theta}}(f_n)) = \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{n=1}^{N} \log \frac{\exp[T_{\boldsymbol{\theta}}(f_n)]^\top e_n}{\sum_{i=1}^{|\mathcal{E}|} \exp[T_{\boldsymbol{\theta}}(f_n)_i]}$$

Question 29

a. With i.i.d. assumption, the MLE objective is:

$$L(\boldsymbol{\theta}) = \frac{1}{N} \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n; \mu_k, \sigma^2).$$

b. The gradient of the MLE objective w.r.t. μ_k is

$$\nabla_{\mu_k} L(\boldsymbol{\theta}) = \frac{1}{N} \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n; \mu_k, \sigma^2)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n; \mu_j, \sigma^2)} \frac{x_n - \mu_k}{\sigma^2}.$$

Setting $\nabla_k L(\boldsymbol{\theta}) = 0$ for all k, we have the fixed-point equation as

$$\mu_k = \frac{1}{N} \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n; \mu_k, \sigma^2)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n; \mu_j, \sigma^2)} x_n.$$

Question 30 Using matrix-vector notations we have $\hat{\mathbf{y}} = \boldsymbol{\Phi}\boldsymbol{\theta}^* = \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\mathbf{y}$. Writing the SVD of $\boldsymbol{\Phi} = \mathbf{U}\Sigma\mathbf{V}^\top$, it is easy to show that $\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top = \mathbf{U}\mathbf{U}^\top$. Notice that $\mathbf{U}$ contains basis vectors which span to the same subspace spanned by the column vectors of $\boldsymbol{\Phi}$.

14 Answers Lecture 5: Gradient Descent Convergence

Question 31 As $\mathbf{A}$ is symmetric, we can write the eigen-decomposition formula as $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^\top$ with $\mathbf{Q}$ containing an orthonormal basis of $\mathbf{R}^{d\times 1}$. Then using the fact that $\mathbf{Q}\mathbf{Q}^\top = \mathbf{I}$, we can define $\mathbf{z} = \mathbf{Q}^\top\mathbf{x}$ and rewrite the Rayleigh quotient as:

$$R(\mathbf{A}, \mathbf{x}) = \frac{\mathbf{x}^\top\mathbf{Q}\Lambda\mathbf{Q}^\top\mathbf{x}}{\mathbf{x}^\top\mathbf{Q}\mathbf{Q}^\top\mathbf{x}} = \frac{\mathbf{z}^\top\Lambda\mathbf{z}}{\mathbf{z}^\top\mathbf{z}}. \tag{48}$$

As $\Lambda = \text{diag}(\lambda_1, ..., \lambda_d)$ is a diagonal matrix, we have (writing $\mathbf{z} = (z_1, ..., z_d)^\top$)

$$\mathbf{z}^\top\Lambda\mathbf{z} = \sum_{i=1}^{d} \lambda_i z_i^2. \tag{49}$$

Therefore the Rayleigh quotient can be written as the following weighted average of the eigenvalues

$$R(\mathbf{A}, \mathbf{x}) = \sum_{i=1}^{d} \frac{z_i^2}{||\mathbf{z}||_2^2} \lambda_i, \quad \text{with} \sum_{i=1}^{d} \frac{z_i^2}{||\mathbf{z}||_2^2} = 1. \tag{50}$$

In summary, these derivations indicate that the Rayleigh quotient is bounded as

$$\begin{aligned} & \lambda_{min}(\mathbf{A}) \leq R(\mathbf{A}, \mathbf{x}) \leq \lambda_{max}(\mathbf{A}) \\ \Rightarrow \quad & \lambda_{min}(\mathbf{A})||\mathbf{x}||_2^2 \leq \mathbf{x}^\top\mathbf{A}\mathbf{x} \leq \lambda_{max}(\mathbf{A})||\mathbf{x}||_2^2, \end{aligned} \tag{51}$$

where $\lambda_{min}(\mathbf{A})$ and $\lambda_{max}(\mathbf{A})$ are the smallest and largest eigenvalues of $\mathbf{A}$, respectively.

Question 32 For linear regression we have $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \mathbf{X}^\top(\mathbf{X}\boldsymbol{\theta}_t - \mathbf{y})$. Therefore in this case the pre-conditioned gradient descent update rule is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\gamma}{\sigma^2}\mathbf{P}^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\theta}_t - \mathbf{y}).$$

Under the formula of Arithmetico-geometric sequence, we have

$$\boldsymbol{\theta}_t = (\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{P}^{-1}\mathbf{X}^\top\mathbf{X})^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*) + \boldsymbol{\theta}^*, \quad \boldsymbol{\theta}^* = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

So this means if pre-conditioned gradient descent converges, it will converge to the right answer $\boldsymbol{\theta}^*$.

Now we need to show that pre-conditioned gradient descent converges with the right choices of γ and $\mathbf{P}$. This requires us to analyse the eigenvalues of the matrix $(\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{P}^{-1}\mathbf{X}^\top\mathbf{X})^2$, and for a given $\mathbf{P}$:

$$\lambda \text{ is an eigenvalue of } \mathbf{P}^{-1}\mathbf{X}^\top\mathbf{X} \quad \Rightarrow \quad (1 - \frac{\gamma}{\sigma^2})^2 \text{ is an eigenvalue of } (\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{P}^{-1}\mathbf{X}^\top\mathbf{X})^2.$$

Following the same idea as to prove convergence for gradient descent, we have the learning rates bounds

$$\gamma_{min} = 2\sigma^2/\lambda_{max}(\mathbf{P}^{-1}\mathbf{X}^\top\mathbf{X}), \quad \gamma_{max} = 2\sigma^2/\lambda_{min}(\mathbf{P}^{-1}\mathbf{X}^\top\mathbf{X}).$$

Now to set $\gamma_{min} = \gamma_{max}$, we should choose $\mathbf{P}$ such that $\lambda_{min}(\mathbf{P}^{-1}\mathbf{X}^\top\mathbf{X}) = \lambda_{max}(\mathbf{P}^{-1}\mathbf{X}^\top\mathbf{X})$, and an easy way to do so is to choose $\mathbf{P} \propto \mathbf{X}^\top\mathbf{X}$.

One can show that for linear regression problems, the Hessian matrix of $L(\boldsymbol{\theta})$ is $\nabla^2_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \propto \mathbf{X}^\top\mathbf{X}$. In general for a given loss function $L(\boldsymbol{\theta})$ we often set $\mathbf{P}_t = \nabla^2_{\boldsymbol{\theta}_t} L(\boldsymbol{\theta}_t)$ if it can be computed in a fast way.

Question 33 First note that $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t) = \frac{1}{\sigma^2}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\theta}_t - \mathbf{y})$ The update equations for both the parameter and the momentum are

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \gamma\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t) + \alpha\Delta\boldsymbol{\theta}_t = \boldsymbol{\theta}_t - \frac{\gamma}{\sigma^2}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\theta}_t - \mathbf{y}) + \alpha\Delta\boldsymbol{\theta}_t \\ \Delta\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t = \alpha\Delta\boldsymbol{\theta}_t - \frac{\gamma}{\sigma^2}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\theta}_t - \mathbf{y}). \end{aligned} \tag{52}$$

Now collecting both equations together into a "joint" linear equation:

$$\begin{bmatrix} \boldsymbol{\theta}_{t+1} \\ \Delta\boldsymbol{\theta}_{t+1} \end{bmatrix} = \begin{bmatrix} \mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X} & \alpha\mathbf{I} \\ -\frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X} & \alpha\mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_t \\ \Delta\boldsymbol{\theta}_t \end{bmatrix} + \begin{bmatrix} \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{y} \\ \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{y} \end{bmatrix}. \tag{53}$$

Then we can apply the derivation of arithmetico–geometric sequences again, and show that

$$\begin{bmatrix} \boldsymbol{\theta}_t \\ \Delta\boldsymbol{\theta}_t \end{bmatrix} = \begin{bmatrix} \mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X} & \alpha\mathbf{I} \\ -\frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X} & \alpha\mathbf{I} \end{bmatrix}^t \begin{bmatrix} \boldsymbol{\theta}_0 - \boldsymbol{\theta}^* \\ \Delta\boldsymbol{\theta}_0 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\theta}^* \\ \mathbf{0} \end{bmatrix}, \tag{54}$$

with $\boldsymbol{\theta}^* = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. This equation also says if momentum GD converges, the momentum $\Delta\boldsymbol{\theta}_t$ will vanish to $\mathbf{0}$, which is as expected as $\Delta\boldsymbol{\theta}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1} \to \mathbf{0}$.