70015 Mathematics for Machine Learning: Exercises

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

December 9, 2022

Contents

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

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.

- 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

- Expectation identities for $X \in \mathbb{R}^D, Y \in \mathbb{R}^E$.

$$\mathbb{E}_X[AX] = \mathbb{E}_X[AX] = A\mathbb{E}_X[X] \tag{7}$$

$$\mathbb{V}_X[X] = \mathbb{E}_X[XX^\intercal] - \mathbb{E}_X[X]\mathbb{E}_X[X]^\intercal \tag{8}$$

$$\mathbb{V}_X[AX] = A\mathbb{V}_X[X]A^\intercal \tag{9}$$

$$\mathbb{C}_{X,Y}[X, Y] = \mathbb{E}_{X,Y}[XY^\intercal] - \mathbb{E}_X[X]\mathbb{E}_Y[Y]^\intercal \tag{10}$$

$$\mathbb{C}_{X,Y}[X, Y] = \mathbb{C}_{X,Y}[Y, X]^\intercal \tag{11}$$

$$\mathbb{C}_{X,Y}[X, Y] = 0 \qquad \text{for } X \perp\!\!\!\perp Y \tag{12}$$

$$\mathbb{E}_{X,Y}[X + Y] = \mathbb{E}_X[X] + \mathbb{E}_Y[Y] \tag{13}$$

$$\mathbb{V}_{X,Y}[X + Y] = \mathbb{V}_X[X] + \mathbb{V}_Y[Y] \qquad \text{for } X \perp\!\!\!\perp Y \tag{14}$$

$$\mathbb{E}_{X,Y}[XY^\intercal] = \mathbb{E}_X[X]\mathbb{E}_Y[Y]^\intercal \qquad \text{for } X \perp\!\!\!\perp Y \tag{15}$$

- For a RV $X \geq 0$, then $\mathbb{E}_X[X] = 0$ if and only if $P(X = 0) = 1$.
- Markov's inequality: For a RV $X \geq 0$, and $a > 0$, then $P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$.
- Chebyshev's inequality: For a RV X with finite mean, and finite non-zero variance, then for any $k > 0$, $P(|X - \mathbb{E}[X]| \geq k\sigma) \leq \frac{1}{k^2}$.
- Weak Law of Large Numbers: For a series of iid RVs X_n, we have for any $\epsilon > 0$ that

$$\lim_{n\to\infty} P(|\frac{1}{N}\sum_{n=1} X_n - \mu| < \epsilon) = 1\,. \tag{16}$$

- Law of the Unconscious Statistician: If $Y = g(X)$ then $\mathbb{E}_Y[Y] = \mathbb{E}_X[g(X)]$.
- Gaussian probability density function (pdf) with input $\mathbf{x} \in \mathbb{R}^D$, denoted as $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right). \tag{17}$$

- For a joint Gaussian density

$$p\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}; \begin{bmatrix}\mathbf{m_x}\\ \mathbf{m_y}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}}\\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}}\end{bmatrix}\right), \tag{18}$$

  we have the conditional density

$$p(\mathbf{x}\,|\,\mathbf{y}) = \mathcal{N}\big(\mathbf{x};\quad \mathbf{m_x} + \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}(\mathbf{y} - \mathbf{m_y}),\quad \boldsymbol{\Sigma}_{\mathbf{xx}} - \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\boldsymbol{\Sigma}_{\mathbf{yx}}\big)\,, \tag{19}$$

  and the marginal density

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{m_x}, \boldsymbol{\Sigma}_{\mathbf{xx}})\,. \tag{20}$$

- Woodbury identity: $(A + UCV)^{-1} = A^{-1} - A^{-1}U\big(C^{-1} + VA^{-1}U\big)^{-1}VA^{-1}$.

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

$$p(\mathbf{x}) = \frac{1}{C}(x_1^2 + x_1x_2 + x_2^2 + 2x_2x_3)\,. \tag{21}$$

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 $\mathrm{X} \in \mathbb{R}^{D\times N}$ where $\mathrm{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$:

$$p(\mathbf{y}|\boldsymbol{\theta}, \mathrm{X}) = \mathcal{N}\big(\mathbf{y}; \boldsymbol{\theta}^\intercal X, \sigma^2\mathrm{I}\big) \tag{22}$$

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 and variance σ^2.

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

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

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. $\mathrm{Tr}(\mathrm{ABC}) = \mathrm{Tr}(\mathrm{CAB})$.

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. $f(\mathrm{X}) = \mathrm{Tr}(\mathrm{AXB})$ 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. $f(z) = \log(1+z), \qquad z = \mathbf{x}^\intercal \mathbf{x}, \qquad \mathbf{x} \in \mathbb{R}^D.$

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

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

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

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

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

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:

$$\begin{aligned}
f &= \mathbf{y}^\intercal (\mathrm{K}_1 + \mathrm{K}_2)^{-1} \mathbf{y}\,, && (23)\\
\mathrm{K}_a &= \exp(\Lambda_a)\,, && (24)\\
\Lambda_a &= -\frac{\mathrm{D}_a}{2\ell_a^2}\,, && (25)\\
\mathrm{D}_a &= (\mathrm{X}[:, \texttt{None}, a] - \mathrm{X}[\texttt{None}, :, a])^2\,, && (26)
\end{aligned}$$

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 $\mathbf{\Phi} = (\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), ..., \phi(\mathbf{x}_N))^\top$ and $\mathbf{y} = (y_1, ..., y_N)^\top$, we have the optimal solution satisfies $\boldsymbol{\theta}^* = (\mathbf{\Phi}^\top \mathbf{\Phi})^{-1} \mathbf{\Phi}^\top \mathbf{y}$. 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 Lectures 6 & 7: Multivariate Probability

Question 34 (Vector independence)**.** While you can probably 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

$$p(\mathbf{x}) = \mathbf{x}^\intercal \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix} \mathbf{x}\,. \tag{27}$$

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.

Question 35 (Linear transform of a Gaussian random variable)**.** If X is a d-dimensional multivariate Gaussian random variable with mean $\boldsymbol{\mu}$ and covariance matrix Σ, then what is the distribution of the random variable $Y = \mathbf{A}X$ with an invertible matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$?

Question 36 (Sum of independent Gaussian random variables)**.** If X, Y are two independent univariate Gaussian random variables (i.e., $X \perp\!\!\!\perp Y$), show that $Z = X + Y$ is also a univariate Gaussian random variable.

Question 37 (KL divergence and change-of-variables rule)**.** Show that KL divergence is invariant to change-of-variables, i.e., $\mathrm{KL}[p_X(x)||q_X(x)] = \mathrm{KL}[p_Y(y)||q_Y(y)]$ for $Y = T(X)$ with an invertible transformation T.

Question 38 (Independence of Gaussian variables)**.** Consider $X = (X_1, ..., X_N)$ as a multivariate random variable, which is distributed as a multivariate Gaussian with covariance matrix Σ. Show that $X_i \perp\!\!\!\perp X_j|X_{-ij}$ where X_{-ij} collect all the other X_n variables, if for the precision matrix $\Lambda := \Sigma^{-1}$ we have $\Lambda_{ij} = \Lambda_{ji} = 0$.

Question 39 (Independent vs uncorrelated variables)**.** Show that for the following definitions of X, Y, these two variables are uncorrelated (i.e., $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$), but not independent to each other: X is a univariate Gaussian variable with mean 0, and $Y = X^2$.

Question 40 (Expectation identities)**.** Prove the following expectation identities:

a. $\mathbb{V}_X[X] = \mathbb{E}_X[XX^\intercal] - \mathbb{E}_X[X]\mathbb{E}_X[X]^\intercal$, for $X \in \mathbb{R}^D$.

b. $\mathbb{C}_{X,Y}[X, Y] = \mathbb{E}_{X,Y}[XY^\intercal] - \mathbb{E}_X[X]\mathbb{E}_Y[Y]^\intercal$, for $X \in \mathbb{R}^D, Y \in \mathbb{R}^E$.

c. $\mathbb{E}_{X,Y}[X + Y] = \mathbb{E}_X[X] + \mathbb{E}_Y[Y]$.

d. $\mathbb{V}_{X,Y}[X + Y] = \mathbb{V}_X[X] + \mathbb{V}_Y[Y]$, for $X \perp\!\!\!\perp Y$.

Question 41 (MML 6.11: Iterated Expectations)**.** Consider random variables X, Y with joint distribution $p(x, y)$. Show that

$$\mathbb{E}_X[X] = \mathbb{E}_Y[\mathbb{E}_X[X|Y]] \tag{28}$$

where $\mathbb{E}_X[X|Y]$ denotes the expectation under the conditional distribution $p(x|y)$.

Question 42 (MML 6.13: Probability Integral Transformation)**.** Given a continous r.v. X, with CDF $F_X(x)$, show that the r.v. $Y := F_X(X)$ is uniformly distributed.

10 Lecture 8 & 9: Generalisation, Test Sets, Monte Carlo

Question 43 (Independence of Losses)**.** Under the iid assumption, the loss can be seen as a transformation of a random variable. Show that the losses are independent.

Question 44 (Basic Monte Carlo Estimate)**.** Consider the following integral over the indicator function $\mathbb{I}(\cdot)$, which takes value 1 when its argument evaluates to `True`:

$$I = \int_{-1}^{1}\int_{-1}^{1} \mathbb{I}(x^2 + y^2 < 1)\mathrm{d}x\mathrm{d}y\,. \tag{29}$$

a. Write this integral as an expectation.

b. Construct a Monte Carlo estimate $\hat{I}$ for this integral.

c. Bonus: Implement this in Python and verify that the value converges to π.

d. Is $\hat{I}^2$ an unbiased estimator for π^2?

11 Lecture 10 & 11: Bayesian Inference

Question 45 (Electrical Communication)**.** Consider the electrical communication example from lectures, where we had a Gaussian distribution on the source voltage, i.e. $p(S = s) = \mathcal{N}(s; 0, 1)$. This time, we make multiple observations, i.e. $V_n|s \overset{\text{iid}}{\sim} \mathcal{N}\big(s, \sigma^2\big)$.

a. Write down Bayes' rule to find the posterior for $p(s|v_1, v_2, \ldots v_N)$.

b. By completing the square, find the density of the posterior $p(s|v_1, v_2, \ldots v_N)$.

c. Show that the likelihood function (which is a function of s!) can be rewritten as

$$p(v_1, v_2, \ldots, v_N|s) = c \cdot p(\bar{v}|s) = c \cdot \mathcal{N}\left(\bar{v}; s, \frac{\sigma^2}{N}\right), \tag{30}$$

where $\bar{v} = \frac{1}{N}\sum_{n=1}^{N} v_n$.

d. Find the joint distribution $p(\bar{v}, s)$.

e. Use the Gaussian conditioning rule to find $p(s|\bar{v})$.

f. Reflect on which method you find easier.

Question 46 (Electrical Communication Errors)**.** Consider the electrical communication example from lectures, where $p(v|s) = \mathcal{N}\big(v; s, \sigma^2\big)$ and a Bernoulli S, i.e. $p(S = s) = p^s(1-p)^{1-s}$. Assume that the noise distribution in the model is the same as that of the data generating process. Now consider a *decision rule*, where we guess the transmitted signal using the rule $\hat{S} = \operatorname{argmax}_s p(s|v)$.

Now consider the true frequency of S to follow $\pi(s) = \mathcal{B}(0.6)$. Calculate the probability of making an error $\mathbb{P}\Big(\hat{S} = S\Big)$, as a function of our prior probability p.

Hint: Remember the difference between the data generating distribution ($\mathbb{P}/\pi$), and our model (P/p). When calculating probabilities w.r.t. the data generating distribution, $\hat{S}$ is a function of S.

Question 47 (Vector Ordering in Gaussians)**.** Consider a joint Gaussian density on a vector $\mathbf{z}$ that can be split up as $\mathbf{z} = [\mathbf{x}^\intercal, \mathbf{y}^\intercal]^\intercal$ (this notation denotes stacking):

$$p(\mathbf{z}) = \mathcal{N}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}; \begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix}, \begin{bmatrix}\mathrm{A} & \mathrm{B}\\ \mathrm{B}^\intercal & \mathrm{C}\end{bmatrix}\right). \tag{31}$$

Now consider the permuted vector $\mathbf{z}' = [\mathbf{y}^\intercal, \mathbf{x}^\intercal]^\intercal$. Show that the Gaussian distribution of $p(\mathbf{z}')$ has a mean with rows, and a covariance matrix with swapped rows and columns.

Hint: Notice that $\mathbf{z}'$ is a permuted version of $\mathbf{z}$. We can write this mathematically as $\mathbf{z}' = \mathrm{P}\mathbf{z}$, where P is a permutation matrix:

$$\mathrm{P} = \begin{bmatrix}0 & \mathrm{I}\\ \mathrm{I} & 0\end{bmatrix}. \tag{32}$$

Note: This is an important skill! Notice that the Gaussian conditioning formula in the formula sheet is only written in one way.

Question 48 (Woodbury Identity)**.** We saw that the two different ways of deriving the Gaussian posterior, gave two different results:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}) &= \mathcal{N}\Big(\boldsymbol{\theta}; \Phi(X)^\intercal\big[\Phi(X)\Phi(X)^\intercal + \sigma^2\mathrm{I}_N\big]^{-1}\mathbf{y}, \qquad \mathrm{I}_M - \Phi(X)^\intercal\big[\Phi(X)\Phi(X)^\intercal + \sigma^2\mathrm{I}_N\big]^{-1}\Phi(X)\Big), \\ p(\boldsymbol{\theta}|\mathbf{y}) &= \mathcal{N}\left(\boldsymbol{\theta}; \left[\frac{1}{\sigma^2}\Phi(X)^\intercal\Phi(X) + \mathrm{I}_M\right]^{-1}\frac{1}{\sigma^2}\Phi(X)^\intercal\mathbf{y}, \ \left[\frac{1}{\sigma^2}\Phi(X)^\intercal\Phi(X) + \mathrm{I}_M\right]^{-1}\right). \end{aligned} \tag{33}$$

Apply the Woodbury identity to $\big[\frac{1}{\sigma^2}\Phi(X)^\intercal\Phi(X) + \mathrm{I}_M\big]^{-1}$ to show that the two solutions are equal.

Note: This is an excellent exercise for your matrix algebra skills. The main difficulty that most uninitiated have, is that multiplication no longer commutes. You need to be careful that you consistently pre/post multiply matrices.

Question 49 (BLR Predictive)**.** For a Bayesian Linear Regression model:

1. Find $p(\mathbf{y}^*, \mathbf{y})$.
2. Find $p(\mathbf{y}^*|\mathbf{y})$.

12 Lecture 12: Bias-variance tradeoff

Question 50 (Variance reduction for Ridge regression)**.** Consider the covariance matrix of the ridge regression estimator $\boldsymbol{\theta}^*_R(\mathcal{D})$ which depends on λ:

$$\mathbf{V}(\lambda) = \mathbb{V}_{\mathcal{D}\sim\pi^N}[\boldsymbol{\theta}^*_R(\mathcal{D})], \quad \boldsymbol{\theta}^*_R(\mathcal{D}) := \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma^2}||\mathbf{y} - \Phi\boldsymbol{\theta}||_2^2 + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2.$$

Show that $\mathbf{V}(\lambda) \preceq \mathbf{V}(0)$ for all $\lambda > 0$. (We assume $\Phi^\top\Phi$ is invertible.)

Question 51 (Bias-Variance tradeoff in Ridge regression)**.** Continuing Question 50, let us write the bias of the ridge regression estimator $\boldsymbol{\theta}^*_R(\mathcal{D})$ as (under no model error assumption and assume the ground-truth parameter is $\boldsymbol{\theta}_0$)

$$\mathbf{b}(\boldsymbol{\theta}^*_R) = \mathbb{E}_{\mathcal{D}\sim p^N_{data}}[\boldsymbol{\theta}^*_R(\mathcal{D})] - \boldsymbol{\theta}_0.$$

Show that when $0 \leq \lambda \leq \frac{2}{||\boldsymbol{\theta}_0||_2^2}$ we have

$$\mathbf{b}(\lambda)\mathbf{b}(\lambda)^\top + \mathbf{V}(\lambda) \preceq \mathbf{V}(0).$$

This result is immediately useful to show that the expected test error of Ridge regression can be smaller than the usual linear regression (i.e., MLE estimate).

Question 52 (Control variate)**.** Consider X as an unbiased estimator of a scalar quantity x_0. Show that for the estimator $X + Y - \mathbb{E}_Y[Y]$ with another random variable Y:

a. It is also an unbiased estimator of x_0;

b. The variance of the estimator is reduced when $\mathbb{V}_Y[Y] + 2Cov_{X,Y}[X,Y] < 0$.

c. Assume $Y = cZ$ where Z is a random variable that is correlated with X, and c is a scaling constant. Choose the best $c \in \mathbb{R}$ such that we achieve the maximum variance reduction.

13 Lecture 13: PCA

Question 53 (Connections to linear auto-encoders)**.** Consider a linear auto-encoder defined as follows for an input $\mathbf{x} \in \mathbb{R}^{D\times 1}$:

$$\text{Encoder:}\quad \mathbf{z} = enc(\mathbf{x}) = \mathbf{B}\mathbf{x},\ \mathbf{B} \in \mathbb{R}^{M\times D}, M < D, \quad \text{Decoder:}\quad \hat{\mathbf{x}} = dec(\mathbf{z}) = \mathbf{A}\mathbf{z},\ \mathbf{A} \in \mathbb{R}^{D\times M}.$$

Let us assume $rank(\mathbf{A}) = rank(\mathbf{B}) = M$. Given a dataset $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^N$ such that mean$(\mathbf{x}_n) = \mathbf{0}$ and the covariance matrix computed on $\mathcal{D}$ is invertible, we wish to train the model by minimising the ℓ_2 reconstruction error:

$$\min_{\mathbf{A},\mathbf{B}} L(\mathbf{A},\mathbf{B}), \quad L(\mathbf{A},\mathbf{B}) := \frac{1}{N}\sum_{n=1}^N ||\mathbf{x}_n - \mathbf{A}\mathbf{B}\mathbf{x}_n||_2^2.$$

a. What is the derivative of the objective w.r.t. $\mathbf{A}$ and $\mathbf{B}$?

b. Given a fixed $\mathbf{A}$, what is the minimiser solution of $\mathbf{B}$?

c. Assume the covariance matrix computed on $\mathcal{D}$ can be eigendecomposed as $\mathbf{Q}\Lambda\mathbf{Q}^\top$ with $\Lambda = diag(\lambda_1, ..., \lambda_D), \lambda_1 \geq ... \geq \lambda_D > 0$. Show that $\mathbf{A}^* = \mathbf{Q}_{1:M}, \mathbf{B}^* = \mathbf{Q}_{1:M}^\top$ is a fixed point of the objective $L(\mathbf{A},\mathbf{B})$, where $\mathbf{Q}_{1:M}$ contains the first M columns of $\mathbf{Q}$.

You might find the following formula useful:

$$\frac{\partial}{\partial\mathbf{X}}\text{tr}(\mathbf{A}\mathbf{X}\mathbf{B}) = \mathbf{A}^\top\mathbf{B}^\top, \quad \frac{\partial}{\partial\mathbf{X}}\text{tr}(\mathbf{X}^\top\mathbf{B}\mathbf{X}\mathbf{C}) = \mathbf{B}\mathbf{X}\mathbf{C} + \mathbf{B}^\top\mathbf{X}\mathbf{C}^\top, \quad \frac{\partial}{\partial\mathbf{X}}\text{tr}(\mathbf{X}^\top\mathbf{X}\mathbf{B}) = \mathbf{X}\mathbf{B} + \mathbf{X}\mathbf{B}^\top.$$

Question 54 (SVD and PCA)**.** Assume we are given a dataset $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^N$ such that mean$(\mathbf{x}_n) = \mathbf{0}$. Write $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]^\top \in \mathbb{R}^{N\times D}$, demonstrate how to use singular value decomposition (SVD) of $\mathbf{X}$ to obtain PCA solutions of M principal components.

14 Lecture 14: Probabilistic PCA

Question 55 (Solving for the optimal $\mathbf{W}$ in Probabilistic PCA)**.** Consider fitting the following Probabilistic PCA model to a dataset $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^N, \mathbf{x}_n \in \mathbb{R}^{D\times 1}$:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}), \quad \mathbf{z} \in \mathbb{R}^{M\times 1}, M < D,$$

$$p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$$

In the lecture we have derived a few fixed point solutions as (with data covariance matrix $\mathbf{S} = \mathbf{Q}\Lambda\mathbf{Q}^\top$ and the eigenvalues in Λ arranged in a descending order):

$$\boldsymbol{\mu}^* = \frac{1}{N}\sum_{n=1}^N \mathbf{x}_n, \quad \mathbf{W}^* = \mathbf{U}\Sigma\mathbf{V}^\top, \quad \mathbf{U} := (\boldsymbol{u}_1, ..., \boldsymbol{u}_D), \boldsymbol{u}_m = \boldsymbol{q}_{i_m}, 1 \le i_m \le D, m = 1, ..., M.$$

a. Show that the corresponding fixed point satisfies $\Sigma_{mm} = \sqrt{\lambda_{i_m} - \sigma^2}$ for $m = 1, ..., M$, and we assume $\lambda_{i_m} \ge \sigma^2$ for appropriately chosen σ.

b. The global maximum solutions satisfy $i_m = m$ for $m = 1, ..., M$ (so that $\boldsymbol{u}_m = \boldsymbol{q}_m$).

15 Warm-up Exercises Answers

15.1 Probability Theory

Question 1 – Set Theory and Probability

a. We can choose any representation denoting the events, e.g. using abstract symbols Ω = {⚀, ⚁, ⚂, ⚃, ⚄, ⚅}. 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. $0 \le P(A) \le 1$

   We use property (c) and axiom 1.

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

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

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

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

   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 $P(B\backslash A) \ge 0$.

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

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:

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

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.

$$A_{i-1} \subseteq A_i \quad \forall i > 0 \implies A = \bigcup_{i=1}^{\infty} A_i \backslash A_{i-1} \tag{34}$$

where the expression holds if we have $A_0 = \varnothing$. We regard 34 as starting with A_1 and adding the new information from $A_2, A_3, \ldots$ (e.g $A_2 \backslash A_1, A_3 \backslash A_2, \ldots$).

$$P(A) = P\left(\bigcup_{i=1}^{\infty} A_i \backslash A_{i-1}\right) = \sum_{i=1}^{\infty} P(A_i \backslash A_{i-1}) \qquad \text{(by axiom 3)} \tag{35}$$

$$P(A) = \sum_{i=1}^{\infty} P(A_i \backslash A_{i-1}) = \lim_{n \longrightarrow \infty} \sum_{i=1}^{n} P(A_i \backslash A_{i-1}) \qquad \text{(the infinite summation is a limit)} \tag{36}$$

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,

$$P(A) = \lim_{n \longrightarrow \infty} \sum_{i=1}^{n} P(A_i) - P(A_{i-1}) = \lim_{n \longrightarrow \infty} \left(\sum_{i=1}^{n} P(A_i) - \sum_{i=1}^{n-1} P(A_i)\right) = \lim_{n \longrightarrow \infty} P(A_n) \tag{37}$$

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

In summary:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P(A) = \lim_{i \longrightarrow \infty} P(A_i) \tag{38}$$

Question 2 – Independent events

a. $E_{1H} = \{HH, HT\}$

b. Show E_{1H} and E_{2T} are independent.

Independence holds if $P(E_{1H} \cap E_{2T}) = P(E_{1H})P(E_{2T})$. If all outcomes have equal probability, i.e. $P(HH) = P(HT) = P(TH) = P(TT) = 0.25$, then we have $P(E_{1H}) = P(HH \cup HT) = P(HH) + P(HT) = 0.5$, and $P(E_{2T}) = P(HT \cup TT) = P(HT) + P(TT) = 0.5$.

The intersection of E_{1H} and E_{2T} is $\{HT\}$, hence $P(E_{1H} \cap E_{2T}) = P(HT) = 0.25$. Since $P(E_{1H})P(E_{2T}) = 0.5^2 = 0.25$, E_{1H} and E_{2T} are independent.

c. Assume E_{1H} and E_{2H} are independent and $P(E_{1H}) = P(E_{2H}) = 0.5$, show all outcomes must have equal probability.

We can take these assumptions to form a system of equations, where the probability of each outcome is treated as a variable.

$$\begin{cases} P(E_{1H}) = P(HH) + P(HT) = 0.5 \\ P(E_{2H}) = P(TH) + P(HH) = 0.5 \\ P(E_{1H} \cap E_{2H}) = P(HH) = P(E_{1H})P(E_{2H}) = 0.25, \quad \text{by assuming independence} \\ P(TT) = 1 - P(HH) - P(HT) - P(TH), \quad \text{from probability axioms} \end{cases}$$

We have a system of 4 equations with 4 variables, which therefore should yield a unique solution (or none). There is in fact a unique solution, which is $P(HH) = P(HT) = P(TH) = P(TT) = 0.25$. Therefore, given the previous assumptions all outcomes must have equal probability.

Question 3 – Random Variables

a. We choose to represent the outcomes of two dice as integer tuples:

$$
\begin{aligned}
\Omega = \{&(\epsdice{1},\epsdice{1}),(\epsdice{1},\epsdice{2}),(\epsdice{1},\epsdice{3}),(\epsdice{1},\epsdice{4}),(\epsdice{1},\epsdice{5}),(\epsdice{1},\epsdice{6}),\\
&(\epsdice{2},\epsdice{1}),(\epsdice{2},\epsdice{2}),(\epsdice{2},\epsdice{3}),(\epsdice{2},\epsdice{4}),(\epsdice{2},\epsdice{5}),(\epsdice{2},\epsdice{6}),\\
&(\epsdice{3},\epsdice{1}),(\epsdice{3},\epsdice{2}),(\epsdice{3},\epsdice{3}),(\epsdice{3},\epsdice{4}),(\epsdice{3},\epsdice{5}),(\epsdice{3},\epsdice{6}),\\
&(\epsdice{4},\epsdice{1}),(\epsdice{4},\epsdice{2}),(\epsdice{4},\epsdice{3}),(\epsdice{4},\epsdice{4}),(\epsdice{4},\epsdice{5}),(\epsdice{4},\epsdice{6}),\\
&(\epsdice{5},\epsdice{1}),(\epsdice{5},\epsdice{2}),(\epsdice{5},\epsdice{3}),(\epsdice{5},\epsdice{4}),(\epsdice{5},\epsdice{5}),(\epsdice{5},\epsdice{6}),\\
&(\epsdice{6},\epsdice{1}),(\epsdice{6},\epsdice{2}),(\epsdice{6},\epsdice{3}),(\epsdice{6},\epsdice{4}),(\epsdice{6},\epsdice{5}),(\epsdice{6},\epsdice{6})\}
\end{aligned}
$$

b. We define random variables A and B to be:

$$
A(s) = \begin{cases}
1 & \text{, if } s \in \{(\epsdice{1},\epsdice{1}),(\epsdice{1},\epsdice{2}),(\epsdice{1},\epsdice{3}),(\epsdice{1},\epsdice{4}),(\epsdice{1},\epsdice{5}),(\epsdice{1},\epsdice{6})\}\\
2 & \text{, if } s \in \{(\epsdice{2},\epsdice{1}),(\epsdice{2},\epsdice{2}),(\epsdice{2},\epsdice{3}),(\epsdice{2},\epsdice{4}),(\epsdice{2},\epsdice{5}),(\epsdice{2},\epsdice{6})\}\\
3 & \text{, if } s \in \{(\epsdice{3},\epsdice{1}),(\epsdice{3},\epsdice{2}),(\epsdice{3},\epsdice{3}),(\epsdice{3},\epsdice{4}),(\epsdice{3},\epsdice{5}),(\epsdice{3},\epsdice{6})\}\\
4 & \text{, if } s \in \{(\epsdice{4},\epsdice{1}),(\epsdice{4},\epsdice{2}),(\epsdice{4},\epsdice{3}),(\epsdice{4},\epsdice{4}),(\epsdice{4},\epsdice{5}),(\epsdice{4},\epsdice{6})\}\\
5 & \text{, if } s \in \{(\epsdice{5},\epsdice{1}),(\epsdice{5},\epsdice{2}),(\epsdice{5},\epsdice{3}),(\epsdice{5},\epsdice{4}),(\epsdice{5},\epsdice{5}),(\epsdice{5},\epsdice{6})\}\\
6 & \text{, if } s \in \{(\epsdice{6},\epsdice{1}),(\epsdice{6},\epsdice{2}),(\epsdice{6},\epsdice{3}),(\epsdice{6},\epsdice{4}),(\epsdice{6},\epsdice{5}),(\epsdice{6},\epsdice{6})\}
\end{cases}
$$

$$
B(s) = \begin{cases}
1 & \text{, if } s \in \{(\epsdice{1},\epsdice{1}),(\epsdice{2},\epsdice{1}),(\epsdice{3},\epsdice{1}),(\epsdice{4},\epsdice{1}),(\epsdice{5},\epsdice{1}),(\epsdice{6},\epsdice{1})\}\\
2 & \text{, if } s \in \{(\epsdice{1},\epsdice{2}),(\epsdice{2},\epsdice{2}),(\epsdice{3},\epsdice{2}),(\epsdice{4},\epsdice{2}),(\epsdice{5},\epsdice{2}),(\epsdice{6},\epsdice{2})\}\\
3 & \text{, if } s \in \{(\epsdice{1},\epsdice{3}),(\epsdice{2},\epsdice{3}),(\epsdice{3},\epsdice{3}),(\epsdice{4},\epsdice{3}),(\epsdice{5},\epsdice{3}),(\epsdice{6},\epsdice{3})\}\\
4 & \text{, if } s \in \{(\epsdice{1},\epsdice{4}),(\epsdice{2},\epsdice{4}),(\epsdice{3},\epsdice{4}),(\epsdice{4},\epsdice{4}),(\epsdice{5},\epsdice{4}),(\epsdice{6},\epsdice{4})\}\\
5 & \text{, if } s \in \{(\epsdice{1},\epsdice{5}),(\epsdice{2},\epsdice{5}),(\epsdice{3},\epsdice{5}),(\epsdice{4},\epsdice{5}),(\epsdice{5},\epsdice{5}),(\epsdice{6},\epsdice{5})\}\\
6 & \text{, if } s \in \{(\epsdice{1},\epsdice{6}),(\epsdice{2},\epsdice{6}),(\epsdice{3},\epsdice{6}),(\epsdice{4},\epsdice{6}),(\epsdice{5},\epsdice{6}),(\epsdice{6},\epsdice{6})\}
\end{cases}
$$

We can find the PMFs by counting the number of occurrences in Ω. For instance:

$$
p_A(3) = \frac{|\{(\epsdice{3},\epsdice{1}),(\epsdice{3},\epsdice{2}),(\epsdice{3},\epsdice{3}),(\epsdice{3},\epsdice{4}),(\epsdice{3},\epsdice{5}),(\epsdice{3},\epsdice{6})\}|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}
$$

Repeating this for all outcomes gives us the full PDFs:

$$
p_A(x) = \begin{cases}
\frac{1}{6} & \text{, if } x = 1\\
\frac{1}{6} & \text{, if } x = 2\\
\frac{1}{6} & \text{, if } x = 3\\
\frac{1}{6} & \text{, if } x = 4\\
\frac{1}{6} & \text{, if } x = 5\\
\frac{1}{6} & \text{, if } x = 6\\
0 & \text{, otherwise}
\end{cases}, \qquad
p_B(x) = \begin{cases}
\frac{1}{6} & \text{, if } x = 1\\
\frac{1}{6} & \text{, if } x = 2\\
\frac{1}{6} & \text{, if } x = 3\\
\frac{1}{6} & \text{, if } x = 4\\
\frac{1}{6} & \text{, if } x = 5\\
\frac{1}{6} & \text{, if } x = 6\\
0 & \text{, otherwise}
\end{cases}
$$

c. To show independence of A and B we must show that $p(A \cap B) = p(A)p(B)$. We have that all outcomes have equal probability $\frac{1}{|\Omega|} = \frac{1}{36}$ and therefore:

$$
p(A \cap B) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = p(A)p(B)
$$

d. We can define a random variable $C = A + B$

$$C(s) = \begin{cases} 2 & \text{, if } s \in \{(\epsdice{1},\epsdice{1})\} \\ 3 & \text{, if } s \in \{(\epsdice{1},\epsdice{2}),(\epsdice{2},\epsdice{1})\} \\ 4 & \text{, if } s \in \{(\epsdice{1},\epsdice{3}),(\epsdice{2},\epsdice{2}),(\epsdice{3},\epsdice{1})\} \\ 5 & \text{, if } s \in \{(\epsdice{1},\epsdice{4}),(\epsdice{2},\epsdice{3}),(\epsdice{3},\epsdice{2}),(\epsdice{4},\epsdice{1})\} \\ 6 & \text{, if } s \in \{(\epsdice{1},\epsdice{5}),(\epsdice{2},\epsdice{4}),(\epsdice{3},\epsdice{3}),(\epsdice{4},\epsdice{2}),(\epsdice{5},\epsdice{1})\} \\ 7 & \text{, if } s \in \{(\epsdice{1},\epsdice{6}),(\epsdice{2},\epsdice{5}),(\epsdice{3},\epsdice{4}),(\epsdice{4},\epsdice{3}),(\epsdice{5},\epsdice{2}),(\epsdice{6},\epsdice{1})\} \\ 8 & \text{, if } s \in \{(\epsdice{2},\epsdice{6}),(\epsdice{3},\epsdice{5}),(\epsdice{4},\epsdice{4}),(\epsdice{5},\epsdice{3}),(\epsdice{6},\epsdice{2})\} \\ 9 & \text{, if } s \in \{(\epsdice{3},\epsdice{6}),(\epsdice{4},\epsdice{5}),(\epsdice{5},\epsdice{4}),(\epsdice{6},\epsdice{3})\} \\ 10 & \text{, if } s \in \{(\epsdice{4},\epsdice{6}),(\epsdice{5},\epsdice{5}),(\epsdice{6},\epsdice{4})\} \\ 11 & \text{, if } s \in \{(\epsdice{5},\epsdice{6}),(\epsdice{6},\epsdice{5})\} \\ 12 & \text{, if } s \in \{(\epsdice{6},\epsdice{6})\} \end{cases}$$

Then the PDF p_C becomes:

$$p_C(x) = \begin{cases} \frac{1}{36} = \frac{1}{36} & \text{, if } x = 2 \\ \frac{2}{36} = \frac{1}{18} & \text{, if } x = 3 \\ \frac{3}{36} = \frac{1}{12} & \text{, if } x = 4 \\ \frac{4}{36} = \frac{1}{9} & \text{, if } x = 5 \\ \frac{5}{36} = \frac{5}{36} & \text{, if } x = 6 \\ \frac{6}{36} = \frac{1}{6} & \text{, if } x = 7 \\ \frac{5}{36} = \frac{5}{36} & \text{, if } x = 8 \\ \frac{3}{36} = \frac{1}{9} & \text{, if } x = 9 \\ \frac{3}{36} = \frac{1}{12} & \text{, if } x = 10 \\ \frac{2}{36} = \frac{1}{18} & \text{, if } x = 11 \\ \frac{1}{36} = \frac{1}{36} & \text{, if } x = 12 \\ 0 & \text{, otherwise} \end{cases}$$

which can be rewritten in more compact form:

$$p_C(x) = \begin{cases} \frac{6-|x-6|}{36} & \text{, if } x = \{2, 3, \ldots, 12\} \\ 0 & \text{, otherwise} \end{cases}$$

Question 4 – Continuous Random Variables

a.

$$1 = \int_{-\infty}^{\infty} p(x)\mathrm{d}x = \int_0^1 Cx\mathrm{d}x = \frac{1}{2}Cx^2\Big|_0^1 = \frac{1}{2}C = 1$$
$$\implies C = 2$$

b.

$$\mathbb{P}(0.3 \leq X \leq 0.75) = \int_{0.3}^{0.75} 2x\mathrm{d}x = x^2\Big|_{0.3}^{0.75} = 0.75^2 - 0.3^2 = 0.4725$$

c.

$$\begin{aligned}\mathbb{P}(X \in [0.3, 0.75] \cup [0.8, 0.9]) &= \int_{0.3}^{0.75} 2x\mathrm{d}x + \int_{0.8}^{0.9} 2x\mathrm{d}x \\ &= x^2\Big|_{0.3}^{0.75} + x^2\Big|_{0.8}^{0.9} = 0.75^2 - 0.3^2 + 0.9^2 - 0.8^2 = 0.6425\end{aligned}$$

d.

$$\mathbb{E}_X[X] = \int xp(x)\mathrm{d}x = \int_0^1 2x^2\mathrm{d}x = \frac{2}{3}x^3\Big|_0^1 = \frac{2}{3}$$

$$\mathbb{E}_X[X^2] = \int x^2 p(x)\mathrm{d}x = \int_0^1 2x^3 \mathrm{d}x = \frac{2}{4}x^4\Big|_0^1 = \frac{1}{2}$$

We can derive the following useful identity that generally holds:

$$\begin{aligned}
\mathbb{V}_X[X] &= \mathbb{E}_X[(X - E_X[X])^2] \\
&= \mathbb{E}_X[X^2 - 2X\mathbb{E}_X[X] + \mathbb{E}_X[X]^2 \\
&= \mathbb{E}_X[X^2] - 2\mathbb{E}_X\mathbb{E}_X[X] + \mathbb{E}_X[X]^2 \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2
\end{aligned}$$

Using this fact $\mathbb{V}_X[X] = \mathbb{E}_X[X^2] - (\mathbb{E}_X[X])^2$ we can calculate the variance:

$$\mathbb{V}_X[X] = \mathbb{E}_X[X^2] - (\mathbb{E}_X[X])^2 = \frac{1}{2} - (\frac{2}{3})^2 = \frac{1}{18}$$

Question 5 – Joint Discrete Random Variables

a.

$P(C = c \mid A = a)$	$a = 1$	$a = 2$	$a = 3$	$a = 4$	$a = 5$	$a = 6$
$c = 2$	$\frac{1}{6}$	0	0	0	0	0
$c = 3$	$\frac{1}{6}$	$\frac{1}{6}$	0	0	0	0
$c = 4$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0	0	0
$c = 5$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0	0
$c = 6$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0
$c = 7$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
$c = 8$	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
$c = 8$	0	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
$c = 8$	0	0	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
$c = 8$	0	0	0	0	$\frac{1}{6}$	$\frac{1}{6}$
$c = 8$	0	0	0	0	0	$\frac{1}{6}$

b.

$P(C = c, A = a)$	$a = 1$	$a = 2$	$a = 3$	$a = 4$	$a = 5$	$a = 6$
$c = 2$	$\frac{1}{36}$	0	0	0	0	0
$c = 3$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0
$c = 4$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0
$c = 5$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0
$c = 6$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0
$c = 7$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
$c = 8$	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
$c = 9$	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
$c = 10$	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
$c = 11$	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$
$c = 12$	0	0	0	0	0	$\frac{1}{36}$

c.

$$\begin{aligned}
\mathbb{P}(2 \leq C \leq 4) &= \sum_{c\in\{2,3,4\}} \sum_{a} \mathbb{P}(C = c, A = a) \\
&= 6 \cdot \frac{1}{36} + 12 \cdot 0 = \frac{1}{6} \\
\mathbb{P}(2 \leq C \leq 4, 2 \leq A \leq 4) &= \sum_{c\in\{2,3,4\}} \sum_{a\in\{2,3,4\}} \mathbb{P}(C = c, A = a) \\
&= 3 \cdot \frac{1}{36} + 6 \cdot 0 = \frac{1}{12}
\end{aligned}$$

Question 6 – Multivariate Integration

a. The probability of the sample space should equal to 1, by the axioms of probability theory. Therefore, if we integrate over all possible outcomes of random variable X, from $-\infty$ to $+\infty$, we should obtain 1. We have that $p(x, y)$ is only defined on a small region, namely when $x \in [0, 1]$ and $y \in [0, 1]$. Outside this region the integral evaluates to zero. By the sum rule of integration, this means we can evaluate the integral by only integrating over the region where $p(x, y)$ is non-zero:

$$
\begin{aligned}
1 &= \int_{-\infty}^{\infty} p(x, y)\mathrm{d}x\mathrm{d}y = \int_0^1 \int_0^1 C(x^2 + xy)\mathrm{d}x\mathrm{d}y \\
&= C \int_0^1 \left(\left(\frac{1}{3}x^3 \Big|_{x=0}^{1} \right) + \left(\frac{1}{2}yx^2 \Big|_{x=0}^{1} \right) \right) \mathrm{d}y \\
&= C \int_0^1 \left(\frac{1}{3} + \frac{1}{2}y \right) \mathrm{d}y \\
&= C \left(\frac{1}{3}y \Big|_0^1 \right) + C \left(\frac{1}{4}y^2 \Big|_0^1 \right) \\
&= C \left(\frac{1}{3} + \frac{1}{4} \right) = C \frac{7}{12} = 1 \\
&\implies C = \frac{12}{7}
\end{aligned}
$$

b. To find $\mathbb{P}(0.3 \leq X \leq 0.5)$, we again integrate over all possible outcomes of random variable Y, that is from $-\infty$ to $+\infty$ and only consider outcomes of X by considering range $x \in [0.3, 0.5]$:

$$
\mathbb{P}(0.3 \leq X \leq 0.5) = \int_{-\infty}^{\infty} \int_{0.3}^{0.5} p(x, y)\mathrm{d}x\mathrm{d}y
$$

Since $p(x, y)$ is only defined on a small region, namely when $x \in [0, 1]$ and $y \in [0, 1]$. Outside this region, we have that $p(x, y)$=0 evaluates to zero. So, by the sum rule of integration, what remains is the integrate over the region where $p(x, y)$ takes the non-zero value $(Cx^2 + Cxy)$:

$$
\begin{aligned}
\mathbb{P}(0.3 \leq X \leq 0.5) &= \int_{-\infty}^{\infty} \int_{0.3}^{0.5} p(x, y)\mathrm{d}x\mathrm{d}y \\
&= \int_0^1 \int_{0.3}^{0.5} p(x, y)\mathrm{d}x\mathrm{d}y \\
&= \int_0^1 \int_{0.3}^{0.5} (Cx^2 + Cxy)\mathrm{d}x\mathrm{d}y \\
&= C \int_0^1 \left(\left(\frac{1}{3}x^3 \Big|_{0.3}^{0.5} \right) + \left(\frac{1}{2}x^2 y \Big|_{x=0.3}^{0.5} \right) \right) \mathrm{d}y \\
&= C \int_0^1 \left(\frac{1}{3}(0.5)^3 + \frac{1}{2}(0.5)^2 y - \frac{1}{3}(0.3)^3 - \frac{1}{2}(0.3)^2 y \right) \mathrm{d}y \\
&= C \int_0^1 \left(\frac{49}{1500} + \frac{2}{25}y \right) \mathrm{d}y \\
&= C \left(\frac{49}{1500}y \Big|_0^1 \right) + C \left(\frac{2}{50}y^2 \Big|_0^1 \right) \\
&= \frac{12}{27} \left(\frac{49}{1500} + \frac{2}{50} \right) = \frac{109}{875}
\end{aligned}
$$

c.

$$
\begin{aligned}
\mathbb{P}(X < Y) &= \int_0^1 \int_0^y (Cx^2 + Cxy)\mathrm{d}x\mathrm{d}y \\
&= C \int_0^1 \left(\left(\frac{1}{3}x^3 \Big|_0^y \right) + \left(\frac{1}{2}yx^2 \Big|_{x=0}^{y} \right) \right) \mathrm{d}y \\
&= C \int_0^1 \left(\frac{1}{3}(y)^3 + \frac{1}{2}y^3 \right) \mathrm{d}y
\end{aligned}
$$

$$
\begin{aligned}
&= C\left(\frac{1}{3}\cdot\frac{1}{4}y^4\Big|_0^1\right) + C\left(\frac{1}{2}\cdot\frac{1}{4}y^4\Big|_0^1\right)\\
&= \frac{12}{7}\frac{1}{12} + \frac{12}{7}\frac{1}{8} = \frac{5}{14}
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{P}(X<Y) &= \int_0^1\int_x^1 (Cx^2 + Cxy)\mathrm{d}y\mathrm{d}x\\
&= C\int_0^1 \left(\left(yx^2\Big|_{y=x}^1\right) + \left(\frac{1}{2}xy^2\Big|_{y=x}^1\right)\right)\mathrm{d}y\\
&= C\int_0^1 \left(x^2 + \frac{1}{2}x - x^3 - \frac{1}{2}x^3\right)\mathrm{d}x\\
&= C\left((\frac{1}{3}x^3 + \frac{1}{4}x^2 - \frac{1}{4}x^4 - \frac{1}{8}x^4)\Big|_0^1\right)\\
&= C\left(\frac{1}{3} + \frac{1}{4} - \frac{1}{4} - \frac{1}{8}\right) = \frac{12}{7}\cdot\frac{5}{24} = \frac{5}{14}
\end{aligned}
$$

d.

$$
\begin{aligned}
\mathbb{P}(X<Y) &= \int_0^1\int_0^1\int_0^1 C(x^2 + xyz)\mathrm{d}x\mathrm{d}y\mathrm{d}z\\
&= C\int_0^1\int_0^1\left((\frac{1}{3}x^3 + \frac{1}{2}x^2yz)\Big|_0^1\right)\mathrm{d}y\mathrm{d}z = C\int_0^1\int_0^1\left(\frac{1}{3} + \frac{1}{2}yz\right)\mathrm{d}y\mathrm{d}z\\
&= C\int_0^1\left((\frac{1}{3}y + \frac{1}{4}y^2z\Big|_0^1\right)\mathrm{d}z = C\int_0^1\left(\frac{1}{3} + \frac{1}{4}z\right)\mathrm{d}z\\
&= C\left((\frac{1}{3}z + \frac{1}{8}z^2)\Big|_0^1\right) = C\left(\frac{1}{3} + \frac{1}{8}\right) = \frac{11}{24}C = 1\\
&\implies C = \frac{24}{11}
\end{aligned}
$$

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.

15.2 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 $\boldsymbol{q} \neq 0$:

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

16 Answers Lecture 1: Probability, Vectors, Differentiation

Question 15 – Vector notation Given $p(\mathbf{x}) = \frac{1}{C}(x_1^2 + x_1 x_2 x_2^2 + 2x_2 x_3)$, we need to rearrange the terms to find an expression as follows:

$$p(\mathbf{x}) = \frac{1}{C}(\mathbf{x}^T A\mathbf{x}), \quad A \in \mathbb{R}^{3\times 3}. \tag{39}$$

Inspection of the terms in $p(\mathbf{x})$ gives the following solution

$$\begin{pmatrix}x_1 & x_2 & x_3\end{pmatrix}^T \begin{pmatrix}1 & \frac{1}{2} & 0\\ \frac{1}{2} & 1 & 1\\ 0 & 1 & 0\end{pmatrix} \begin{pmatrix}x_1 & x_2 & x_3\end{pmatrix} = \begin{pmatrix}x_1 + \frac{x_2}{2}\\ \frac{x_1}{2} + x_2 + x_3\\ x_2\end{pmatrix} \begin{pmatrix}x_1 & x_2 & x_3\end{pmatrix} = x_1^2 + x_1 x_2 x_2^2 + 2x_2 x_3.$$

Thus,

$$p(\mathbf{x}) = \frac{1}{C}(\mathbf{x}^T A \mathbf{x}), \quad A = \begin{pmatrix}1 & \frac{1}{2} & 0\\ \frac{1}{2} & 1 & 1\\ 0 & 1 & 0\end{pmatrix}. \tag{40}$$

Question 16 – Noise conditional independence Given inputs $X \in \mathbb{R}^{D\times N}$, where $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, we want to show that all y_ns are independent, i.e,

$$p(\mathbf{y}|\boldsymbol{\theta}, \mathrm{X}) = \prod_{n=1}^{N} p(y_n|\boldsymbol{\theta}, \mathbf{x}_n). \tag{41}$$

We can show this by rearranging terms of the Gaussian distribution:

$$p(\mathbf{y}|\boldsymbol{\theta}, \mathrm{X}) = \mathcal{N}\big(\mathbf{y}; \boldsymbol{\theta}^\intercal X, \sigma^2 \mathrm{I}\big) = \frac{1}{\sqrt{(2\pi)^N |\sigma^2 I|}} \exp\big(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\theta}^\intercal X)^\intercal \sigma^{-2} I (\mathbf{y} - \boldsymbol{\theta}^\intercal X)\big) \tag{42}$$

$$= \frac{1}{\sqrt{(2\pi\sigma^2)^N}} \exp\big(-\frac{1}{2\sigma^2}||\mathbf{y} - \boldsymbol{\theta}^\intercal X||^2\big) = \frac{1}{\sqrt{(2\pi\sigma^2)^N}} \exp\big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \boldsymbol{\theta}^\intercal \mathbf{x}_n)^2\big) \tag{43}$$

$$= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big(-\frac{1}{2\sigma^2}(y_n - \boldsymbol{\theta}^\intercal \mathbf{x}_n)^2\big) = \prod_{n=1}^{N} p(y_n|\boldsymbol{\theta}, \mathbf{x}_n). \tag{44}$$

Question 17 a) Our model assumes Gaussian density function defined over N iid data samples collected in $x \in \mathbb{R}^N$ as $x = [x_1, \ldots, x_N]^\intercal$. Given the iid assumption, we can find the joint distribution by multiplying the same distribution N times:

$$p(x \mid \mu, \sigma) = \prod_{i=1}^{N} \mathcal{N}(x_i; \mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{(x_i - \mu)^2}{2\sigma^2}\Big)$$

b) The log-likelihood of the above distribution can be defined as:

$$\mathcal{L}(\mu, \sigma) = \log p(x|\mu, \sigma^2) = \log \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x_i - \mu)^2}{2\sigma^2}\Big)$$
$$= -N\log(\sqrt{2\pi\sigma^2}) - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}$$

Maximum likilehood estimates for μ, σ are obtained by maximizing the above likelihood function, which corresponds to:

$$\mu^*, \sigma^{2*} = \underset{\mu, \sigma^2}{\operatorname{argmax}} -N\log(\sqrt{2\pi\sigma^2}) - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}$$

To find a maximum, we need to ensure that the derivatives w.r.t. μ and σ^2 are *simultaneously* zero. To do this, we should start by finding both partial derivatives. We start with μ:

$$\frac{\partial \mathcal{L}}{\partial \mu} = \frac{\partial(-N\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2})}{\partial \mu} = 0$$
$$2\sum_{i=1}^{N} \frac{(x_i - \mu)}{2\sigma^2} = 0$$
$$\mu^* = \frac{1}{N}\sum_{i=1}^{N} x_i$$

We find that this can be set to zero *independently* of σ^2.

For variance estimation we differentiate the likelihood function w.r.t σ^2 at the obtained μ^* and set that to zero:

$$\frac{\partial(-N\log(\sqrt{2\pi\sigma^2}) - \sum_{i=1}^{N}\frac{(x_i-\mu^*)^2}{2\sigma'^2})}{\partial\sigma^2} = 0 \tag{45}$$

$$-\frac{N}{2\sigma^2} + \sum_{i=1}^{N}\frac{(x_i-\mu^*)^2}{2\sigma^4} = 0 \tag{46}$$

$$\sigma^{2*} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu^*)^2 \tag{47}$$

Finally, we compute the Hessian to verify that our solution is a maximum:

$$H = \begin{bmatrix} \frac{\partial^2(-N\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma'^2})}{\partial\mu^2} & \frac{\partial^2(-N\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2})}{\partial(\sigma^2)\partial\mu} \\ \frac{\partial^2(-N\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma'^2})}{\partial\mu\partial(\sigma^2)} & \frac{\partial^2(-N\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2})}{\partial(\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} -\frac{N}{\sigma^2} & -\sum_{i=1}^{N}\frac{(x_i-\mu)}{\sigma^4} \\ -\sum_{i=1}^{N}\frac{(x_i-\mu)}{\sigma^4} & \frac{N}{2\sigma^4} - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{\sigma^6} \end{bmatrix}$$

Determinant of the computed Hessian at μ^*, σ^{2*}:

$$\det(H) = -\frac{N^2}{2(\sigma^{2*})^3} + \frac{N}{(\sigma^{2*})^4}\sum_{i=1}^{N}(x_i-\mu^*)^2 - \frac{1}{(\sigma^{2*})^4}(\sum_{i=1}^{N}(x_i-\mu^*))^2 \tag{48}$$

$$= -\frac{N^2}{2(\sigma^{2*})^3} + \frac{N^2}{(\sigma^{2*})^4}\sum_{i=1}^{N}\frac{(x_i-\mu^*)^2}{N} - \frac{1}{(\sigma^{2*})^4}(\sum_{i=1}^{N}(x_i-\mu^*))^2 \tag{49}$$

$$= \frac{N^2}{2(\sigma^{2*})^3} - \frac{N}{(\sigma^{2*})^3} \tag{50}$$

$$= \frac{N^2-2N}{2(\sigma^{2*})^3} \tag{51}$$

For $N \geq 2$, we have that $\det(H) \geq 0$. Keeping in mind that $\det(H) = \prod_n \lambda_n$, where λ_n are the eigenvalues of H, and that $H[0,0]$ is negative, we conclude the μ^* and σ^{2*} results in maximizing logliklihood of the distribution.

Question 18 Let X be data matrix of all inputs with $\mathbf{y}$ as the outputs and $f(\mathbf{x};\mathbf{w},\beta) = \mathbf{w}^T\mathbf{x} + \beta = \theta^T[\mathbf{x}, 1]$ be a linear regression model, where $\theta = [\mathbf{w}, \beta]$ is a vector of parameters including bias.

Let's consider two cases for estimating parameters θ

case 1: Maximum Likelihood Estimate (MLE): $\theta^* = \text{argmax}_\theta\, \mathcal{L}(y; f(\mathbf{x};\theta), \mathbb{I}\sigma^2)$

case 2: Minimum Squared Estimate (MSE): $\theta^* = \text{argmin}_\theta \frac{1}{N}\sum_{i=1}^{N}(y_i - f(\mathbf{x}_i;\theta))^2$

Task is to show that both estimates are same.

Consider case 1: $\theta^* = \text{argmax}_\theta\, \mathcal{L}(y; f(\mathbf{x};\theta), \mathbb{I}\sigma^2)$

$$\Rightarrow \frac{\partial\mathcal{L}(y; f(\mathbf{x};\theta), \mathbb{I}\sigma^2)}{\partial\theta)} = 0$$

$$\Rightarrow \frac{-0.5\sigma^{-2}\partial\sum_{i=1}^{N}(y_i - \theta^T\mathbf{x}_i)^2}{\partial\theta} = 0$$

$$\Rightarrow \frac{\partial\frac{1}{N}\sum_{i=1}^{N}(y_i - \theta^T\mathbf{x}_i)^2}{\partial\theta} = 0$$

As the above expression is linear in θ, it can be also be seen as: $\theta^* = \text{argmin}_\theta \frac{1}{N}\sum_{i=1}^{N}(y_i - f(\mathbf{x}_i;\theta))^2$, resulting in MSE estimates.

Question 19

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

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = \sin(x^3)\frac{\mathrm{d}\log(x^4)}{\mathrm{d}x} + \log(x^4)\frac{\mathrm{d}\sin(x^3)}{\mathrm{d}x}$$
$$\Rightarrow f'(x) = \sin(x^3)\frac{4}{x} + 12\log(x)\cos(x^3)x^2$$

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

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = \frac{\mathrm{d}(1+\exp(-x))^{-1}}{\mathrm{d}x}$$
$$\Rightarrow f'(x) = \frac{\mathrm{d}(1+\exp(-x))^{-1}}{\mathrm{d}(1+\exp(-x))}\frac{\mathrm{d}(1+\exp(-x))}{\mathrm{d}x}$$
$$\Rightarrow f'(x) = \frac{\exp(-x)}{(1+\exp(-x))^2}$$

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

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = \frac{\mathrm{d}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{\mathrm{d}x}$$
$$\Rightarrow f'(x) = -\frac{(x-\mu)}{\sigma^2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

17 Answers Lecture 2: Vector Differentiation

Question 20 – Circle Answer discussed in lectures.

Question 21 – Index notation Matrix-vector expressions to index notation:

a. $\mathrm{ABC}\mathbf{x} = \sum_{jkl} A_{ij}B_{jk}C_{kl}x_l$

b. $\mathrm{Tr}(\mathrm{A}) = \sum_i A_{ii}$

c. $\mathrm{Tr}(\mathrm{AB}) = \sum_{ij} A_{ij}B_{ji}$

d. $\mathbf{y}^\intercal \mathrm{A}^\intercal \mathbf{x} = \sum_{ij} y_i A_{ji} x_j$

Index notation to matrix-vector expressions:

a. $\sum_{ijk} A_{ij}B_{jk}C_{ki} = \mathrm{Tr}(\mathrm{ABC})$

b. $b_i + \sum_j A_{ij}b_j = \mathbf{b} + \mathrm{A}\mathbf{b}$

c. $x_i x_j = \mathbf{x}\mathbf{x}^T$

d. $\sum_j \delta_{ij}a_j = \mathbf{a}$

Question 22 – Index notation proofs

1. $\mathbf{x}^T A\mathbf{y} = \mathbf{y}^T A\mathbf{x}$ if $A = A^T$.

   Note that $A = A^T \implies a_{ij} = a_{ij}$. We first write the product in terms of two vectors, $\mathbf{x}$ and $A\mathbf{y}$, then rearrange the terms considering $A\mathbf{y}$ as a vector with components $(A\mathbf{y})_i = \sum_{j=1}^N a_{ij}y_j$, and use $a_{ij} = a_{ji}$:

$$\mathbf{x}^T A\mathbf{y} = \sum_{i=1}^N x_i(A\mathbf{y})_i = \sum_{i=1}^N x_i \sum_{j=1}^N a_{ij}y_j = \sum_{j=1}^N\sum_{i=1}^N y_j a_{ji} x_i = \sum_{j=1}^N y_j(A\mathbf{x})_j = \mathbf{y}^T A\mathbf{x}.$$

2. $\mathbf{x}^T\mathbf{y} = Tr(\mathbf{x}^T\mathbf{y}) = Tr(\mathbf{y}^T\mathbf{x}), \mathbf{x}, \mathbf{y} \in \mathbb{R}^D$.

   Considering $\mathbf{x}^T\mathbf{y} \in \mathbb{R}$, we have $\mathbf{x}^T\mathbf{y} = Tr(\mathbf{x}^T\mathbf{y})$. Then, we need to check whether $\mathbf{x}^T\mathbf{y} = \mathbf{y}^T\mathbf{x}$, which we can show by rearranging the terms of the vector product

$$\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{N} x_i y_i = \sum_{i=1}^{N} y_i x_i = \mathbf{y}^T\mathbf{x}.$$

3. $Tr(ABC) = Tr(CAB)$. We assume $A \in \mathbb{R}^{D\times E}, B \in \mathbb{R}^{E\times F}$, and $C \in \mathbb{R}^{F\times D}$.

   We start by inspecting the terms involving $Tr(ABC)$,

$$Tr(ABC) = \sum_{i=1}^{D} \left((AB)C\right)_{ii} = \sum_{i=1}^{D}\sum_{j=1}^{F} (AB)_{ij} c_{ji} = \sum_{i=1}^{D}\sum_{j=1}^{F}\sum_{k=1}^{E} a_{ik} b_{kj} c_{ji}.$$

   Just by swapping the summations we can get the following identity: $Tr(ABC) = Tr(CAB) = Tr(BCA)$. We show $Tr(ABC) = Tr(CAB)$ as an example:

$$Tr(ABC) = \sum_{i=1}^{D}\sum_{j=1}^{F}\sum_{k=1}^{E} a_{ik} b_{kj} c_{ji} = \sum_{j=1}^{F}\sum_{k=1}^{E}\sum_{i=1}^{D} c_{ji} a_{ik} b_{kj} = \sum_{j=1}^{F}\sum_{k=1}^{E} (CA)_{jk} b_{kj} = \sum_{j=1}^{F} (CAB)_{jj} = Tr(CAB).$$

Question 23 – MML 5.5-5.6

a $f(\mathbf{x}) = \sin(x_1)\cos(x_2), \quad \mathbf{x} \in \mathbb{R}^2$

$$\frac{\partial f}{\partial \mathbf{x}} \in \mathbb{R}^{1\times 2}$$

$$\begin{aligned}\frac{\partial f}{\partial \mathbf{x}} &= \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}\right] \\ &= \left[\cos(x_1)\cos(x_2), -\sin(x_1)\sin(x_2)\right]\end{aligned}$$

b $f(\mathbf{x}) = \mathbf{x}^T\mathbf{y}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$\frac{\partial f}{\partial \mathbf{x}} \in \mathbb{R}^{1\times n}$$

   We can solve this directly using basic rules of vector calculus

$$\frac{\partial f}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}^T\mathbf{y})}{\partial \mathbf{x}} = \mathbf{y}^T$$

   We can confirm this result holds with index notation. First, let us calculate the value $f(\mathbf{x})$

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

$$\frac{\partial f}{\partial x_j} = \frac{\partial}{\partial x_j}\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} \frac{\partial x_j}{\partial x_i} y_i = \sum_{i=1}^{n} \delta_{ij} y_i = y_j$$

$$\frac{\partial f}{\partial \mathbf{x}} = \left[\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right] = [y_1, \ldots, y_n] = \mathbf{y}^T$$

c $\mathbf{f}(x) = \mathbf{x}\mathbf{x}^T, \quad \mathbf{x} \in \mathbb{R}^n$

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} \in \mathbb{R}^{(n\times n)\times n}$$

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = C \quad \text{where } C \text{ is a 3D tensor.}$$

$$c_{ijk} = \frac{\partial f(\mathbf{x})_{ij}}{\partial x_k}$$

$$\mathbf{x}\mathbf{x}^T = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \begin{pmatrix} x_1 & \dots & x_n \end{pmatrix} = \begin{pmatrix} x_1^2 & x_1 x_2 & \dots & x_1 x_n \\ x_2 x_1 & x_2^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ x_n x_1 & \dots & \dots & x_n^2 \end{pmatrix}$$

$$c_{ijk} = \frac{\partial(x_i x_j)}{\partial x_k} = \frac{\partial x_i}{\partial x_k} x_j + \frac{\partial x_j}{\partial x_k} x_i = \delta_{ik} x_j + \delta_{jk} x_i = \begin{cases} 0 & \text{if } k \neq i \text{ and } k \neq j \\ x_i & \text{if } k = j \text{ and } i \neq j \\ x_j & \text{if } k = i \text{ and } i \neq j \\ 2x_i & \text{if } k = i = j \end{cases}$$

d $f(\mathbf{t}) = \sin\big(\log(\mathbf{t}^T\mathbf{t})\big) \quad \mathbf{t} \in \mathbb{R}^D$

We directly apply the chain rule

$$\frac{\partial f}{\partial \mathbf{t}} = \frac{\partial \sin\big(\log(\mathbf{t}^T\mathbf{t})\big)}{\partial \log(\mathbf{t}^T\mathbf{t})} \cdot \frac{\partial \log(\mathbf{t}^T\mathbf{t})}{\partial(\mathbf{t}^T\mathbf{t})} \cdot \frac{\partial(\mathbf{t}^T\mathbf{t})}{\partial \mathbf{t}}$$

All of the terms are one dimensional except for $\frac{\partial(\mathbf{t}^T\mathbf{t})}{\partial \mathbf{t}} \in \mathbb{R}^{1\times D}$. Let us calculate the value using the notation for vector calculus in the lectures. As in 5.5, we first calculate the value of $\mathbf{t}^T\mathbf{t}$ and its derivative w.r.t. t_i.

$$\mathbf{t}^T\mathbf{t} = \sum_{i=1}^{D} t_i^2, \qquad \frac{\partial(\mathbf{t}^T\mathbf{t})}{\partial t_i} = 2t_i$$

$$\frac{\partial(\mathbf{t}^T\mathbf{t})}{\partial \mathbf{t}} = \left[\frac{\partial(\mathbf{t}^T\mathbf{t})}{\partial t_1}, \dots, \frac{\partial(\mathbf{t}^T\mathbf{t})}{\partial t_D}\right] = [2t_1 \dots, 2t_D] = 2\mathbf{t}^T$$

We can now use this result to proceed with the derivative of $f(\mathbf{t})$.

$$\frac{\partial f}{\partial \mathbf{t}} = \cos\big(\log(\mathbf{t}^T\mathbf{t})\big) \cdot \frac{1}{\mathbf{t}^T\mathbf{t}} \cdot 2\mathbf{t}^T$$

$$\frac{\partial f}{\partial \mathbf{t}} = 2\mathbf{t}^T \frac{\cos\big(\log(\mathbf{t}^T\mathbf{t})\big)}{\mathbf{t}^T\mathbf{t}}$$

e $f(X) = tr(AXB), \quad A \in \mathbb{R}^{D\times E}, X \in \mathbb{R}^{E\times F}, B \in \mathbb{R}^{F\times D}$

Use index notation:

$$f(X) = tr(AXB) = \sum_{i=1}^{D} (AXB)_{ii}$$

In order to fully compute $f(X)$, we need to calculate $(AXB)_{ii}$

$$(AXB)_{ii} = \sum_{k=1}^{F} (AX)_{ik} b_{ki} = \sum_{k=1}^{F} \left(\sum_{l=1}^{E} a_{il} x_{lk}\right) b_{ki}$$

Thus

$$f(X) = \sum_{i=1}^{D} \sum_{k=1}^{F} \sum_{l=1}^{E} a_{il} x_{lk} b_{ki}$$

Now we can just calculate the derivative using index notation

$$\frac{\partial f}{\partial x_{nm}} = \frac{\partial}{\partial x_{nm}} \sum_{i=1}^{D} \sum_{k=1}^{F} \sum_{l=1}^{E} a_{il} x_{lk} b_{ki} = \sum_{i=1}^{D} \sum_{k=1}^{F} \sum_{l=1}^{E} a_{il} \frac{\partial x_{lk}}{\partial x_{nm}} b_{ki} = \sum_{i=1}^{D} \sum_{k=1}^{F} \sum_{l=1}^{E} a_{il} \delta_{ln} \delta_{km} b_{ki}$$

Notice that in the last expression, all the terms in the summation cancel except when $k = m$ and $l = n$. Therefore

$$\frac{\partial f}{\partial x_{nm}} = \sum_{i=1}^{D} a_{in} b_{mi} = \sum_{i=1}^{D} b_{mi} a_{in} = (BA)_{mn}$$

Using this last result, we can calculate the derivative w.r.t. X.

$$\frac{\partial f}{\partial X} = (BA)^T = A^T B^T$$

Alternative proof: Use properties 4.19 and 5.100 from the MML book. From 4.19

$$f(X) = tr(AXB) = tr(XBA) = tr(XC), \quad C = BA$$

and from 5.100

$$\frac{\partial f}{\partial X} = \frac{\partial tr(XC)}{\partial X} = tr\left(\frac{\partial (XC)}{\partial X}\right), \quad \text{where } \frac{\partial (XC)}{\partial X} \in \mathbb{R}^{(E\times E)\times(E\times F)}$$

We need to calculate $\frac{\partial (XC)_{ij}}{\partial x_{kl}}$, and we find convenient to write the pairs i, j of the product IXC, where $I \in \mathbb{R}^{E\times E}$ is the identity matrix.

$$(IXC)_{ij} = \sum_{e=1}^{E}\sum_{f=1}^{F} \delta_{ie} x_{ef} c_{fj}$$

$$\frac{\partial (XC)_{ij}}{\partial x_{kl}} = \frac{\partial (IXC)_{ij}}{\partial x_{kl}} = \delta_{ik} c_{lj}$$

in the prevous expression, all the terms in the sum vanish except the ones that contain x_{kl} in it. Now, we take into account the definition of the trace for any 4D tensor $T \in \mathbb{R}^{(N\times N)\times(P\times Q)}$ given in the MML book:

$$tr(T)_{ij} = \sum_{k=1}^{N} a_{kkij}, \quad \text{where } tr(T) \in \mathbb{R}^{P\times Q}$$

We use this definition to calculate our result.

$$tr\left(\frac{\partial (XC)}{\partial X}\right)_{ij} = \sum_{k=1}^{E} \frac{\partial (XC)_{kk}}{\partial x_{ij}} = \sum_{k=1}^{E} \delta_{ki} c_{jk} = c_{ji}$$

all the terms will be 0 except when $k = i$.

$$tr\left(\frac{\partial (XC)}{\partial X}\right) = C^T = (BA)^T = A^T B^T$$

Question 24 – MML 5.7-5.8: Chain rule

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

$$\frac{\partial f}{\partial \mathbf{x}} = \frac{\partial f}{\partial z}\frac{\partial z}{\partial \mathbf{x}} = \frac{\partial \log(1+z)}{\partial z}\frac{\partial (\mathbf{x}^T\mathbf{x})}{\partial \mathbf{x}} = \frac{2\mathbf{x}^T}{1+z} = \frac{2\mathbf{x}^T}{1+\mathbf{x}^T\mathbf{x}}$$

Dimensions are

$$\frac{\partial f}{\partial \mathbf{x}} \in \mathbb{R}^D, \quad \frac{\partial f}{\partial z} \in \mathbb{R}, \quad \frac{\partial z}{\partial \mathbf{x}} \in \mathbb{R}^D$$

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

$$\frac{\partial f}{\partial \mathbf{x}} = \frac{\partial f}{\partial \mathbf{z}}\frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \frac{\partial \sin(\mathbf{z})}{\partial \mathbf{z}}\frac{\partial (A\mathbf{x}+\mathbf{b})}{\partial \mathbf{x}}$$

Notice that $\frac{\partial f}{\partial \mathbf{z}} \in \mathbb{R}^{E\times E}$. We already know that $\sin(\cdot)$ is applied to each element independently, thus

$$\frac{\partial f_i}{\partial z_j} = \begin{cases} 0 & \text{if } i \neq j \\ \cos(z_i) & \text{if } i = j \end{cases}$$

We also have $\frac{\partial \mathbf{z}}{\partial \mathbf{x}} \in \mathbb{R}^{E\times D}$. Although this has already shown in the lectures, let us review the result $\frac{\partial \mathbf{z}}{\partial \mathbf{x}}$ using the notation of the course.

$$z_i = \sum_{j=1} a_{ij}x_j + b_i$$

We can now easily compute $\frac{\partial z_i}{\partial x_j}$

$$\frac{\partial z_i}{\partial x_j} = a_{ij}, \quad \frac{\partial \mathbf{z}}{\partial \mathbf{x}} = A$$

Let us use all the previous results to compute the derivative of $f(\mathbf{z})$ w.r.t. $\mathbf{x}$.

$$\frac{\partial f}{\partial \mathbf{x}} = diag(\cos(\mathbf{z}))A, \quad \text{where e.g. } diag(\mathbf{v}) = \begin{pmatrix} v_1 & 0 & \dots & 0 \\ 0 & v_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & v_N \end{pmatrix}, \quad \mathbf{v} \in \mathbb{R}^N$$

$$\frac{\partial f}{\partial \mathbf{x}} = diag(\cos(A\mathbf{x} + \mathbf{b}))A$$

Dimensions are

$$\frac{\partial f}{\partial \mathbf{x}} \in \mathbb{R}^{E\times D}, \quad \frac{\partial f}{\partial \mathbf{z}} \in \mathbb{R}^{E\times E}, \quad \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \in \mathbb{R}^{E\times D}$$

c. $f(z) = \exp(-\frac{1}{2}z), \quad z = g(\mathbf{y}) = \mathbf{y}^T S^{-1}\mathbf{y}, \quad \mathbf{y} = h(\mathbf{x}) = \mathbf{x} - \boldsymbol{\mu}, \quad \mathbf{x}, \boldsymbol{\mu} \in \mathbb{R}^D, S \in \mathbb{R}^{D\times D}$

$$\begin{aligned}\frac{\partial f}{\partial \mathbf{x}} &= \frac{\partial f}{\partial z}\frac{\partial z}{\partial \mathbf{y}}\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial \exp(-\frac{1}{2}z)}{\partial z}\frac{\partial (\mathbf{y}^T S^{-1}\mathbf{y})}{\partial \mathbf{y}}\frac{\partial (\mathbf{x}-\boldsymbol{\mu})}{\partial \mathbf{x}} = \exp\Big(-\frac{1}{2}z\Big)\Big(-\frac{1}{2}\Big)\mathbf{y}^T(S^T + S^{-T})I \\ &= -\frac{1}{2}\exp\Big(-\frac{1}{2}\Big((\mathbf{x}-\boldsymbol{\mu})^T S^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big)\Big)(\mathbf{x}-\boldsymbol{\mu})^T(S^{-1} + S^{-T})\end{aligned}$$

where $S^{-T} = \big(S^{-1}\big)^T$, and we use (5.107) to calculate $\frac{\partial (\mathbf{y}^T S^{-1}\mathbf{y})}{\partial \mathbf{y}}$.
Dimensions are

$$\frac{\partial f}{\partial \mathbf{x}} \in \mathbb{R}^D, \quad \frac{\partial f}{\partial z} \in \mathbb{R}, \quad \frac{\partial z}{\partial \mathbf{y}} \in \mathbb{R}^D, \quad \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \in \mathbb{R}^{D\times D}$$

d. $f(\mathbf{x}) = tr(\mathbf{x}\mathbf{x}^T + \sigma^2 I), \quad \mathbf{x} \in \mathbb{R}^D$
Let us expand $f(x)$.

$$\begin{aligned}f(x) &= \sum_{i=1}^{D}\Big((\mathbf{x}\mathbf{x}^T)_{ii} + \sigma^2\Big) \\ &= \sum_{i=1}^{D}(\mathbf{x}\mathbf{x}^T)_{ii} + D\sigma^2 = \sum_{i=1}^{D} x_i^2 + D\sigma^2\end{aligned}$$

We already know that $(\mathbf{x}\mathbf{x}^T)_{ij} = x_i x_j$. Therefore

$$\frac{\partial f}{\partial \mathbf{x}} = \frac{\partial\Big(\sum_{i=1}^{D} x_i^2 + D\sigma^2\Big)}{\partial \mathbf{x}} = 2\mathbf{x}^T$$

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

$$\begin{aligned}\frac{\partial f}{\partial \mathbf{x}} &= \frac{\partial f}{\partial \mathbf{z}}\frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \frac{\partial \tanh(\mathbf{z})}{\partial \mathbf{z}}\frac{\partial (A\mathbf{x} + \mathbf{b})}{\partial \mathbf{x}} = diag\big(1 - \tanh^2(\mathbf{z})\big)\mathbf{A} \\ &= diag\big(1 - \tanh^2(A\mathbf{x} + \mathbf{b})\big)A\end{aligned}$$

where we used $\frac{d\tanh(v)}{dv} = 1 - \tanh^2(v)$.
Dimensions are

$$\frac{\partial f}{\partial \mathbf{x}} \in \mathbb{R}^{M\times N}, \quad \frac{\partial f}{\partial \mathbf{z}} \in \mathbb{R}^{M\times M}, \quad \frac{\partial z}{\partial \mathbf{x}} \in \mathbb{R}^{M\times N}$$

f. $f(A) = \mathbf{x}^\intercal A\mathbf{x}, \quad A = \mathbf{x}\mathbf{x}^\intercal, \quad A \in \mathbb{R}^{N\times N}$

Note that A is symmetric. We apply the chain rule straightforwardly.

$$\frac{df}{d\mathbf{x}} = \frac{\partial f}{\partial \mathbf{x}} + \frac{\partial f}{\partial A}\frac{\partial A}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}^T A\mathbf{x})}{\partial \mathbf{x}} + \sum_{i,j}\frac{\partial(\mathbf{x}^T A\mathbf{x})}{\partial a_{ij}}\frac{\partial a_{ij}}{\partial \mathbf{x}}$$

The first term can be computed using vector calculus rules

$$\frac{\partial(\mathbf{x}^T A\mathbf{x})}{\partial \mathbf{x}} = 2A\mathbf{x} = 2\mathbf{x}\mathbf{x}^T\mathbf{x} = 2||\mathbf{x}||^2\mathbf{x}.$$

We can compute the second term using index notation, i.e. the derivative with respect to x_k.

$$\begin{aligned}\sum_{i,j}\frac{\partial(\mathbf{x}^T A\mathbf{x})}{\partial a_{ij}}\frac{\partial a_{ij}}{\partial x_k} &= \sum_{i,j} a_{ij}(\delta_{ik}x_j + \delta_{jk}x_i) = \sum_{i,j} a_{ij}\delta_{ik}x_j + \sum_{i,j} a_{ij}\delta_{jk}x_i \\ &= \sum_j a_{kj}x_j + \sum_i a_{ik}x_i = 2\sum_i x_k x_i^2 = 2||\mathbf{x}||^2 x_k.\end{aligned}$$

Where we used vector calculus rules and Question 23 – MML 5.5-5.6 c) to derive the following

$$\frac{\partial(\mathbf{x}^T A\mathbf{x})}{\partial A} = \mathbf{x}\mathbf{x}^T = A, \quad \frac{\partial a_{ij}}{\partial x_k} = \delta_{ik}x_j + \delta_{jk}x_i, \quad a_{ij} = x_i x_j.$$

In conclusion, we have

$$\frac{df}{d\mathbf{x}} = \frac{\partial(\mathbf{x}^T A\mathbf{x})}{\partial \mathbf{x}} + \sum_{i,j}\left(\frac{\partial(\mathbf{x}^T A\mathbf{x})}{\partial a_{ij}}\right)\frac{\partial a_{ij}}{\partial \mathbf{x}} = 2||\mathbf{x}||^2\mathbf{x} + 2||\mathbf{x}||^2\mathbf{x} = 4||\mathbf{x}||^2\mathbf{x}.$$

Extra: one can check the previous result is true by simplifying the initial form of f, which we maintained for illustrative purposes, i.e. $f(\mathbf{x}) = ||\mathbf{x}||^4$. We can compute the derivative using index notation

$$\frac{df}{dx_i} = \frac{d}{dx_i}\left(\sum_{i=1}^N x_i^2\right)^2 = 2||\mathbf{x}||^2 2x_i = 4||\mathbf{x}||^2 x_i,$$

which in vector form is expressed as follows

$$\frac{df}{dx} = 4||\mathbf{x}||^2\mathbf{x}.$$

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

$$L(\boldsymbol{\theta}) = |\mathbf{y} - \Phi(X)\boldsymbol{\theta}|^2\,, \qquad \frac{\mathrm{d}L}{\mathrm{d}\boldsymbol{\theta}} = 2(\Phi(X)\boldsymbol{\theta} - \mathbf{y})^\intercal\Phi(X)\,. \tag{52}$$

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:

$$\frac{\partial}{\partial\theta_j}\left(\frac{\partial L}{\partial\theta_i}\right) = \frac{\partial}{\partial\theta_j}\left(2\sum_k\left(\sum_m \Phi_{km}\theta_m - y_k\right)\Phi_{ki}\right) = \frac{\partial}{\partial\theta_j}\left(2\sum_k\left(\sum_m \Phi_{km}\theta_m - y_k\right)\Phi_{ki}\right) \tag{53}$$

$$= 2\sum_{km}\Phi_{km}\delta_{mj}\Phi_{ki} = 2\sum_k \Phi_{kj}\Phi_{ki}\,, \tag{54}$$

$$\implies \mathrm{H}_{\boldsymbol{\theta}}(L) = 2\Phi(X)^\intercal\Phi(X)\,. \tag{55}$$

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

$$\mathbf{v}^\intercal\mathrm{H}\mathbf{v} = 2\mathbf{v}^\intercal\Phi(X)^\intercal\Phi(X)\mathbf{v} \tag{56}$$

$$= \mathbf{w}^\intercal \mathbf{w} = \sum_i w_i^2 \,, \qquad\qquad \text{with } \mathbf{v} = \Phi(X)\mathbf{v} \,. \tag{57}$$

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

$$\mathbf{v}^\intercal \mathrm{H}\mathbf{v} = \mathbf{v}^\intercal \mathrm{Q}\Lambda\mathrm{Q}^{-1}\mathbf{v} \qquad \text{(eigenvalue decomposition)} \tag{58}$$

$$= \mathbf{v}^\intercal \mathrm{Q}\Lambda\mathrm{Q}^\intercal\mathbf{v} \qquad (\mathrm{H} = \mathrm{H}^\intercal, \text{ so } \mathrm{Q}\Lambda\mathrm{Q}^{-1} = (\mathrm{Q}\Lambda\mathrm{Q}^{-1})^\intercal, \text{ so } \mathrm{Q}^{-1} = Q^\intercal) \tag{59}$$

$$= \mathbf{z}^\intercal \Lambda \mathbf{z} \,, \tag{60}$$

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.

18 Answers Lecture 3: Automatic Differentiation

18.1 Question 26 – Product rule

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

$$v_0 = x \qquad \frac{\partial v_0}{\partial x} = 1 \tag{61}$$

$$v_1 = a(x) \qquad \frac{\partial v_1}{\partial x} = \frac{\partial v_1}{\partial v_0}\frac{\partial v_0}{\partial x} = \frac{\partial a(x)}{\partial x} \tag{62}$$

$$v_2 = b(x) \qquad \frac{\partial v_2}{\partial x} = \frac{\partial v_2}{\partial v_0}\frac{\partial v_0}{\partial x} = \frac{\partial b(x)}{\partial x} \tag{63}$$

$$v_3 = v_1 \cdot v_2 \qquad \frac{\partial v_3}{\partial x} = \sum_{j \in \mathtt{inputs}(3)} \frac{\partial v_3}{\partial v_j}\frac{\partial v_j}{\partial x} = v_2 \frac{\partial a(x)}{\partial x} + v_1 \frac{\partial b(x)}{\partial x} \,. \tag{64}$$

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

$$\frac{\mathrm{d}f}{\mathrm{d}x} = b(x)\frac{\mathrm{d}a(x)}{\mathrm{d}x} + a(x)\frac{\mathrm{d}b(x)}{\mathrm{d}x} \,. \tag{65}$$

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.

18.2 Question 27 – Multivariate Autodiff

This is a rather big question. It is designed to test *every* single differentiation skill we taught, and how it fits together. As such, I think it's great preparation for the exam. However, it is long. You may want to try to do parts of it yourself, and perhaps part in a group.

[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.

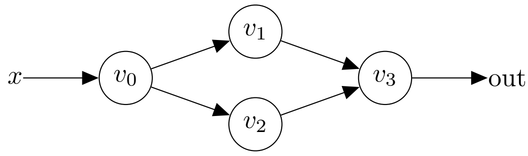

Figure 1: Computational graph for Question 26 – Product rule.

Part a)

$$
\begin{aligned}
D_a &: \mathbb{R}^{N\times 2} && \to \mathbb{R}^{N\times N} && (66)\\
\Lambda_a &: \mathbb{R}^{N\times N} \times \mathbb{R}^2 && \to \mathbb{R}^{N\times N} && (67)\\
K_a &: \mathbb{R}^{N\times N} && \to \mathbb{R}^{N\times N} && (68)\\
f &: \mathbb{R}^{N\times N} \times \mathbb{R}^{N\times N} && \to \mathbb{R} && (69)
\end{aligned}
$$

Part b) See fig. 2. We use names evident from the question for some nodes, but give new names to some additional intermediate notes. How should one decide to split up the computational graph? You should at least split according to the sequence of functions given in the question. Hence why we have the nodes named the same as intermediate functions. We should also split up depending on for which operations gradients are defined in our autodiff framework. So for example, in the computation of f, we need to compute a matrix sum, matrix inverse, and then a vector quadratic. We know how to differentiate each one of these operations separately, so we split them up so we can chain them together. For the vector quadratic, we have a choice. We could see this as a single operation, or we could split this into two matrix-vector multiplications. Autodiff frameworks probably do the latter, but for this question, we consider it done together. In an exam situation, the split will be defined clearly for you, or you would explicitly be asked to state your assumptions when making splits.

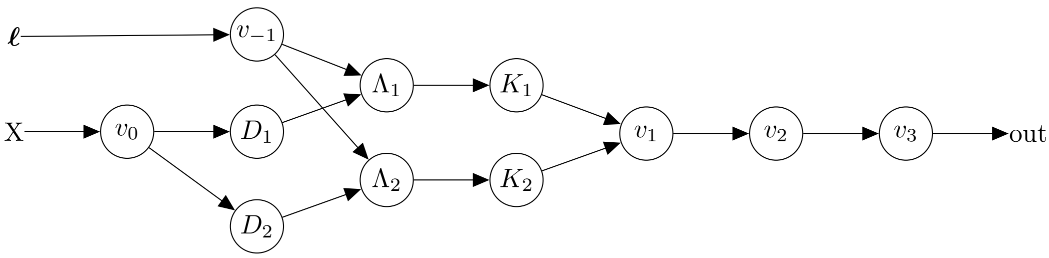

Figure 2: Computational graph for Question 27 – Multivariate Autodiff, where we define $v_1 = \mathrm{K}_1 + \mathrm{K}_2$, $v_2 = v_1^{-1}$, and $v_3 = \mathbf{y}^\intercal v_1 \mathbf{y}$.

Part c) Let's first consider **forward mode**. For the derivatives w.r.t. X, we initialise as:

Primal	Forward tangent	Notes	
$v_{-1} = \boldsymbol{\ell}$	$\dot{v}_{-1,iab} = \dfrac{\partial [v_{-1}]_i}{\partial X_{ab}} = 0$	$v_{-1} \in \mathbb{R}^{2\times(N\times 2)}, O(N)$	(70)
$v_0 = X$	$\dot{v}_0 = \dfrac{\partial [v_0]_{ij}}{\partial X_{ab}} = \delta_{ia}\delta_{jb}$	$\dot{v}_0 \in \mathbb{R}^{(N\times 2)\times(N\times 2)}, O(N^2)$	(71)

While for the derivatives w.r.t. $\boldsymbol{\ell}$, we initialise as:

Primal	Forward tangent	Notes	
$v_{-1} = \boldsymbol{\ell}$	$\dot{v}_{-1,ia} = \dfrac{\partial [v_{-1}]_i}{\partial \ell_a} = \delta_{ai}$	$v_{-1} \in \mathbb{R}^{2\times 2}, O(1)$	(72)
$v_0 = X$	$\dot{v}_0 = \dfrac{\partial [v_0]_{ij}}{\partial \ell_a} = 0$	$\dot{v}_0 \in \mathbb{R}^{(N\times 2)\times 2}, O(N)$	(73)
			(74)

Then we follow the forward mode computations. In the following calculation, we consider the gradient of X, so all forward tangents will have indices a and b corresponding to values in X. The computations w.r.t. $\boldsymbol{\ell}$ are exactly the same, but without the index b.

Primal	Forward tangent	Notes	
$D_z = \ldots$	$\frac{\partial [D_z]_{nm}}{\partial v_{0ij}} = \frac{\partial}{\partial v_{0ij}}(v_{0nz} - v_{0mz})^2$	$\dot{D}_z \in \mathbb{R}^{(N\times N)\times(N\times 2)}$	
	$= 2(v_{0nz} - v_{0mz})(\delta_{ni}\delta_{zj} - \delta_{mi}\delta_{zj})$		(75)
	$\dot{D}_{znmab} = \left[\frac{\partial D_z}{\partial v_0}\dot{v}_0\right]_{nmab}$	$O(N)$ for sum, so total $O(N^4)$.	(76)
	$= 2(v_{0nz} - v_{0mz})(\dot{v}_{0nzab} - \dot{v}_{0mzab})$	Structure allows $O(N^3)$.	(77)
$\Lambda_z = -\frac{D_z}{2v_{-1z}^2}$	$\frac{\partial \Lambda_{zij}}{\partial D_{znm}} = -\frac{1}{2v_{-1z}^2}\delta_{in}\delta_{jm}$	$\dot{\Lambda}_z \in \mathbb{R}^{(N\times N)\times(N\times 2)}$	(78)
	$\frac{\partial \Lambda_{zij}}{\partial v_{-1,k}} = \frac{D_{zij}}{v_{-1z}^3}\delta_{zk}$		(79)
	$\dot{\Lambda}_{zijab} = \left[\frac{\partial \Lambda_z}{\partial D_z}\dot{D}_z + \frac{\partial \Lambda_z}{\partial v_{-1}}\dot{v}_{-1}\right]_{ijab}$	$O(N^2)$ for sum, so total $O(N^5)$.	(80)
	$= -\frac{1}{2v_{-1z}^2}\dot{D}_{zijab} + \frac{D_{zij}}{v_{-1z}^3}\dot{v}_{-1,zab}$	Structure allows $O(N^3)$.	(81)
$K_z = \exp(\Lambda_z)$	$\frac{\partial K_{znm}}{\partial \Lambda_{zij}} = \exp(\Lambda_{zij})\delta_{ni}\delta_{mj}$	$\dot{K}_z \in \mathbb{R}^{(N\times N)\times(N\times 2)}$	(82)
	$\dot{K}_{znmab} = \exp(\Lambda_{znm})\dot{\Lambda}_{znmab}$	Structure allows $O(N^3)$.	(83)
$v_1 = \mathrm{K}_1 + \mathrm{K}_2$	$\frac{\partial v_{1ij}}{\partial K_{znm}} = \delta_{in}\delta_{jm}$		(84)
	$\dot{v}_1 = \frac{\partial v_1}{\partial \mathrm{K}_1}\dot{\mathrm{K}}_1 + \frac{\partial v_1}{\partial \mathrm{K}_2}\dot{\mathrm{K}}_2$		(85)
	$\dot{v}_{1,ijab} = \dot{K}_{1,ijab} + \dot{K}_{z,ijab}$		(86)
$v_2 = v_1^{-1}$	$\frac{\partial v_2}{\partial v_1} = -v_1^{-1}\frac{\partial v_1}{\partial v_1}v_1^{-1}$	This is an identity.	(87)
	$\frac{\partial v_{2,nm}}{\partial v_{1,ij}} = \sum_{pq} -v_{1,np}^{-1}\delta_{pi}\delta_{qj}v_{1,qm}^{-1}$	$A_{pq}^{-1} = [A^{-1}]_{pq}$	(88)
	$\dot{v}_{2,nmab} = \sum_{pqij} -v_{1,np}^{-1}\delta_{pi}\delta_{qj}v_{1,qm}^{-1}\dot{v}_{1,ijab}$		(89)
	$= -\sum_p v_{1,np}\sum_q \dot{v}_{1,pqab}v_{1,qm}$	$O(N^4)$	(90)
$v_3 = \mathbf{y}^\mathsf{T} v_2 \mathbf{y}$	$\frac{\partial v_3}{\partial v_{2,nm}} = y_n y_m$		(91)
	$\dot{v}_{3,ab} = \sum_{nm} y_n y_m \dot{v}_{2,nmab}$	Done!	(92)

The bottleneck computational cost we see is $O(N^4)$. Let's compare this to the cost of computing the objective function. We see a few steps that are $O(N^2)$ (like computing the distances), and the matrix inverse, which is $O(N^3)$. We know that forward mode gives us a guarantee of computing the gradient that is linear in the number of inputs, multiplied by the cost of evaluating the function. Since we have $O(N)$ inputs, we see that the $O(N^4)$ bottleneck is consistent with this!

The bottleneck comes from the identity for differentiating an inverse. If we would have differentiated through the actual algorithm that computes the inverse, rather than using the identity, we would have obtained the same computational cost.

Let's next consider **reverse mode**. We use the bar notation to denote the derivative w.r.t. the input of a node, i.e. $\bar{v}_3 = \frac{\partial f}{\partial v_2}$. The primal mode column is executed top to bottom, after which the reverse adjoint column is executed bottom to top.

Primal	Reverse adjoint	Notes	
$D_z = \dots$	$\bar{v}_{0,ij} = 2\left(\sum_m \bar{D}_{jim}(v_{0ij} - v_{0mj}) - \sum_n \bar{D}_{jni}(v_{0nj} - v_{0,ij})\right)$	Done! $O(N^2)$	(93)
	$\frac{\partial D_{znm}}{\partial v_{0ij}} = 2(v_{0nz} - v_{0mz})(\delta_{ni}\delta_{zj} - \delta_{mi}\delta_{zj})$		(94)
$\Lambda_z = -\frac{D_z}{2v_{-1z}^2}$	$\bar{D}_{znm} = -\bar{\Lambda}_{znm}\frac{D_{znm}}{2v_{-1z}^2}$		(95)
	$\bar{v}_{-1,z} = \sum_{ij} \bar{\Lambda}_{zij}\frac{D_{zij}}{v_{-1,z}^3}$	Done! $O(N^2)$	(96)
	$\frac{\partial \Lambda_{zij}}{\partial D_{znm}} = -\frac{D_{znm}}{2v_{-1z}^2}\delta_{in}\delta_{jm}$		(97)
	$\frac{\partial \Lambda_{zij}}{\partial v_{-1,k}} = \frac{D_{zij}}{v_{-1z}^3}\delta_{zk}$		(98)
$K_z = \exp(\Lambda_z)$	$\bar{\Lambda}_{zij} = \bar{K}_{zij}K_{zij}$		(99)
	$\frac{\partial K_{znm}}{\partial \Lambda_{zij}} = \exp(\Lambda_{zij})\delta_{ni}\delta_{mj}$		(100)
$v_1 = \mathrm{K}_1 + \mathrm{K}_2$	$\bar{K}_{znm} = \sum_{ij} \bar{v}_{1,ij}\delta_{in}\delta_{jm} = \bar{v}_{1,nm}$		(101)
	$\frac{\partial v_{1ij}}{\partial K_{znm}} = \delta_{in}\delta_{jm}$		(102)
$v_2 = v_1^{-1}$	$\bar{v}_{1,ij} = \sum_{nm} \bar{v}_{2,nm}\frac{\partial v_{2,nm}}{\partial v_{1,ij}} = -\sum_n v_{1,ni}^{-1}\sum_m \bar{v}_{2,nm}v_{1,jm}^{-1}$	$O(N^3)$	(103)
	$\frac{\partial v_{2,nm}}{\partial v_{1,ij}} = -v_{1,ni}^{-1}v_{1,jm}^{-1}$	See forward mode.	(104)
$v_3 = \mathbf{y}^\intercal v_2 \mathbf{y}$	$\bar{v}_{2,nm} = \bar{v}_3 y_m y_n$		(105)
	$\bar{v}_2 = \bar{v}_3\frac{\partial v_3}{\partial v_2}$		(106)
	$\frac{\partial v_3}{\partial v_{2,nm}} = y_n y_m$		(107)
$\mathrm{out} = v_3$	$\bar{v}_3 = 1$		(108)

The computational complexity this way round is cheaper that for the forward mode, since we have many variables that we are differentiating with respect to. The cost for each reverse step is the same as the forward step.

19 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

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

Notice that by Bayes' rule, we have the posterior probability of cluster assignment as

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

This means

$$\nabla_{\mu_k} L(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} p(k|x_n) \frac{x_n - \mu_k}{\sigma^2}.$$

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

$$\mu_k = \frac{\sum_{n=1}^{N} p(k|x_n) x_n}{\sum_{n=1}^{N} p(k|x_n)}.$$

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 $\boldsymbol{\Phi}$.

20 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:

$$R(\mathbf{A}, \mathbf{x}) = \frac{\mathbf{x}^\top\mathbf{Q}\Lambda\mathbf{Q}^\top\mathbf{x}}{\mathbf{x}^\top\mathbf{Q}\mathbf{Q}^\top\mathbf{x}} = \frac{\mathbf{z}^\top\Lambda\mathbf{z}}{\mathbf{z}^\top\mathbf{z}}. \tag{109}$$

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

$$\mathbf{z}^\top\Lambda\mathbf{z} = \sum_{i=1}^{d} \lambda_i z_i^2. \tag{110}$$

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

$$R(\mathbf{A}, \mathbf{x}) = \sum_{i=1}^{d} \frac{z_i^2}{||\mathbf{z}||_2^2}\lambda_i, \quad \text{with} \quad \sum_{i=1}^{d} \frac{z_i^2}{||\mathbf{z}||_2^2} = 1. \tag{111}$$

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

$$\begin{aligned} & \lambda_{min}(\mathbf{A}) \leq R(\mathbf{A}, \mathbf{x}) \leq \lambda_{max}(\mathbf{A}) \\ \Rightarrow \quad & \lambda_{min}(\mathbf{A})||\mathbf{x}||_2^2 \leq \mathbf{x}^\top\mathbf{A}\mathbf{x} \leq \lambda_{max}(\mathbf{A})||\mathbf{x}||_2^2, \end{aligned} \tag{112}$$

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

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \gamma\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t) + \alpha\Delta\boldsymbol{\theta}_t = \boldsymbol{\theta}_t - \frac{\gamma}{\sigma^2}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\theta}_t - \mathbf{y}) + \alpha\Delta\boldsymbol{\theta}_t \\ \Delta\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t = \alpha\Delta\boldsymbol{\theta}_t - \frac{\gamma}{\sigma^2}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\theta}_t - \mathbf{y}). \end{aligned} \tag{113}$$

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

$$\begin{bmatrix} \boldsymbol{\theta}_{t+1} \\ \Delta\boldsymbol{\theta}_{t+1} \end{bmatrix} = \begin{bmatrix} \mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X} & \alpha\mathbf{I} \\ -\frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X} & \alpha\mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_t \\ \Delta\boldsymbol{\theta}_t \end{bmatrix} + \begin{bmatrix} \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{y} \\ \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{y} \end{bmatrix}. \tag{114}$$

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

$$\begin{bmatrix} \boldsymbol{\theta}_t \\ \Delta\boldsymbol{\theta}_t \end{bmatrix} = \begin{bmatrix} \mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X} & \alpha\mathbf{I} \\ -\frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X} & \alpha\mathbf{I} \end{bmatrix}^t \begin{bmatrix} \boldsymbol{\theta}_0 - \boldsymbol{\theta}^* \\ \Delta\boldsymbol{\theta}_0 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\theta}^* \\ \mathbf{0} \end{bmatrix}, \tag{115}$$

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 $\Delta\boldsymbol{\theta}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1} \to \mathbf{0}$.

21 Answers Lectures 6 & 7: Multivariate Probability

Question 35 Y is distributed as a multivariate Gaussian with mean $\mathbf{A}\boldsymbol{\mu}$ and covariance matrix $\mathbf{A}\Sigma\mathbf{A}^\top$.

To see this, we first define $T(X) = \mathbf{A}X$, so that we have $Y = T(X)$. Using change of variables rule, the PDF of Y needs to satisfy the following:

$$p_Y(\mathbf{y}) = p_X(T^{-1}(\mathbf{y}))|\frac{dT^{-1}(\mathbf{y})}{dy}|.$$

As $T^{-1}(\mathbf{y}) = \mathbf{A}^{-1}\mathbf{y}$, we have $|\frac{dT^{-1}(y)}{dy}| = |\mathbf{A}^{-1}| = |\mathbf{A}|^{-1}$. Plugging in these results:

$$\begin{aligned} p_Y(\mathbf{y}) &= |\mathbf{A}|^{-1}\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp[-\frac{1}{2}(\mathbf{A}^{-1}\mathbf{y} - \boldsymbol{\mu})^\top\Sigma^{-1}(\mathbf{A}^{-1}\mathbf{y} - \boldsymbol{\mu})] \\ &= \frac{1}{\sqrt{(2\pi)^d|\mathbf{A}\Sigma\mathbf{A}^\top|}}\exp[-\frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu})^\top(\mathbf{A}\Sigma\mathbf{A}^\top)^{-1}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu})]. \end{aligned}$$

Question 36 The PDF of $Z = X + Y$ needs to satisfy the following:

$$p_Z(z) = \int_{x+y=z} p_X(x)p_Y(y)dxdy.$$

Expanding the right hand side expression, we have (assuming the mean and variance of X, Y are $\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2$):

$$\begin{aligned}
p_Z(z) &= \int_{x+y=z} p_X(x)p_Y(y)dxdy \\
&= \int_{x\in\mathbb{R}} p_X(x)p_Y(z-x)dx \\
&= \int \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp[-\frac{1}{2\sigma_Y^2}(x-\mu_X)^2] \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp[-\frac{1}{2\sigma_Y^2}(z-x-\mu_Y)^2]dx \\
&= \int \frac{1}{2\pi\sqrt{\sigma_X^2\sigma_Y^2}} \exp[-\frac{1}{2}[(\sigma_X^{-2}+\sigma_Y^{-2})x^2 - 2x(\sigma_X^{-2}\mu_X + \sigma_Y^{-2}(z-\mu_Y)) + \sigma_X^{-2}\mu_X^2 + \sigma_Y^{-2}(z-\mu_Y)^2]]dx.
\end{aligned}$$

Define two new parameters μ, σ satisfying $\sigma^{-2} = \sigma_X^{-2} + \sigma_Y^{-2}$ and $\sigma^{-2}\mu = \sigma_X^{-2}\mu_X + \sigma_Y^{-2}(z-\mu_Y)$, we have

$$\begin{aligned}
p_Z(z) &= \int \frac{1}{2\pi\sqrt{\sigma_X^2\sigma_Y^2}} \exp[-\frac{1}{2}[\sigma^{-2}(x-\mu)^2 + \sigma_X^{-2}\mu_X^2 + \sigma_Y^{-2}(z-\mu_Y)^2 - \sigma^{-2}\mu^2]]dx \\
&= \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2]dx \frac{\sqrt{2\pi\sigma^2}}{2\pi\sqrt{\sigma_X^2\sigma_Y^2}} \exp[-\frac{1}{2}[\sigma_X^{-2}\mu_X^2 + \sigma_Y^{-2}(z-\mu_Y)^2 - \sigma^{-2}\mu^2]] \\
&= \frac{1}{\sqrt{2\pi\sigma_X^2\sigma_Y^2\sigma^{-2}}} \exp[-\frac{1}{2}[\sigma_X^{-2}\mu_X^2 + \sigma_Y^{-2}(z-\mu_Y)^2 - \sigma^2[\sigma_X^{-2}\mu_X + \sigma_Y^{-2}(z-\mu_Y)]^2]].
\end{aligned}$$

Up to this point, we see that the PDF of Z has a Gaussian form (as the terms inside the exponent is a quadratic function of z), so this confirms that Z is distributed as a univariate Gaussian. Continuing the derivation by completing the squares, we can work out the corresponding mean and variance of Z:

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2, \quad \mu_Z = \mu_X + \mu_Y.$$

Question 37 Discrete case: by using $x = T^{-1}(y)$,

$$\begin{aligned}
\mathrm{KL}[p_Y(y)||q_Y(y)] &= \sum_y p_Y(y) \log \frac{p_Y(y)}{q_Y(y)} = \sum_y p_X(T^{-1}(y)) \log \frac{p_X(T^{-1}(y))}{q_X(T^{-1}(y))} \\
&= \sum_x p_X(x) \log \frac{p_X(x)}{q_X(x)} = \mathrm{KL}[p_X(x)||q_X(x)].
\end{aligned}$$

Continuous case: again by using $x = T^{-1}(y)$,

$$\begin{aligned}
\mathrm{KL}[p_Y(y)||q_Y(y)] &= \int p_Y(y) \log \frac{p_Y(y)}{q_Y(y)} dy = \int p_X(T^{-1}(y))|\frac{dT^{-1}(y)}{dy}| \log \frac{p_X(T^{-1}(y))|\frac{dT^{-1}(y)}{dy}|}{q_X(T^{-1}(y))|\frac{dT^{-1}(y)}{dy}|} dy \\
&= \int p_X(T^{-1}(y))|\frac{dx}{dy}| \log \frac{p_X(T^{-1}(y))}{q_X(T^{-1}(y))} dy = \int p_X(x) \log \frac{p_X(x)}{q_X(x)} dx = \mathrm{KL}[p_X(x)||q_X(x)].
\end{aligned}$$

Question 38 As we define $\Lambda = \Sigma^{-1}$, the PDF of the multivariate Gaussian is (with mean $\boldsymbol{\mu}$):

$$\begin{aligned}
p(X = \mathbf{x}) &= \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Lambda (\mathbf{x}-\boldsymbol{\mu})^\top] \\
&= \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp[-\frac{1}{2}\sum_k\sum_l (x_k-\mu_k)\Lambda_{kl}(x_l-\mu_l)] \\
&= \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp[-\frac{1}{2}[f(x_i) + g(x_j) + \sum_{k\neq i,j}\sum_l (x_k-\mu_k)\Lambda_{kl}(x_l-\mu_l)]],
\end{aligned}$$

where $\quad f(x_i) = \Lambda_{ii}(x_i-\mu_i)^2 + (x_i-\mu_i)\sum_{l\neq i.j}\Lambda_{il}(x_l-\mu_l),$

$$g(x_j) = \Lambda_{jj}(x_j-\mu_j)^2 + (x_j-\mu_j)\sum_{l\neq i,j}\Lambda_{jl}(x_l-\mu_l).$$

Now to show $X_i \perp\!\!\!\perp X_j | X_{-ij}$, we need to compute the marginal distribution $p(X_{-ij})$. This can be done using the sum rule:

$$
\begin{aligned}
p(X_{-ij}) &= \int p(X_{-ij}, X_i = x_i, X_j = x_j) dx_i dx_j \\
&= \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp[-\frac{1}{2}[\sum_{k \neq i,j} \sum_{l} (X_k - \mu_k) \Lambda_{kl} (X_l - \mu_l)]] \int \exp[-\frac{1}{2}[f(x_i) + g(x_j)]] dx_i dx_j.
\end{aligned} \tag{116}
$$

Without explicitly computing the integral and instead denoting it as $\int \exp[-\frac{1}{2}[f(x_i) + g(x_j)]] dx_i dx_j = F(X_{-ij})$, we have

$$
p(X_i, X_j | X_{-ij}) = \frac{P(X)}{P(X_{-ij})} = \frac{\exp[-\frac{1}{2}[f(X_i) + g(X_j)]]}{F(X_{-ij})}.
$$

Using the sum rule again, we can show that:

$$
p(X_i | X_{-ij}) = \int p(X_i, X_j = x_j | X_{-ij}) dx_j = \frac{\exp[-\frac{1}{2} f(X_i)]}{F(X_{-ij})} \int \exp[-\frac{1}{2} g(x_j)] dx_j.
$$

$$
p(X_j | X_{-ij}) = \int p(X_i = x_i, X_j | X_{-ij}) dx_j = \frac{\exp[-\frac{1}{2} g(X_j)]}{F(X_{-ij})} \int \exp[-\frac{1}{2} f(x_i)] dx_i.
$$

Lastly, notice that:

$$
F(X_{-ij}) = \int \exp[-\frac{1}{2}[f(x_i) + g(x_j)]] dx_i dx_j = \int \exp[-\frac{1}{2} f(x_i)] dx_i \int \exp[-\frac{1}{2} g(x_j)] dx_j.
$$

Combining all the above equations proves that

$$
p(X_i, X_j | X_{-ij}) = p(X_i | X_{-ij}) p(X_j | X_{-ij}) \quad \Rightarrow \quad X_i \perp\!\!\!\perp X_j | X_{-ij}.
$$

Question 39 As $Y = X^2$ this means Y is completely determined by X hence X and Y are dependent. However, $\mathbb{E}[XY] = \mathbb{E}[X^3] = 0$ as the PDF of a univariate Gaussian is symmetric around the mean (which is 0 in this case). Notice that $\mathbb{E}[X] = 0$, therefore $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, meaning that X and Y are uncorrelated.

Question 41 – MML 6.11: Iterated Expectations Let us consider the random variables X, Y with joint distribution $p(x, y)$.

$$
\begin{aligned}
\mathbb{E}_X[X] &= \mathbb{E}_Y\Big[\mathbb{E}_X[X|Y]\Big] \\
&= \int_{y \in \mathbb{R}} \mathbb{E}_X[X|Y] p(y) dy \\
&= \int_{y \in \mathbb{R}} \int_{x \in \mathbb{R}} x p(x|y) p(y) dx dy \\
&= \int_{x \in \mathbb{R}} x \Big(\int_{y \in \mathbb{R}} p(x, y) dy \Big) dx \\
&= \int_{x \in \mathbb{R}} x p(x) dx = \mathbb{E}_X[X]
\end{aligned}
$$

Notice that we just swapped the integrals and marginalized y over the joint distribution to obtain $p(x)$.

Question 42 – MML 6.13: Probability Integral Transformation Given a continuous random variable X, with *cdf* $F_X(x)$, show that the random variable $Y := F_X(X)$ is uniformly distributed.

$$
\begin{aligned}
F_Y(y) &= P(Y \leq y) \\
&= P(F_X(X) \leq y) \\
&= P(X \leq F_X^{-1}(y)) \\
&= F_X(F_X^{-1}(y)) \\
&= y
\end{aligned}
$$

Thus, we have that $Y \sim Uniform(0, 1)$ because $F_Y(y)$ is the *cdf* of a uniform distribution.

22 Answers Lecture 8: Generalisation, Test Sets, Monte Carlo

Question 44 – Basic Monte Carlo Estimate

a. Let us write the integral as and expectation. Loosely speaking, we assume x and y are uniformly distributed random variables in the interval $[-1, 1]$. From the previous assumption, we know $p(x) = p(y) = \frac{1}{1-(-1)} = \frac{1}{2}$. Consider now the expectation over $f(x, y) = \mathbb{I}(x^2 + y^2 < 1)$.

$$\mathbb{E}_{p(x,y)}[f(x,y)] = \int_{-1}^{1}\int_{-1}^{1} p(x,y)f(x,y)dxdy = \int_{-1}^{1}\int_{-1}^{1} \frac{1}{2}\frac{1}{2}\mathbb{I}(x^2+y^2<1)dxdy = \frac{1}{4}I \tag{117}$$

$$I = 4\mathbb{E}_{p(x,y)}[f(x,y)] \tag{118}$$

where we use $p(x, y) = p(x)p(y)$.

b. We can compute a Monte Carlo estimate of this integral as follows

$$I = 4\mathbb{E}_{p(x,y)}[f(x,y)] \approx \frac{4}{N}\sum_{i=1}^{N} I(x_i^2 + y_i^2 < 1) = \hat{I}_N \tag{119}$$

where $x_i \sim U(-1, 1), y_i \sim U(-1, 1), i = 1, \ldots, N$ are i.i.d samples. Law of large numbers ensures that

$$\lim_{N\to\infty} \hat{I}_N = I \tag{120}$$

c. Try it and see if the relative error scales as $\frac{1}{\sqrt{N}}$.

d. We start just with verifying unbiasedness:

$$\mathbb{E}_{X_1,\ldots,X_N}\left[\left(\hat{I}\right)^2\right] = \mathbb{E}_{X_1,X_2,\ldots,X_N}\left[\left(\frac{1}{N}\sum_{n=1}^{N} f(x_n)\right)^2\right] \tag{121}$$

$$= \mathbb{E}_{X_1,X_2,\ldots,X_N}\left[\left(\frac{1}{N}\sum_{n=1}^{N} f(x_n)\right)\left(\frac{1}{N}\sum_{m=1}^{N} f(x_m)\right)\right] \tag{122}$$

$$= \frac{1}{N^2}\sum_{n}\sum_{m}\mathbb{E}_{X_1,X_2,\ldots,X_N}[f(x_n)f(x_m)] \tag{123}$$

$$= \frac{1}{N^2}\left(\sum_{n\neq m}\mathbb{E}_X[f(X)]^2 + \sum_{n}\mathbb{E}_X\left[f(X)^2\right]\right) \tag{124}$$

We see that this only holds if $\mathbb{E}_X\left[f(X)^2\right] = \mathbb{E}_X[f(X)]^2$, which would imply that $\mathbb{V}_X[f(X)] = 0$. This is not the case for our problem.

An alternative proof comes from Jensen's inequality. Let us compute the bias of the estimate $\hat{I}^2$. Since $(\cdot)^2$ is a convex function, we can use Jensen's inequality to show

$$\mathbb{E}\left[(\hat{I})^2\right] < \left(\mathbb{E}\left[\hat{I}\right]\right)^2 \tag{125}$$

where the equality holds iff $(\cdot)^2$ is linear, or if the variance of the RV is zero. Neither is the case.

23 Answers Lecture 10 & 11: Bayesian Inference

Question 45 – Electrical Communication

a. Write down Bayes' rule to find the posterior.

Consider the likelihood $p(v_1, \ldots, v_N|s)$, prior $p(s)$ and evidence $p(v_1, \ldots, v_N)$. Bayes' rule gives

$$p(s|v_1,\ldots,v_N) = \frac{p(v_1,\ldots,v_N|s)p(s)}{p(v_1,\ldots,v_N)} \tag{126}$$

$$= \frac{\prod_{n=1}^{N} p(v_n|s)p(s)}{p(v_1,\ldots,v_N)} \tag{127}$$

b. Find the density of the posterior.

First, let us write the exponential terms which depend on s. Equivalently, we take the logarithm of the previous expression.

$$\begin{aligned}
\log p(s|v_1,\dots,v_N) &= \log p(v_1,\dots,v_N|s) + \log p(s) + const \\
&= \sum_{n=1}^{N} \log p(v_n|s) + \log p(s) + const \\
&= \sum_{n=1}^{N} \left(\frac{-1}{2\sigma^2}(v_n - s)^2 \right) + \frac{-1}{2}s^2 + const \\
&= \frac{-1}{2\sigma^2} \sum_{n=1}^{N} (v_n^2 - 2v_n s + s^2) + \frac{-1}{2}s^2 \\
&= -\frac{1}{2}s^2 \left(\frac{N}{\sigma^2} + 1 \right) + \frac{s}{\sigma^2} \sum_n v_n + const
\end{aligned}$$

We recognise this log density as that of a Gaussian, because it's quadratic. Assuming $p(s|v_1,\dots,v_N)$ is Gaussian distributed $s|v_1,\dots,v_N \sim \mathcal{N}(m,\lambda^2)$, and we know

$$p(s|v_1,\dots,v_N) \propto p(v_1,\dots,v_N|s)p(s), \tag{128}$$

we can complete the square wrt s to obtain m and λ^2.

We can now use the quadratic term wrt s to obtain the variance

$$-\frac{1}{2}\left(\frac{N}{\sigma^2} + 1 \right) = \frac{-1}{2\lambda^2} \implies \lambda^2 = \frac{1}{1+\frac{N}{\sigma^2}}, \tag{129}$$

and the linear term wrt s to obtain the mean

$$\begin{aligned}
\frac{s}{\sigma^2} \sum_n v_n &= \frac{sm}{\lambda^2} \\
\frac{1}{\sigma^2} \sum_n v_n &= m\left(1 + \frac{N}{\sigma^2}\right) \\
m &= \frac{1}{\sigma^2 + N} \sum_n v_n
\end{aligned}$$

Thus, by completing the square we have the density of the posterior:

$$p(s|v_1,\dots,v_N) = \mathcal{N}\left(\frac{1}{\sigma^2+N}\sum_n v_n, \frac{1}{1+\frac{N}{\sigma^2}}\right). \tag{130}$$

c. Given that we know $v_n|s \sim \mathcal{N}(s,\sigma^2)$, from the i.i.d. assumption we know that the joint probability expressed as a univariate Gaussian has mean s. We can now complete the square wrt s to find the variance. We will see that this shares the same expression as the distribution of $\frac{1}{N}(v_1+v_2+\dots+v_N)$ up to some constant wrt s.

$$\begin{aligned}
\log p(v_1,\dots,v_N|s) &= \log \prod_{n=1}^{N} p(v_n|s) \\
&= \sum_{n=1}^{N} \left(\frac{-1}{2\sigma^2}(v_n - s)^2 \right) + const \\
&= \frac{-1}{2\sigma^2} \sum_{n=1}^{N} (v_n^2 - 2v_n s + s^2) + const \\
&= \frac{-N}{2\sigma^2}s^2 + \frac{1}{2\sigma^2}s \sum_{n=1}^{N} v_n + const
\end{aligned}$$

$$p(v_1, \ldots, v_N|s) = c \cdot \mathcal{N}(\bar{v}; s, \frac{\sigma^2}{N})$$

where we find the variance using the quadratic term and from the linear term observe that the corresponding random variable for the univariate distribution is $\bar{v} = \frac{1}{N}(v_1 + v_2 + \cdots + v_N)$.

d. From Gaussian conditioning, we know that the joint distribution can be expressed as follows

$$p(\bar{v}, s) = \mathcal{N}\left(\left[\begin{array}{c} \bar{v} \\ s \end{array}\right]; \left[\begin{array}{c} a \\ b \end{array}\right], \left(\begin{array}{cc} A & B \\ B & C \end{array}\right)\right), \tag{131}$$

where the conditional $p(\bar{v}|s)$ is

$$p(\bar{v}|s) = \mathcal{N}\left(\bar{v}; \frac{B}{C}(s - b) + a, A - \frac{B^2}{C}\right). \tag{132}$$

We can use the mean from $p(\bar{v}|s)$ to get $B = C = 1$ and $b = a = 0$, and the variance from $p(\bar{v}|s)$ to get $A = 1 + \frac{\sigma^2}{N}$. Thus, the joint probability $p(\bar{v}, s)$ is

$$p(\bar{v}, s) = \mathcal{N}\left(\left[\begin{array}{c} \bar{v} \\ s \end{array}\right]; \left[\begin{array}{c} 0 \\ 0 \end{array}\right], \left(\begin{array}{cc} 1 + \frac{\sigma^2}{N} & 1 \\ 1 & 1 \end{array}\right)\right). \tag{133}$$

e. From conditioning, we can obtain the posterior distribution $p(s|\bar{v})$.

$$p(s|\bar{v}) = \mathcal{N}\left(s; \frac{\bar{v}}{1 + \frac{\sigma^2}{N}}, 1 - \frac{1}{1 + \frac{\sigma^2}{N}}\right) = \mathcal{N}\left(s; \frac{\bar{v}}{1 + \frac{\sigma^2}{N}}, \frac{1}{+\frac{N}{\sigma^2}}\right), \tag{134}$$

which is equivalent to the posterior obtained by completing the square.

Question 46 – Electrical Communication Errors Let's work backwards from what we are asked to compute: $\mathbb{P}(\hat{S} = S)$. Remember that $\hat{S}$ is actually a function of v, and then let's try to find it in terms of probabilities we can find:

$$\begin{aligned}
\mathbb{P}(\hat{S}(v) = S) &= \mathbb{P}(\hat{S}(v) = 1|S = 1)\mathbb{P}(S = 1) + \mathbb{P}(\hat{S}(v) = 1|S = 0)\mathbb{P}(S = 0) && (135)\\
&= \mathbb{P}(\hat{S}(v) = 1|S = 1)\pi + \mathbb{P}(\hat{S}(v) = 1|S = 0)(1 - \pi)\,. && (136)
\end{aligned}$$

Now we are left to find $\mathbb{P}(\hat{S}(v) = s|S = s)$. To find this probability, we will need to understand the function $\hat{S}(v)$ a bit better.

We know that $\hat{S}(v) = \text{argmax}_s\, p(s|v)$, and so it will have a binary output. Let's understand the region where $\hat{S}(v) = 1$. We start by noting $\hat{S}(v) = \text{argmax}_s\, p(s|v) = \text{argmax}_s\, p(v|s)p(s)$, so we will have $\hat{S}(v) = 1$ if

$$\begin{aligned}
\log p(v|S = 1)p(S = 1) &> \log p(v|S = 0)p(S = 0) && (137)\\
-\frac{1}{2\sigma^2}(v - 1)^2 + \log p &> -\frac{1}{2\sigma^2}v^2 + \log 1 - p && (138)\\
\frac{1}{2\sigma^2}2v &> \log\frac{1 - p}{p} + \frac{1}{2\sigma^2} && (139)\\
v &> \sigma^2 \log\frac{1 - p}{p} + \frac{1}{2} := t\,. && (140)
\end{aligned}$$

This allows us to compute the two probabilities we are after:

$$\begin{aligned}
\mathbb{P}(\hat{S}(v) = 0|S = 0) &= \int_{-\infty}^{t} \pi(v|S = 0)\mathrm{d}v & \mathbb{P}(\hat{S}(v) = 1|S = 1) &= \int_{t}^{\infty} \pi(v|S = 1)\mathrm{d}v && (141)\\
&= \int_{-\infty}^{t} \mathcal{N}(v; 0, \sigma^2)\mathrm{d}v & &= \int_{t}^{\infty} \mathcal{N}(v; 1, \sigma^2)\mathrm{d}v && (142)\\
&= \Phi\left(\frac{t}{\sigma}\right) & &= 1 - \Phi\left(\frac{t - 1}{\sigma}\right) && (143)
\end{aligned}$$

Here, we did assume that the true noise distribution is equal to the noise distribution we assume in the model, i.e. $\pi(v|s) = p(v|s)$. This may not be true of all models!

So, overall, we get:

$$\mathbb{P}(\hat{S} = S) = \pi\left(1 - \Phi\left(\frac{t-1}{\sigma}\right)\right) + (1-\pi)\Phi\left(\frac{t}{\sigma}\right) \tag{144}$$

Although the question is done, let's get a bit more insight by investigating what happens when we vary p, for a true frequency $\mathbb{P}(S = 1) = \pi = 0.6$. In section 23, we plot $\mathbb{P}(\hat{S} = S)$ for varying p. We see that the probability of getting it right is maximised *when our model prior matches the frequency of reality.*

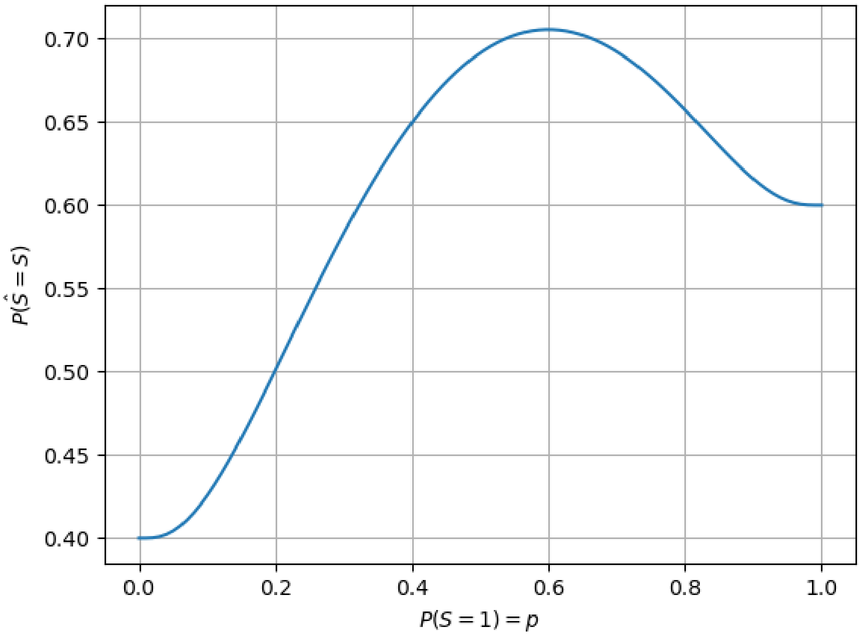

Figure 3: $\mathbb{P}(\hat{S} = S)$ for varying p.

24 Answers Lecture 12: Bias-Variance Tradeoff

Question 50 We have the covariance of the ridge regression estimator as

$$\mathbf{V}(\lambda) = \mathbb{E}_{\mathbf{X}_{\text{train}}}[\sigma^2(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\Phi^\top\Phi(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}], \quad \mathbf{V}(0) = \mathbb{E}_{\mathbf{X}_{\text{train}}}[\sigma^2(\Phi^\top\Phi)^{-1}].$$

Note that as we assume $\Phi^\top\Phi$ is invertible, we can rewrite

$$\begin{aligned}\mathbf{V}(0) &= \mathbb{E}_{\mathbf{X}_{\text{train}}}[\sigma^2(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)(\Phi^\top\Phi)^{-1}(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}] \\ &= \mathbb{E}_{\mathbf{X}_{\text{train}}}[\sigma^2(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}(2\sigma^2\lambda\mathbf{I} + \sigma^4\lambda^2(\Phi^\top\Phi)^{-1} + \Phi^\top\Phi)(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}].\end{aligned}$$

So one can also show that:

$$\mathbf{V}(\lambda) - \mathbf{V}(0) = \mathbb{E}_{\mathbf{X}_{\text{train}}}[-\sigma^2(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\underbrace{(2\sigma^2\lambda\mathbf{I} + \sigma^4\lambda^2(\Phi^\top\Phi)^{-1})}_{\text{positive definite}}(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}] \preceq 0,$$

To see why $2\sigma^2\lambda\mathbf{I} + \sigma^4\lambda^2(\Phi^\top\Phi)^{-1}$ is positive definite: $\Psi^\top\Psi$ is positive semi-definite and invertible, so that $(\Psi^\top\Psi)^{-1}$ is positive definite, and scaling it with $\sigma^4\lambda^2 > 0$ still maintain its positive definite property. Then the summation of two positive definite matrix ($2\sigma^2\lambda\mathbf{I}$ is positive definite) is a positive definite matrix, and the expectation of a positive definite matrix (as a random variable) still remains positive definite.

Question 51 We have the bias of the estimate is

$$\mathbf{b}(\lambda) := \mathbb{E}_{\mathcal{D}\sim p_{data}^N}[\boldsymbol{\theta}_R^*(\mathcal{D})] - \boldsymbol{\theta}_0 = -\mathbb{E}_{\mathbf{X}_{\text{train}}}[\sigma^2\lambda(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}]\boldsymbol{\theta}_0,$$

This means:

$$\begin{aligned}\mathbf{b}(\lambda)\mathbf{b}(\lambda)^\top &= \mathbb{E}_{\mathbf{X}_{\text{train}}}[\sigma^4\lambda^2(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}]\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\mathbb{E}_{\mathbf{X}_{\text{train}}}[(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}] \\ &\preceq \mathbb{E}_{\mathbf{X}_{\text{train}}}[\sigma^4\lambda^2(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}].\end{aligned}$$

Using the result we have for $\mathbf{V}(\lambda) - \mathbf{V}(0)$ in Question 50, we have:

$$\mathbf{b}(\lambda)\mathbf{b}(\lambda)^\top + \mathbf{V}(\lambda) - \mathbf{V}(0) \preceq -\mathbb{E}_{\mathbf{X}_{\text{train}}}[\sigma^2\lambda(\Phi^\top\Phi + \sigma^2\lambda\mathbf{I})^{-1}\underbrace{(\sigma^2[2\mathbf{I} - \lambda\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top + \sigma^2\lambda(\Phi^\top\Phi)^{-1}])}_{:=\mathbf{E}}(\Phi^\top\Phi + \sigma^2\lambda\mathbf{I})^{-1}].$$

Furthermore, one can show that

$$\mathbf{E} \text{ is positive semi-definite} \quad \Rightarrow \quad \mathbf{b}(\lambda)\mathbf{b}(\lambda)^\top + \mathbf{V}(\lambda) - \mathbf{V}(0) \preceq \mathbf{0},$$

which can be achieved by e.g. setting $0 \leq \lambda \leq \frac{2}{||\boldsymbol{\theta}_0||_2^2}$. To see this, a close inspection on $\mathbf{E}$ shows that if we make $2\mathbf{I} - \lambda\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top$ positive semi-definite then $\mathbf{E}$ will also be positive semi-definite. As $\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top$ is a rank-1 matrix, the only non-zero eigenvalue of $\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top$ is $||\boldsymbol{\theta}_0||_2^2$. Using the discussed indications of eigen-decomposition, we can show that $2\mathbf{I} - \lambda\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top$ is positive semi-definite when all of its eigenvalues are non-negative, which can be achieved by $0 \leq \lambda \leq \frac{2}{||\boldsymbol{\theta}_0||_2^2}$.

Question 52 (a) As X is assumed to be an unbiased estimator of x_0, we have $\mathbb{E}_X[X] = x_0$. Now for the estimator $X + Y - \mathbb{E}_Y[Y]$:

$$\mathbb{E}_{X,Y}[X + Y - \mathbb{E}_Y[Y]] = \mathbb{E}_{X,Y}[X] + \mathbb{E}_{X,Y}[Y] - \mathbb{E}_Y[Y] = \mathbb{E}_X[X] = x_0,$$

therefore this estimator is also unbiased.

(b) We compute the variance of this estimator:

$$\begin{aligned}\mathbb{V}_{X,Y}[X + Y - \mathbb{E}_Y[Y]] &= \mathbb{E}_{X,Y}[(X + Y - \mathbb{E}_Y[Y] - x_0)^2] \\ &= \mathbb{E}_{X,Y}[((X - \mathbb{E}_X[X]) + (Y - \mathbb{E}_Y[Y]))^2] \\ &= \mathbb{E}_{X,Y}[(X - \mathbb{E}_X[X])^2 + (Y - \mathbb{E}_Y[Y])^2 + 2(X - \mathbb{E}_X[X])(Y - \mathbb{E}_Y[Y])] \\ &= \mathbb{V}_X[X] + \mathbb{V}_Y[Y] + 2Cov_{X,Y}[X, Y].\end{aligned}$$

This immediately implies that $\mathbb{V}_{X,Y}[X + Y - \mathbb{E}_Y[Y]] \leq \mathbb{V}_X[X]$ if $\mathbb{V}_Y[Y] + 2Cov_{X,Y}[X, Y] < 0$.

(c) Following (b) and assume $Y = cZ$:

$$\mathbb{V}_{X,Z}[X + cZ - \mathbb{E}_Z[cZ]] = \mathbb{V}_X[X] + c^2\mathbb{V}_Z[Z] + 2cCov_{X,Z}[X, Z].$$

To maximise variance reduction, we need to compute the derivative of $c^2\mathbb{V}_Z[Z] + 2cCov_{X,Z}[X, Z]$ w.r.t. c and make it equal to zero. This leads to:

$$\frac{\partial}{\partial c}c^2\mathbb{V}_Z[Z] + 2cCov_{X,Z}[X, Z] = 2c\mathbb{V}_Z[Z] + 2Cov_{X,Z}[X, Z] = 0 \quad \Rightarrow c = -\frac{Cov_{X,Z}[X, Z]}{\mathbb{V}_Z[Z]}.$$

25 Answers Lecture 13: PCA

Question 53 (a) Let us rewrite the objective $L(\mathbf{A},\mathbf{B})$:

$$
\begin{aligned}
L(\mathbf{A},\mathbf{B}) :=& \frac{1}{N}\sum_{n=1}^{N}||\mathbf{x}_n - \mathbf{A}\mathbf{B}\mathbf{x}_n||_2^2 = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \mathbf{A}\mathbf{B}\mathbf{x}_n)^\top(\mathbf{x}_n - \mathbf{A}\mathbf{B}\mathbf{x}_n) \\
=& \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n^\top\mathbf{x}_n - 2\mathbf{x}_n^\top\mathbf{A}\mathbf{B}\mathbf{x}_n + \mathbf{x}_n^\top\mathbf{B}^\top\mathbf{A}^\top\mathbf{A}\mathbf{B}\mathbf{x}_n) \\
=& \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n^\top\mathbf{x}_n - 2\mathrm{tr}(\mathbf{x}_n^\top\mathbf{A}\mathbf{B}\mathbf{x}_n) + \mathrm{tr}(\mathbf{x}_n^\top\mathbf{B}^\top\mathbf{A}^\top\mathbf{A}\mathbf{B}\mathbf{x}_n)) \\
=& \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n^\top\mathbf{x}_n - 2\mathrm{tr}(\mathbf{A}\mathbf{B}\mathbf{x}_n\mathbf{x}_n^\top) + \mathrm{tr}(\mathbf{B}^\top\mathbf{A}^\top\mathbf{A}\mathbf{B}\mathbf{x}_n x_n^\top)) \\
=& \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n^\top\mathbf{x}_n - 2\mathrm{tr}(\mathbf{A}\mathbf{B}\mathbf{x}_n\mathbf{x}_n^\top) + \mathrm{tr}(\mathbf{B}^\top\mathbf{A}^\top\mathbf{A}\mathbf{B}\mathbf{x}_n x_n^\top)) \\
=& \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n^\top\mathbf{x}_n - 2\mathrm{tr}(\mathbf{A}\mathbf{B}\underbrace{\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\top}_{:=\mathbf{S}}) + \mathrm{tr}(\mathbf{B}^\top\mathbf{A}^\top\mathbf{A}\mathbf{B}\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n x_n^\top) \\
=& \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n^\top\mathbf{x}_n - 2\mathrm{tr}(\mathbf{A}\mathbf{B}\mathbf{S}) + \mathrm{tr}(\mathbf{B}^\top\mathbf{A}^\top\mathbf{A}\mathbf{B}\mathbf{S})
\end{aligned}
$$

Now we derive the derivative of $L(\mathbf{A},\mathbf{B})$ w.r.t. $\mathbf{A}$, and notice that $\mathrm{tr}(\mathbf{B}^\top\mathbf{A}^\top\mathbf{A}\mathbf{B}\mathbf{S}) = \mathrm{tr}(\mathbf{A}^\top\mathbf{A}\mathbf{B}\mathbf{S}\mathbf{B}^\top)$:

$$\frac{\partial}{\partial\mathbf{A}}L = 2\mathbf{A}\mathbf{B}\mathbf{S}\mathbf{B}^\top - 2\mathbf{S}\mathbf{B}^\top.$$

Similarly we derive the derivative of $L(\mathbf{A},\mathbf{B})$ w.r.t. $\mathbf{B}$:

$$\frac{\partial}{\partial\mathbf{B}}L = 2\mathbf{A}^\top\mathbf{A}\mathbf{B}\mathbf{S} - 2\mathbf{A}^\top\mathbf{S}.$$

(b) As we assume $rank(\mathbf{A}) = M$, $\mathbf{A}^\top\mathbf{A}$ is invertible. As $\mathbf{S}$ is also assumed to be invertible, then the optimal solution of $\mathbf{B}^*$ satisfies:

$$\mathbf{A}^\top\mathbf{A}\mathbf{B}^*\mathbf{S} = \mathbf{A}^\top\mathbf{S} \quad\Rightarrow\quad \mathbf{A}^\top\mathbf{A}\mathbf{B}^* = \mathbf{A}^\top \quad\Rightarrow\quad \mathbf{B}^* = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top.$$

(c) We first consider, when $\mathbf{B}$ is a given rank-M matrix, the optimal solution of $\mathbf{A}^*$ satisfies:

$$\mathbf{A}^*\mathbf{B}\mathbf{S}\mathbf{B}^\top = \mathbf{S}\mathbf{B}^\top \quad\Rightarrow\quad \mathbf{A}^* = \mathbf{S}\mathbf{B}^\top(\mathbf{B}\mathbf{S}\mathbf{B}^\top)^{-1}.$$

Combining (b), this means the optimal solution $\mathbf{A}^*, \mathbf{B}^*$ satisfies:

$$\mathbf{A}^* = \mathbf{S}(\mathbf{B}^*)^\top(\mathbf{B}^*\mathbf{S}(\mathbf{B}^*)^\top)^{-1}, \quad \mathbf{B}^* = ((\mathbf{A}^*)^\top\mathbf{A}^*)^{-1}(\mathbf{A}^*)^\top.$$

Note that $\mathbf{B}\mathbf{S}\mathbf{B}^\top$ is invertible because both $\mathbf{B}$ and $\mathbf{S}$ has full row rank. Now writing $\mathbf{S} = \mathbf{Q}\Lambda\mathbf{Q}^\top$, we verify in below that $\mathbf{A}^* = \mathbf{Q}_{1:M}, \mathbf{B}^* = \mathbf{Q}_{1:M}^\top$ is a fixed point of the objective $L(\mathbf{A},\mathbf{B})$.

$$
\begin{aligned}
\mathbf{S}(\mathbf{B}^*)^\top(\mathbf{B}^*\mathbf{S}(\mathbf{B}^*)^\top)^{-1} &= \mathbf{Q}\Lambda\mathbf{Q}^\top\mathbf{Q}_{1:M}(\mathbf{Q}_{1:M}^\top\mathbf{Q}\Lambda\mathbf{Q}^\top\mathbf{Q}_{1:M})^{-1} \\
&= \mathbf{Q}\Lambda\begin{bmatrix}\mathbf{I}_M & 0\\ 0 & 0\end{bmatrix}\left(\begin{bmatrix}\mathbf{I}_M & 0\\ 0 & 0\end{bmatrix}\Lambda\begin{bmatrix}\mathbf{I}_M & 0\\ 0 & 0\end{bmatrix}\right)^{-1} \\
&= \mathbf{Q}\Lambda\begin{bmatrix}\Lambda_{1:M}^{-1} & 0\\ 0 & 0\end{bmatrix} = \mathbf{Q}\begin{bmatrix}\mathbf{I}_M & 0\\ 0 & 0\end{bmatrix} = \mathbf{Q}_{1:M} = \mathbf{A}^*,
\end{aligned}
$$

$$((\mathbf{A}^*)^\top\mathbf{A}^*)^{-1}(\mathbf{A}^*)^\top = (\mathbf{Q}_{1:M}^\top\mathbf{Q}_{1:M})^{-1}\mathbf{Q}_{1:M}^\top = \mathbf{Q}_{1:M}^\top = \mathbf{B}^*.$$

In general a fixed point of the objective satisfies:

$$\mathbf{A}^* = \mathbf{Q}_{1:M}\mathbf{C}^{-1}, \mathbf{B}^* = \mathbf{C}\mathbf{Q}_{1:M}^\top, \quad \forall \text{ invertible matrix } \mathbf{C} \in \mathbb{R}^{M\times M}.$$

Question 54 Let us write an SVD of $\mathbf{X}$ as $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top$ with $\mathbf{U} \in \mathbb{R}^{N\times N}, \Sigma \in \mathbb{R}^{N\times D}$ and $\mathbf{V} \in \mathbb{R}^{D\times D}$. Note that the covariance on $\mathcal{D}$ can be computed as $\mathbf{S} = \mathbf{X}^\top\mathbf{X}$. Plugging in the SVD of $\mathbf{X}$:

$$\mathbf{S} = \mathbf{X}^\top\mathbf{X} = \mathbf{V}\Sigma^\top\mathbf{U}^\top\mathbf{U}\Sigma\mathbf{V}^\top = \mathbf{V}\Sigma^\top\Sigma\mathbf{V}^\top.$$

Note that in an SVD, Σ is a rectangular diagonal matrix, i.e., only the leading diagonal terms have non-zero values. This also means $\Sigma^\top\Sigma \in \mathbb{R}^{D\times D}$ is a diagonal matrix with non-negative diagonal values. Therefore $\mathbf{V}\Sigma^\top\Sigma\mathbf{V}^\top$ is an eigendecomposition of $\mathbf{S}$, therefore by sorting the diagonal values in $\Sigma^\top\Sigma$ in descending order, we can retrieve the corresponding columns in $\mathbf{V}$ as the principal components.

26 Answers Lecture 14: Probabilistic PCA

Question 55 (a) As shown in the lecture, the optimal $\mathbf{W}^*$ with its SVD $\mathbf{W}^* = \mathbf{U}\Sigma\mathbf{V}^\top$ satisfies

$$(\mathbf{S}\boldsymbol{u}_1, ..., \mathbf{S}\boldsymbol{u}_M, \mathbf{0}, ..., \mathbf{0}) = ((\sigma_1^2 + \sigma^2)\boldsymbol{u}_1, ..., (\sigma_M^2 + \sigma^2)\boldsymbol{u}_M, \mathbf{0}, ..., \mathbf{0}), \quad \sigma_m = \Sigma_{mm}$$

Plugging-in $\mathbf{S} = \mathbf{Q}\Lambda\mathbf{Q}^\top$, the above equation means $\boldsymbol{u}_m = \boldsymbol{q}_{i_m}, 1 \leq i_m \leq D$ for $i = 1, ..., M$. Also due to the orthogonality constraint of $\mathbf{U}$ column vectors (by definition of SVD), we know that $\{i_m\}_{m=1}^M$ contains distinct indices. This leads to:

$$\mathbf{S}\boldsymbol{u}_m = \lambda_{i_m}\boldsymbol{q}_{i_m} = (\sigma_m^2 + \sigma^2)\boldsymbol{q}_{i_m}.$$

which proves statement (a).

(b) We first compute $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$ for $\mathbf{W}^*$ using SVD of $\mathbf{W}$:

$$\mathbf{C} = \mathbf{U}\begin{bmatrix} \sigma_1 & 0 & \dots \\ 0 & \ddots & \\ \vdots & & \sigma_M \\ & & 0 \\ & & \vdots \\ & & 0 \end{bmatrix}\mathbf{V}^\top\mathbf{V}\begin{bmatrix} \sigma_1 & 0 & \dots \\ 0 & \ddots & \\ \vdots & & \sigma_M \\ & & 0 \\ & & \vdots \\ & & 0 \end{bmatrix}^\top\mathbf{U}^\top + \sigma^2\mathbf{I}$$

$$= \mathbf{U}\begin{bmatrix} \sigma_1^2 + \sigma^2 & 0 & \dots & & & \\ 0 & \ddots & & & & \\ \vdots & & \sigma_M^2 + \sigma^2 & & & \\ & & & \sigma^2 & & \\ & & & & \ddots & \\ & & & & & \sigma^2 \end{bmatrix}\mathbf{U}^\top$$

Plugging-in the optimal $\boldsymbol{\mu}^*$ and a fixed point of $\mathbf{W}^*$ from (a):

$$\mathbf{C} = (\boldsymbol{q}_{i_1}, ..., \boldsymbol{q}_{i_M}, \boldsymbol{u}_{M+1}, ..., \boldsymbol{u}_D)\begin{bmatrix} \lambda_{i_1} & 0 & \dots & & & \\ 0 & \ddots & & & & \\ \vdots & & \lambda_{i_M} & & & \\ & & & \sigma^2 & & \\ & & & & \ddots & \\ & & & & & \sigma^2 \end{bmatrix}(\boldsymbol{q}_{i_1}, ..., \boldsymbol{q}_{i_M}, \boldsymbol{u}_{M+1}, ..., \boldsymbol{u}_D)^\top,$$

which satisfies $\log|C| = (D-M)\log\sigma^2 + \sum_{m=1}^M \log\lambda_{i_m}$. Notice that the above equation returns the same matrix for $\mathbf{C}$ no matter how we choose $\boldsymbol{u}_{M+1}, ..., \boldsymbol{u}_D$ as long as $\mathbf{U}$ contains an orthonormal basis (by definition of SVD). This means for $m \geq M+1$ we can simply choose $\boldsymbol{u}_m = \boldsymbol{q}_j$ for some $j \notin \{i_1, i_2, ..., i_M\}$. Therefore with such choices of $\boldsymbol{u}_{M+1}, ..., \boldsymbol{u}_D$, there exists a permutation matrix $\mathbf{P}$ such that $\mathbf{U} = \mathbf{QP}$ which gives the corresponding permutation of the $\mathbf{Q}$ basis vectors. Then the corresponding MLE objective is:

$$
\begin{aligned}
&\log p_{\boldsymbol{\theta}^*}(\mathbf{x}) = \log \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}^*, \mathbf{C}), \quad \mathbf{C} = \mathbf{QP}(\Sigma\Sigma^\top + \sigma^2\mathbf{I})\mathbf{P}^\top\mathbf{Q}^\top \\
\Rightarrow \quad &\mathcal{L}(\boldsymbol{\theta}^*) = \frac{1}{N}\sum_{n=1}^{N} \log \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}^*, \mathbf{C}) \\
&= -\frac{D}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}^*)^\top \mathbf{C}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}^*) \\
&= -\frac{D}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}\mathrm{tr}(\mathbf{C}^{-1}\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}^*)(\mathbf{x}_n - \boldsymbol{\mu}^*)^\top) \\
&= -\frac{D}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}\mathrm{tr}(\mathbf{C}^{-1}\mathbf{S}) \\
&= -\frac{D}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}\mathrm{tr}(\mathbf{QP}(\Sigma\Sigma^\top + \sigma^2\mathbf{I})^{-1}\mathbf{P}^\top\mathbf{Q}^\top\mathbf{Q}\Lambda\mathbf{Q}^\top) \\
&= -\frac{D}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}\mathrm{tr}(\mathbf{P}(\Sigma\Sigma^\top + \sigma^2\mathbf{I})^{-1}\mathbf{P}^\top\Lambda\mathbf{Q}^\top\mathbf{Q}) \\
&= -\frac{D}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}\mathrm{tr}((\Sigma\Sigma^\top + \sigma^2\mathbf{I})^{-1}\mathbf{P}^\top\Lambda\mathbf{P})
\end{aligned}
$$

We have used permutation invariance of matrix trace. Under the condition in question (a) and the property that $\mathbf{P}^\top\Lambda\mathbf{P}$ permutes the diagonals of Λ using the permutation defined by $\mathbf{P}$, we can show that the top-left $M \times M$ blocks of $(\mathbf{P}^{-1}\Lambda\mathbf{P})$ and $(\Sigma\Sigma^\top + \sigma^2\mathbf{I})$ are equal. This means we only need to consider the last $M+1$ to D diagonal terms in $(\Sigma\Sigma^\top + \sigma^2\mathbf{I})^{-1}\mathbf{P}^{-1}\Lambda\mathbf{P}$ for the trace term, which reads:

$$
\mathrm{tr}((\Sigma\Sigma^\top + \sigma^2\mathbf{I})^{-1}\mathbf{P}^{-1}\Lambda\mathbf{P}) = M + \sigma^{-2}\sum_{j \notin \{i_1,\ldots,i_M\}} \lambda_j.
$$

Also recall that $\log|C| = (D-M)\log\sigma^2 + \sum_{m=1}^{M}\log\lambda_{i_m}$. Combining both results, this means we would like to find an permutation mapping $\rho(\cdot)$ to minimise

$$
\begin{aligned}
\log|\mathbf{C}| + \mathrm{tr}(\mathbf{C}^{-1}\mathbf{S}) &= M + \sigma^{-2}\sum_{j \notin \{i_1,\ldots,i_M\}} \lambda_j + \sum_{m=1}^{M}\log\lambda_{i_m} + (D-M)\log\sigma^2 \\
&= M + \sum_{j \notin \{i_1,\ldots,i_M\}} \frac{\lambda_j}{\sigma^2} + \sum_{m=1}^{M}\log\frac{\lambda_{i_m}}{\sigma^2} + D\log\sigma^2.
\end{aligned}
$$

Recall the assumption of descending eigenvalues $\lambda_1 \geq ... \geq \lambda_D \geq 0$. Since we can show that $x > \log x$ (natural logarithm) for all $x > 0$, then to minimise the above expression, we want to use larger eigenvalues for the $\log\frac{\lambda_{i_m}}{\sigma^2}$ terms and smaller eigenvalues for the $\frac{\lambda_j}{\sigma^2}$ terms. This means the global maximum solutions satisfy $i_m = m$ for $m = 1, ..., M$.