Machine Learning Exercise Sheet 1

Math Refresher

The machine learning lecture relies heavily on your knowledge of undergraduate mathematics, especially linear algebra and probability theory. You should think of this exercise sheet as a test to see if you meet the prerequisites for taking this course. If you struggle with a large fraction of the exercises you should reconsider taking this lecture at this point and instead first prepare by taking a course that reinforces your mathematical foundations (e.g. "Basic Mathematical Tools for Imaging and Visualization" (IN2124)).

Homework

Reading

We strongly recommend that you review the following documents to refresh your knowledge. You should already be familiar with most of their content from your previous studies.

- Linear algebra `http://cs229.stanford.edu/section/cs229-linalg.pdf` (except sections 4.4, 4.5, 4.6), and `http://ee263.stanford.edu/notes/matrix_crimes.pdf` (common linear algebra mistakes)
- Probability theory `http://cs229.stanford.edu/summer2020/cs229-prob.pdf`

Linear Algebra

Notation. We use the following notation in this lecture:

- Scalars are denoted with lowercase letters, e.g. a, x, μ.
- Vectors are denoted with bold lowercase letters, e.g. $\boldsymbol{a}$, $\boldsymbol{x}$, $\boldsymbol{\mu}$.
- Matrices are denoted with bold uppercase letters, e.g. $\boldsymbol{A}$, $\boldsymbol{X}$, $\boldsymbol{\Sigma}$.
- $\mathbb{R}^N$ denotes N-dimensional Euclidean space, i.e. the set of N-dimensional vectors with real-valued entries. For example, $\boldsymbol{x} = (2, \sqrt{2}, 6.5, -7)^T$ is an element of $\mathbb{R}^4$, which we denote as $\boldsymbol{x} \in \mathbb{R}^4$.
- $\mathbb{R}^{M \times N}$ is the set of matrices with M rows and N columns. For example, the matrix $\boldsymbol{A} = \begin{pmatrix} 2 & 3 & 1 \\ 1 & 4 & 5 \end{pmatrix}$ is an element of $\mathbb{R}^{2\times 3}$, which we denote as $\boldsymbol{A} \in \mathbb{R}^{2\times 3}$.
- A function $f : \mathcal{X} \to \mathcal{Y}$ maps elements of the set $\mathcal{X}$ into the set $\mathcal{Y}$. An example would be a function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ defined as $f(x, y) = 2x^2 + xy - 4$.

Problem 1: Let $\boldsymbol{x} \in \mathbb{R}^M$, $\boldsymbol{y} \in \mathbb{R}^N$ and $\boldsymbol{Z} \in \mathbb{R}^{P\times Q}$. The function $f : \mathbb{R}^M \times \mathbb{R}^N \times \mathbb{R}^{P\times Q} \to \mathbb{R}$ is defined as

$$f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{Z}) = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{y} + \boldsymbol{B}\boldsymbol{x} - \boldsymbol{y}^T \boldsymbol{C} \boldsymbol{Z} \boldsymbol{D} - \boldsymbol{y}^T \boldsymbol{E}^T \boldsymbol{y} + \boldsymbol{F}.$$

What should be the dimensions (shapes) of the matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{E}, \boldsymbol{F}$ for the expression above to be a valid mathematical expression?

$$\boldsymbol{A} \in \mathbb{R}^{M \times N}, \boldsymbol{B} \in \mathbb{R}^{1 \times M}, \boldsymbol{C} \in \mathbb{R}^{N \times P}, \boldsymbol{D} \in \mathbb{R}^{Q \times 1}, \boldsymbol{E} \in \mathbb{R}^{N \times N}, \boldsymbol{F} \in \mathbb{R}^{1 \times 1}$$

Problem 2: Let $\boldsymbol{x} \in \mathbb{R}^N, \boldsymbol{M} \in \mathbb{R}^{N \times N}$. Express the function $f(\boldsymbol{x}) = \sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j M_{ij}$ using **only** matrix-vector multiplications.

$$f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{M} \boldsymbol{x}$$

Problem 3: Let $\boldsymbol{A} \in \mathbb{R}^{M \times N}, \boldsymbol{x} \in \mathbb{R}^N$ and $\boldsymbol{b} \in \mathbb{R}^M$. We are interested in solving the following system of linear equations for $\boldsymbol{x}$

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{1}$$

a) Under what conditions does the system of linear equations have a **unique** solution $\boldsymbol{x}$ for **any** choice of $\boldsymbol{b}$?

$M < N$ or $\text{rank}(\boldsymbol{A}) < N$? No, the solution $\boldsymbol{x}$ would not always be unique.
$M > N$ or $\text{rank}(\boldsymbol{A}) < M$? No, a solution $\boldsymbol{x}$ would not exist for every $\boldsymbol{b} \in \mathbb{R}^M$.
Hence, $M = N$ and $\boldsymbol{A}$ has full rank and is therefore invertible.

b) Assume that $M = N = 5$ and that $\boldsymbol{A}$ has the following eigenvalues: $\{-5, 0, 1, 1, 3\}$. Does Equation 1 have a unique solution $\boldsymbol{x}$ for any choice of $\boldsymbol{b}$? Justify your answer.

No, because $\boldsymbol{A}$ has an eigenvalue 0 and therefore does not have full rank.

Problem 4: Let $\boldsymbol{A} \in \mathbb{R}^{N \times N}$. Assume that there exists a matrix $\boldsymbol{B} \in \mathbb{R}^{N \times N}$ such that $\boldsymbol{B}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{B} = \boldsymbol{I}$. What can you say about the eigenvalues of $\boldsymbol{A}$? Justify your answer.

By definition, $\boldsymbol{B}$ is the inverse of $\boldsymbol{A}$. Therefore, $\boldsymbol{A}$ is invertible, i.e. the determinant of $\boldsymbol{A}$ is not equal to zero, i.e. none of the eigenvalues of $\boldsymbol{A}$ are equal to zero.

Problem 5: A symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ is positive semi-definite (PSD) if and only if for any $\boldsymbol{x} \in \mathbb{R}^N$ it holds that $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \geq 0$. Prove that a symmetric matrix $\boldsymbol{A}$ is PSD **if and only if** it has no negative eigenvalues.

1. No negative eigenvalues ⇒ PSD:

Since $\boldsymbol{A}$ is symmetric we can choose orthonormal eigenvectors and then express any vector in $\mathbb{R}^N$ as $\boldsymbol{x} = \sum_{i=1}^{N} w_i \boldsymbol{v}_i$, with the eigenvectors $\boldsymbol{v}_i$ and some coefficients w_i. Hence,

$$\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} = \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \boldsymbol{v}_i^T \boldsymbol{A} \boldsymbol{v}_j = \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \boldsymbol{v}_i^T \lambda_j \boldsymbol{v}_j = \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \lambda_j \delta_{ij} = \sum_{i=1}^{N} w_i^2 \lambda_i \geq 0,$$

since the eigenvalues $\lambda_i \geq 0$. $\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j \end{cases}$ denotes the Kronecker delta.

2. PSD $\Rightarrow$ no negative eigenvalues:

Let $\boldsymbol{v}$ be an eigenvector of $\boldsymbol{A}$ with eigenvalue λ. By the PSD property of $\boldsymbol{A}$ we have

$$0 \leq \boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} = \boldsymbol{v}^T \lambda \boldsymbol{v} = \lambda \|\boldsymbol{v}\|_2^2.$$

Because the Euclidean norm $\|\boldsymbol{v}\|_2^2$ is non-negative we have $\lambda \geq 0$. □

Problem 6: Let $\boldsymbol{A} \in \mathbb{R}^{M \times N}$. Prove that the matrix $\boldsymbol{B} = \boldsymbol{A}^T \boldsymbol{A}$ is positive semi-definite for any choice of $\boldsymbol{A}$.

$$\boldsymbol{x}^T \boldsymbol{B} \boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{x} = (\boldsymbol{A}\boldsymbol{x})^T (\boldsymbol{A}\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x}\|_2^2 \geq 0.$$

Since the norm is always non-negative.

Calculus

Problem 7: Consider the following function $f : \mathbb{R} \to \mathbb{R}$

$$f(x) = \frac{1}{2} a x^2 + b x + c$$

We are interested in solving the following optimization problem

$$\min_{x \in \mathbb{R}} \; f(x)$$

a) Under what conditions does this optimization problem have (i) a unique solution, (ii) infinitely many solutions or (iii) no solution? Justify your answer.

We obtain a solution by setting the derivative to zero, i.e. $f'(x) = ax + b = 0$, and checking if the second derivative is positive, i.e. $f''(x) = a > 0$. Hence, we obtain
(i) a single solution if $a > 0$,
(ii) infinitely many solutions if $a = b = 0$, and
(iii) no solution if $a = 0$, $b \neq 0$ or $a < 0$.

b) Assume that the optimization problem has a unique solution. Write down the closed-form expression for $x^\star$ that minimizes the objective function, i.e. find $x^\star = \arg\min_{x \in \mathbb{R}} f(x)$.

This is an important question that we will encounter in multiple disguises throughout the lecture. Since we know that $a > 0$ we can solve it by simply setting the derivative to zero:

$$f'(x) = ax + b = 0 \quad \Leftrightarrow \quad x = -\frac{b}{a}$$

Problem 8: Consider the following function $g : \mathbb{R}^N \to \mathbb{R}$

$$g(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + c$$

where $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is a symmetric, PSD matrix, $\boldsymbol{b} \in \mathbb{R}^N$ and $c \in \mathbb{R}$.

We are interested in solving the following optimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^N} \; g(\boldsymbol{x})$$

a) Compute the Hessian $\nabla^2 g(\boldsymbol{x})$ of the objective function. Under what conditions does this optimization problem have a unique solution?

The Hessian is defined as the matrix of partial derivatives (see Section 4.2 of the Stanford Linear Algebra Review and Reference).

$$\begin{aligned}\partial_{x_l}\partial_{x_k} g(\boldsymbol{x}) &= \partial_{x_l}\partial_{x_k}\left(\frac{1}{2}\sum_i\sum_j x_i A_{ij} x_j + \sum_i b_i x_i + c\right) = \partial_{x_l}\left(\frac{1}{2}\sum_i A_{ik}x_i + \frac{1}{2}\sum_j A_{kj}x_j + b_k\right)\\ &= \partial_{x_l}\left(\sum_j A_{kj}x_j + b_k\right) = A_{kl}\\ \nabla^2 g(\boldsymbol{x}) &= \boldsymbol{A}\end{aligned}$$

A differentiable function has a unique minimum if the Hessian $\nabla^2 g(\boldsymbol{x})$ is positive definite for all $\boldsymbol{x}$. For our case, this means that $g(\boldsymbol{x})$ has a unique minimum if $\boldsymbol{A}$ is positive definite.
Note that if we extend the definition of PSD to non-symmetric matrices we would obtain $\nabla^2 g(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{A} + \boldsymbol{A}^T)$*, which is the symmetric part of the matrix* $\boldsymbol{A}$ *and PSD if* $\boldsymbol{A}$ *is PSD. The rest of the exercise would work exactly the same way.*

b) Why is it necessary for a matrix $\boldsymbol{A}$ to be PSD for the optimization problem to be well-defined? *Hint: What happens if* $\boldsymbol{A}$ *has a negative eigenvalue?*

The problem is ill-defined if it does not have a minimum. We assume that the matrix $\boldsymbol{A}$ is not PSD and use the logical equivalence $a \Rightarrow b \Leftrightarrow \neg b \Rightarrow \neg a$.
Not being PSD means that the matrix $\boldsymbol{A}$ has a negative eigenvalue. The eigenvector associated with that eigenvalue gives a direction in which the function always decreases, i.e.

$$a\boldsymbol{v}^T\boldsymbol{A}a\boldsymbol{v} + \boldsymbol{b}^T a\boldsymbol{v} + c = a^2\boldsymbol{v}^T\lambda\boldsymbol{v} + a\boldsymbol{b}^T\boldsymbol{v} + c = a^2\lambda\|\boldsymbol{v}\|_2^2 + a\boldsymbol{b}^T\boldsymbol{v} + c,$$

with the eigenvalue $\lambda < 0$, the corresponding eigenvector $\boldsymbol{v}$, and $a \in \mathbb{R}$. As we take $a \to \infty$ the first term will dominate at some point $a > a_0$ since $a\boldsymbol{b}^T\boldsymbol{v} + c \in \mathcal{O}(a)$, with $\mathcal{O}(g(x)) = \{f \,|\, \forall c > 0, \exists x_0 > 0, \forall x > x_0 : |f(x)| \leq c|g(x)|\}$. Because $\lambda < 0$ this function will thus keep decreasing and therefore does not have a minimum. In summary, if $\boldsymbol{A}$ and therefore the Hessian is not PSD the problem does not have a solution, i.e. g does not have a minimum. Hence the problem is not well-defined.
Note that convexity is not related to the problem being well-defined. A non-convex function can still have a global minimum.

c) Assume that the matrix $\boldsymbol{A}$ is positive definite (PD). Write down the closed-form expression for $\boldsymbol{x}^\star$ that minimizes the objective function, i.e. find $\boldsymbol{x}^\star = \arg\min_{\boldsymbol{x}\in\mathbb{R}^N} g(\boldsymbol{x})$.

We solve this by setting the gradient to zero (as in the previous exercise). For this, we first calculate the gradient:

$$\begin{aligned}\partial_{x_k} g(\boldsymbol{x}) &= \partial_{x_k}\left(\frac{1}{2}\sum_i\sum_j x_i A_{ij} x_j + \sum_i b_i x_i + c\right)\\ &= \frac{1}{2}\sum_i A_{ik}x_i + \frac{1}{2}\sum_j A_{kj}x_j + b_k = \sum_j A_{kj}x_j + b_k,\\ \nabla g(\boldsymbol{x}) &= \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}\end{aligned}$$

Setting $\nabla g(\boldsymbol{x}) = 0$, we obtain

$$\boldsymbol{x}^\star = -\boldsymbol{A}^{-1}\boldsymbol{b}.$$

Since $\boldsymbol{A}$ is symmetric and PD it has no zero eigenvalues (see Exercise 5) and is therefore invertible.

Probability Theory

Notation. We use the following notation in our lecture

- For conciseness and to avoid clutter, we use $p(x)$ to denote multiple things
  1. If X is a discrete random variable, $p(x)$ denotes the probability mass function (PMF) of X at point x (usually denoted as $p_X(x)$ or $p(X = x)$ in the statistics literature).
  2. If X is a continuous random variable, $p(x)$ denotes the probability density function (PDF) of X at point x (usually denoted as $f_X(x)$ in the statistics literature).
  3. If $A \in \Omega$ is an event, $p(A)$ denotes the probability of this event (usually denoted as $\Pr(\{A\})$ or $\mathbb{P}(\{A\})$ in the statistics literature)

  You will mostly encounter (1) and (2) throughout the lecture. Usually, the meaning is clear from the context.
- Given the distribution $p(x)$, we may be interested in computing the expected value $\mathbb{E}_{p(x)}[f(x)]$ or, equivalently, $\mathbb{E}_X[f(x)]$. Usually, it is clear with respect to which distribution we are computing the expectation, so we omit the subscript and simply write $\mathbb{E}[f(x)]$.
- $x \sim p$ means that x is distributed (sampled) according to the distribution p. For example, $x \sim \mathcal{N}(\mu, \sigma^2)$ (or equivalently $p(x) = \mathcal{N}(x|\mu, \sigma^2)$) means that x is distributed according to the normal distribution with mean μ and variance σ^2.

Problem 9: Prove or disprove the following statement

$$p(a|b, c) = p(a|c) \Rightarrow p(a|b) = p(a)$$

Disprove by counterexample: We have a coin and do not know if it is fair, i.e. $p(A = T) = 0.5$, or unfair, i.e. $p(A = T) = 1$. C denotes the event whether the coin is fair $C = F$ or unfair $C = U$. Both of these events have equal probability $p(C = F) = p(C = U) = 0.5$. We perform two coin tosses

A and B. When we know which coin we have, coin tosses A and B are of course independent, i.e. $p(a|b,c) = p(a|c)$. However, if we do not observe C, then

$$p(A=T) = p(A=T|C=F)p(C=F) + p(A=T|C=U)p(C=U) = \frac{3}{4},$$

$$\begin{aligned} p(A=T|B=T) &= \frac{p(A=T,B=T)}{p(B=T)} \\ &= \frac{p(A=T,B=T|C=F)p(C=F) + p(A=T,B=T|C=U)p(C=U)}{p(B=T)} \\ &= \frac{\begin{array}{l} p(A=T|B=T,C=F)p(B=T|C=F)p(C=F) \\ \quad + p(A=T|B=T,C=U)p(B=T|C=U)p(C=U) \end{array}}{p(B=T)} \\ &= \frac{1/2 \cdot 1/2 \cdot 1/2 + 1 \cdot 1 \cdot 1/2}{3/4} = \frac{5/8}{3/4} = \frac{5}{6}. \end{aligned}$$

Therefore, $p(A=T) \neq p(A=T|B=T)$ and A and B are not independent. □

Problem 10: Prove or disprove the following statement

$$p(a|b) = p(a) \Rightarrow p(a|b,c) = p(a|c)$$

Disprove by counterexample: Let the random variables A and B denote two independent dice rolls. $C = A + B$ denotes the sum of these dice rolls. Clearly, A and B are independent and therefore $p(a|b) = p(a)$. However, when we observe their sum the two become dependent. E.g. if we observe $C = 3$, then $p(A = 1) = 1/2$ and $p(A = 2) = 1/2$. However, if we observe $B = 2$, then $p(A = 1|B = 2) = 1$. □

This fact is quite interesting. It means that two independent random variables can *become dependent* when observing a different variable, which is known as the *explaining away effect* (or selection bias, Berkson's paradox) in the literature. For example, if students are admitted to a university either by having good school grades or by showing excellent athletic performance, then these two attributes will be negatively correlated in the student population, even though they are independent in general.

Problem 11: You are given the joint PDF $p(a,b,c)$ of three continuous random variables. Show how the following expressions can be obtained using the rules of probability

1. $p(a)$
2. $p(c|a,b)$
3. $p(b|c)$

$$p(a) = \int\int p(a,b,c)dbdc$$

$$p(c \mid a,b) = \frac{p(a,b,c)}{p(a,b)} = \frac{p(a,b,c)}{\int p(a,b,c)dc}$$

$$p(b \mid c) = \frac{p(b,c)}{p(c)} = \frac{\int p(a,b,c)da}{\int\int p(a,b,c)dadb}$$

Problem 12: Researchers have developed a test which determines whether a person has a rare disease. The test is fairly reliable: if a person is sick, the test will be positive with 95% probability, if a person is healthy, the test will be negative with 95% probability. It is known that $\frac{1}{1000}$ of the population have this rare disease. A person (chosen uniformly at random from the population) takes the test and obtains a positive result. What is the probability that the person has the disease?

We can use Bayes' theorem to solve this. We denote the event of having the disease by D and a positive test by T.

$$p(D|T) = \frac{p(T|D)p(D)}{p(T)} = \frac{p(T|D)p(D)}{p(T|D)p(D) + p(T|\neg D)p(\neg D)} = \frac{0.95 \cdot 0.001}{0.95 \cdot 0.001 + 0.05 \cdot 0.999} = 0.0187$$

Problem 13: Let $X \sim \mathcal{N}(\mu, \sigma^2)$, and $f(x) = ax + bx^2 + c$. What is $\mathbb{E}[f(x)]$?

Using the linearity of expectation and $\sigma^2 = \mathbb{E}[x^2] - \mathbb{E}[x]^2$ we obtain

$$\mathbb{E}[f(x)] = \mathbb{E}[ax + bx^2 + c] = a\,\mathbb{E}[x] + b\,\mathbb{E}[x^2] + c = a\mu + b(\sigma^2 + \mu^2) + c$$

Problem 14: Let $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $g(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$ (where $\boldsymbol{A} \in \mathbb{R}^{N\times N}$). What are the values of the following expressions:

- $\mathbb{E}[g(\boldsymbol{x})]$,
- $\mathbb{E}[g(\boldsymbol{x})g(\boldsymbol{x})^T]$,
- $\mathbb{E}[g(\boldsymbol{x})^T g(\boldsymbol{x})]$,
- the covariance matrix $\mathrm{Cov}[g(\boldsymbol{x})]$.

$$\begin{aligned}
\mathbb{E}[g(\boldsymbol{x})] &= \boldsymbol{A}\,\mathbb{E}[\boldsymbol{x}] = \boldsymbol{A}\boldsymbol{\mu} \\
\mathbb{E}[g(\boldsymbol{x})g(\boldsymbol{x})^T] &= \mathbb{E}[\boldsymbol{A}\boldsymbol{x}(\boldsymbol{A}\boldsymbol{x})^T] = \mathbb{E}[\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{A}^T] = \boldsymbol{A}\,\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]\boldsymbol{A}^T = \boldsymbol{A}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T)\boldsymbol{A}^T \\
\mathbb{E}[g(\boldsymbol{x})^T g(\boldsymbol{x})] &= \mathbb{E}[(\boldsymbol{A}\boldsymbol{x})^T\boldsymbol{A}\boldsymbol{x}] = \mathbb{E}[\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}] = \mathbb{E}[\mathrm{Tr}(\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x})] \\
&= \mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{A}^T)] = \mathrm{Tr}(\mathbb{E}[\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{A}^T]) \\
&= \mathrm{Tr}(\boldsymbol{A}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T)\boldsymbol{A}^T) \\
\mathrm{Cov}[g(\boldsymbol{x})] &= \mathbb{E}[g(\boldsymbol{x})g(\boldsymbol{x})^T] - \mathbb{E}[g(\boldsymbol{x})]\,\mathbb{E}[g(\boldsymbol{x})]^T = \boldsymbol{A}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T)\boldsymbol{A}^T - \boldsymbol{A}\boldsymbol{\mu}\boldsymbol{\mu}^T\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T
\end{aligned}$$

Note that we have used the fact that taking the trace of a scalar does not have any effect, the trace's cyclic property and that it is defined as a sum and can therefore be pulled out of the expectation when proving $\mathbb{E}[g(\boldsymbol{x})^T g(\boldsymbol{x})]$.