Mathematical foundations for ML - UM6P class

Final Exam

February 9th, 2024

The goal of the exam is to consolidate the concepts we have seen in class. Some rules:

- **You don't need to solve all the exercises to get the maximal score of 20/20.**
- Correctly answering all the items of the *multiple choice questions* (Exercise 1) should get you around **one third** of the maximal score.
- If you have a partial answer to a question, write it down: it might get you a partial grade.
- Be succinct, but as mathematically thorough as you can.
- Don't waste too much time on **bonus questions** - try them only once you get bored, or if you are confident you can solve them

Section 1: Multiple Choice Questions

Exercise 1.
There is exactly one correct answer per question

1. What is the formula for the variance $Var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$ of a discrete random variable X taking values in a set $\mathcal{X}$, with probability mass function p?
   (a) $\sum_{x\in\mathcal{X}} |x - \sum_{x'\in\mathcal{X}} x'p(x')|p(x)$
   (b) $\sum_{x\in\mathcal{X}} x^2p(x) - \sum_{(x,x')\in\mathcal{X}\times\mathcal{X}} xx'p(x)p(x')$
   (c) $\int_{\mathcal{X}} \left(x - \int_{\mathcal{X}} x'p(x')dx'\right)^2 p(x)$
   (d) $\int_{\mathcal{X}} |x - \sum_{x'\in\mathcal{X}} p(x)|dx$
2. Consider a multiclass classification task with $C > 2$ possible labels. Assume that you use a neural network to transform your input. What should be the output size?
   (a) 1 (b) C (c) $2C$ (d) $C - 1$
3. Consider a multiclass classification task with $C > 2$ possible labels. Assume that you use a neural network to transform your input. What function should be used at the last layer to obtain class probabilities?
   (a) The sigmoid function σ (b) The softmax function $softmax$
   (c) The hyperbolic tangent tanh (d) A linear function $\boldsymbol{h} \mapsto \boldsymbol{A}\boldsymbol{h} + \boldsymbol{b}$
4. Given a real-valued function defined as $f(x) = \boldsymbol{b}^\top\boldsymbol{x} + \boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{b} \in \mathbb{R}^n$, and $\boldsymbol{A}$ is an $n \times n$ symmetric matrix. What is the gradient of f with respect to $\boldsymbol{x}$, $\nabla_x f(\boldsymbol{x})$?
   (a) $\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ (b) $2\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ (c) $\boldsymbol{A} + 2\boldsymbol{b}^\top\boldsymbol{x}$ (d) $\boldsymbol{A}\boldsymbol{x}$
5. What is the objective of the K-means algorithm as it iterates through the assignment and re-estimation steps?
   (a) To maximize the distance between different clusters' centroids.
   (b) To minimize the variance within each cluster.
   (c) To decrease the distortion, which is the average squared distance from a point to the center of its cluster.
   (d) To increase the number of clusters until each point forms its own cluster.
6. In the context of HMMs, which statement is true regarding the observations?
   (a) Observations are assumed to be independent of each other.
   (b) Observations are directly influenced by the hidden states.
   (c) Observations can only be continuous variables.
   (d) Observations can predict future stock market movements with 100% accuracy.
7. An HMM is designed with 4 hidden states and is used to model sequences of observations drawn from a 5-category multinomial distribution. How many parameters are in this HMM?
   (a) 36 (b) 40 (c) 44 (d) $\log(\pi^\pi) - \frac{e^\pi}{3}$
8. One key step in the EM (Expectation-Maximization) algorithm involves maximizing a lower bound on the log-likelihood with respect to a distribution q over the latent variables. What does maximizing this lower bound with respect to q imply in the context of the EM algorithm?
   (a) Maximizing the lower bound with respect to q means setting q equal to the conditional distribution of the latent variables given the observed data and current parameter estimates, i.e., the posterior distribution.

(b) Maximizing the lower bound with respect to q means setting q to a uniform distribution over all possible values of the latent variables.

(c) Maximizing the lower bound with respect to q means ignoring the latent variables and only focusing on the observed data.

(d) Maximizing the lower bound with respect to q involves assuming that the latent variables and observed data are independent.

9. What is the Frobenius norm of a matrix equal to?

(a) The square root of the sum of the squares of its singular values.

(b) The sum of its singular values.

(c) The maximum singular value.

(d) The sum of the squares of its elements.

10. Consider a matrix $\boldsymbol{A}$ that has an SVD given by $\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$. Given $k \leq r$, we obtain a rank-k matrix $\boldsymbol{A}_k$ by "truncating" the SVD after the first k terms:

$$\boldsymbol{A}_k := \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$$

According to the Eckhart-Young Theorem, which of the following best describes the relationship between $\boldsymbol{A}$, $\boldsymbol{A}_k$, and any other rank-k approximation $\boldsymbol{B}$?

(a) The Frobenius norm $\|\boldsymbol{A} - \boldsymbol{A}_k\|_F$ is always greater than $\|\boldsymbol{A} - \boldsymbol{B}\|_F$ for any $\boldsymbol{B}$ of rank at most k.

(b) The Frobenius norm $\|\boldsymbol{A} - \boldsymbol{A}_k\|_F$ is always equal to $\|\boldsymbol{A} - \boldsymbol{B}\|_F$ for any $\boldsymbol{B}$ of rank at most k.

(c) The Frobenius norm $\|\boldsymbol{A} - \boldsymbol{A}_k\|_F$ is the minimum Frobenius norm difference between $\boldsymbol{A}$ and any real $m \times n$ matrix $\boldsymbol{B}$ of rank at most k.

(d) The Frobenius norm $\|\boldsymbol{A} - \boldsymbol{A}_k\|_F$ increases as the rank k of $\boldsymbol{B}$ decreases.

Section 2

Exercise 2 (Introducing Regularization to a Linear Regression Problem)**.** Algorithmic trading strategies use factor models to quantify the relationship between the return of an asset and the sources of risk that are the main drivers of these returns. We wish to predict the return of an asset based on M features called **factor premia**.

For each time step t in $\{1, \ldots, T\}$, the return of the asset is denoted $r^{<t>}$, and the M features are denoted $(f_i^{<t>})_{1 \leq i \leq M}$.

The training data is composed of the **feature matrix** $\boldsymbol{F}$ of shape $T \times M$ and the output observation vector $\boldsymbol{R}$ of shape T:

1. **Introducing a basic Regression Model**

A **factor model** simply decomposes the return of the asset at time t (denoted $r^{<t>}$) into the set of **factor premia** $(f_i^{<t>})_{1 \leq i \leq M}$ as follows:

$$\forall t \in \{1, \ldots, T\} \quad r^{<t>} = \sum_{i=1}^{M} \beta_i f_i^{<t>} + \alpha + \epsilon, \tag{1}$$

where $(\beta_1, \ldots, \beta_M, \alpha)^\top \in \mathbb{R}^{M+1}$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Let $\boldsymbol{F}^{<1>}, \ldots, \boldsymbol{F}^{<T>}$ be the rows of the matrix $\boldsymbol{F}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M, \alpha)^\top \in \mathbb{R}^{M+1}$ be the vector of all the parameters we want to estimate using the training data.

We want to re-write the equation 1 in a matrix form as follows:

$$\forall t \in \{1, \ldots, T\} \quad r^{<t>} = \boldsymbol{\beta}^\top \tilde{\boldsymbol{F}}^{<t>} + \epsilon, \text{ with } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Let $\tilde{\boldsymbol{F}}$ be the matrix composed of the rows $\tilde{\boldsymbol{F}}^{<t>} \quad \forall t \in \{1, \ldots, T\}$

(a) Express $\tilde{\boldsymbol{F}}$ using the matrix $\boldsymbol{F}$.

(b) Show that the optimal vector of parameters $\boldsymbol{\beta}^* \in \mathbb{R}^{M+1}$ is defined as follows:

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta} \in \mathbb{R}^{M+1}}{\arg\min} \frac{1}{T} \|\tilde{\boldsymbol{F}}\boldsymbol{\beta} - \boldsymbol{R}\|_2^2$$

(c) Calculate the gradient of the function $J(\boldsymbol{\beta}) := \frac{1}{T} \|\tilde{\boldsymbol{F}}\boldsymbol{\beta} - \boldsymbol{R}\|_2^2$ with respect to $\boldsymbol{\beta}$

(d) Deduce the value of the optimal $\boldsymbol{\beta}^*$ if $\tilde{\boldsymbol{F}}^\top \tilde{\boldsymbol{F}}$ is invertible.

(e) We would like to optimize the function J using the Gradient Descent Algorithm. Describe the optimization process.

2. **Introducing regularization techniques to the Regression Model**

An important theoretical result of Machine Learning is the fact that a model's generalization error can be expressed as the sum of the three following errors: the **bias**, the **variance** and the **irreducible error**.

One popular approach to control the overfitting problem is that of **regularization**, which involves the addition of a penalty term to the error function to discourage the regression coefficients from reaching large values.

The added penalty turns the optimal linear regression coefficients into the solution to the following minimization problem:

$$\boldsymbol{\beta}^*_{\text{reg}} = \underset{\beta_{\text{reg}} \in \mathbb{R}^{M+1}}{\arg\min} \frac{1}{T} \left(||\tilde{\boldsymbol{F}}\boldsymbol{\beta}_{\text{reg}} - \boldsymbol{R}||_2^2 + \lambda \mathcal{S}(\boldsymbol{\beta}_{\text{reg}}) \right) \tag{2}$$

These **shrinkage methods** differ by how they calculate the penalty term. The most common versions for the linear regression model are the **ridge regression** and the **lasso regression**.

In this section, let us consider the **ridge regression**. The penalty term is then defined as follows:

$$\mathcal{S}(\boldsymbol{\beta}_{\text{reg}}) = \boldsymbol{\beta}_{\text{reg}}^\top \boldsymbol{\beta}_{\text{reg}}$$

Equation 2 becomes:

$$\boldsymbol{\beta}^*_{\text{reg}} = \underset{\boldsymbol{\beta}_{\text{reg}} \in \mathbb{R}^{M+1}}{\arg\min} \frac{1}{T} \left(||\tilde{\boldsymbol{F}}\boldsymbol{\beta}_{\text{reg}} - \boldsymbol{R}||_2^2 + \lambda \boldsymbol{\beta}_{\text{reg}}^\top \boldsymbol{\beta}_{\text{reg}} \right) \tag{3}$$

Show that the closed solution to the ridge regression problem defined in equation 3 is:

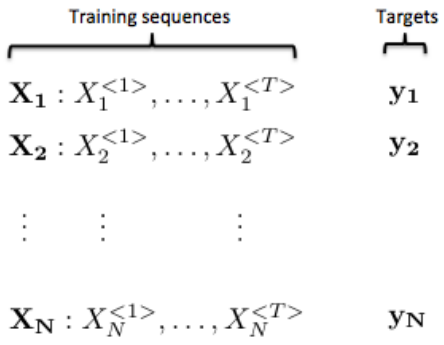
$$\boldsymbol{\beta}^*_{\text{reg}} = (\tilde{\boldsymbol{F}}^\top \tilde{\boldsymbol{F}} + \lambda \boldsymbol{I}_{M+1})^{-1} \tilde{\boldsymbol{F}}^\top \boldsymbol{R}$$

Exercise 3 (Building a Generative Classifier for Sentiment Analysis)**.**
We wish to create a sentiment analysis model to classify financial news into three possible labels: **positive**, **negative**, or **neutral**.

The training dataset is composed of N sentences $(\boldsymbol{X}_i)_{1 \leq i \leq N}$. Each sentence is composed of T words.

Let V be the vocabulary size. The first step of the processing consists of creating a dictionary to map each word to a discrete category in $\{1, \dots, V\}$.

We end up with N sequences $(\boldsymbol{X}_i)_{1 \leq i \leq N}$ of categories $X_i^{<t>}$ in $\{1, \dots, V\}$ for all $i \in \{1, \dots, N\}$ and for all $t \in \{1, \dots, T\}$ as shown in the following figure:



The three possible labels are encoded as follows: 0 for the negative sentiment, 1 for the neutral sentiment and 2 for the positive sentiment.

We would like to create a generative classifier. For that, we need to train three class conditional density functions, one for each target value $k \in \{0, 1, 2\}$

Each class conditional density function associated with the target $k \in \{0, 1, 2\}$ is parameterized by $\boldsymbol{\theta}_k$, which enables us to calculate $\mathbb{P}_{\boldsymbol{\theta}_k}(\boldsymbol{X}|y = k)$ for a given a sequence $\boldsymbol{X} = (X^{<1>}, \dots, X^{<T>}) \in \{1, \dots, V\}^T$.

Let $\boldsymbol{X} = (X^{<1>}, \dots, X^{<T>}) \in \{1, \dots, V\}^T$ be a new sequence.

1. Express $\mathbb{P}(y = k|\boldsymbol{X})$ as a function of $(\mathbb{P}(y = j))_{j \in \{0,1,2\}}$ and $(\mathbb{P}_{\theta_j}(\boldsymbol{X}|y = j))_{j \in \{0,1,2\}}$

2. **Using a Hidden Markov Model as a class conditional discrete density estimator on the discrete categories**

   Let k be in $\{0, 1, 2\}$. We would like to use a Hidden Markov Model (HMM) as a class conditional discrete density estimator $\mathbb{P}_{\boldsymbol{\theta}_k}(\boldsymbol{X}|y = k)$ (where $\boldsymbol{X} = (X^{<1>}, \dots, X^{<T>}) \in \{1, \dots, V\}^T$).

   Let M be the number of hidden states.

   (a) What are the parameters $\boldsymbol{\theta_k}$ of the HMM class conditional discrete density estimator?

   (b) By choosing reasonable values of V and M, compare the number of parameters of the HMM model with the number of parameters of a simple Markov Model on the observations.

   (c) Describe briefly the training method we use to estimate the parameters of the HMM?

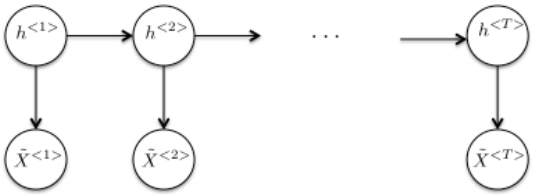
3. **Using a Hidden Markov Model as a class conditional density estimator on the embedding vectors**

   Instead of modelling $\mathbb{P}_{\boldsymbol{\theta}_k}(\boldsymbol{X}|y = k)$ for $\boldsymbol{X} = (X^{<1>}, \dots, X^{<T>}) \in \{1, \dots, V\}^T$, we would like to consider instead a class conditional density estimator on the sequences of some embedding vectors.

   Each category in $\{1, \dots, V\}$ represents a word and can be mapped into a D-dimensional vector, where each dimension encodes part of the information about the corresponding word.

   For each new sequence $\boldsymbol{X} = (X^{<1>}, \dots, X^{<T>}) \in \{1, \dots, V\}^T$, we define $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{X}}^{<1>}, \dots, \tilde{\boldsymbol{X}}^{<T>})$ the sequence of the embedding vectors $\tilde{\boldsymbol{X}}^{<t>} \in \mathbb{R}^D$ associated with the categories $X^{<t>} \in \{1, \dots, V\}$ for each $t \in \{1, \dots, T\}$.

   We would like to use an HMM model in order to learn the distribution of a sequence of D-dimensional embedding vectors $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{X}}^{<1>}, \dots, \tilde{\boldsymbol{X}}^{<T>})$ as represented in the following figure:



For each $t \in \{1, \dots, T\}$, the embeding vector $\tilde{\boldsymbol{X}}^{<t>} \in \mathbb{R}^D$ is associated with a hidden state $h^{<t>} \in \{1, \dots, M\}$.

Let m be in $\{1, \dots, M\}$. The emission probability distribution of the observation $\tilde{\boldsymbol{X}}^{<t>}$ conditioned on the hidden state $h^{<t>} = m$ is parameterized by a multivariate normal distribution with a mean vector $\boldsymbol{\mu}_m \in \mathbb{R}^D$ and a covariance matrix $\boldsymbol{\Sigma}_m \in \mathbb{R}^{D \times D}$, as explained in equation 4

$$\forall t \in \{1, \dots, T\} \; \forall m \in \{1, \dots, M\} \quad \tilde{\boldsymbol{X}}^{<t>} | h^{<t>} = m \sim \mathcal{N}_D(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \quad (4)$$

Let $\tilde{\boldsymbol{\theta}_k}$ be the set of the parameters of the HMM class conditional density estimator $\mathbb{P}_{\tilde{\boldsymbol{\theta}_k}}(\tilde{\boldsymbol{X}} | y = k)$ (where $\tilde{\boldsymbol{X}}$ is a sequence of embedding vectors $(\tilde{\boldsymbol{X}}^{<1>}, \dots, \tilde{\boldsymbol{X}}^{<T>})$

(a) What are the parameters $\tilde{\boldsymbol{\theta}_k}$ of the HMM class conditional density estimator on the embedding vectors?

(b) By choosing reasonable values for M, D and V, compare the number of parameters of the HMM model on the discrete observations in $\{1, \dots, V\}$ with the number of parameters of the HMM on the D-dimensional embedding vectors.

Section 3

Exercise 4 (Probability)**.**

1. **Covariance between two random variables:** We recall the definition of the covariance of two random variables X and Y is $Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$.

   Show that, when all quantities are defined:

   $$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

2. **Conditional independence:** We recall the definition of conditional independence between two discrete random variables X and Y on $\mathcal{X}$ and $\mathcal{Y}$ respectively, with probability mass functions p_X and p_Y respectively, conditional on a discrete random variable Z with probability mass function p_Z, taking values in a set $\mathcal{Z}$, and satisfying: $\forall z \in \mathcal{Z}, \quad p_Z(z) > 0$ (**We write** $X \perp Y \mid Z$):

   $$\forall (x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}, \quad p_{X,Y|Z}(x, y \mid z) = p_{X|Z}(x \mid z) p_{Y|Z}(y \mid z)$$

   (a) Verify that the $X \perp Y \mid Z$ is **equivalent** to the following:

   $$\forall (x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}, \quad p_{X|Y,Z}(x \mid y, z) = p_{X|Z}(x \mid z)$$

   (b) Prove that, for 4 discrete random variables X, Y, Z, W, if $X \perp Y \mid Z$ and $(X, Y) \perp W \mid Z$, then $X \perp W \mid Z$

3. **Information theory:** We define the cross-entropy of a distribution q (probability density function or probability mass function) relative to a distribution p over a given set as $H(p, q) = -\mathbb{E}_{X \sim p}[\log q(X)]$. This is linked to the Kullback-Leibler (KL) divergence of p from q as follows:

   $$H(p, q) = H(p) + D_{\mathrm{KL}}(p \| q),$$

   where $H(p)$ is the entropy (or differential entropy) of the distribution p: $H(p) = -\mathbb{E}_{X \sim p}[\log p(X)]$.

   Consider two univariate normal distributions $q(x) = \mathcal{N}(x \mid \mu, \sigma^2)$ and $p(x) = \mathcal{N}(x \mid m, s^2)$.

   (a) Evaluate $H(p, q)$ as a function of m, s, μ, σ

   (b) Deduce $D_{\mathrm{KL}}(p \| q)$

4. **The Beta distribution:** The Beta distribution is a useful distribution to model continuous bounded variables that are not uniform. Its support is the interval $[0, 1]$. It has two parameters, $a, b > 0$, called the *concentration* parameters. Its probability density function is:

   $$x \in [0, 1] \mapsto \mathcal{B}(x \mid a, b) = \frac{1}{B(a, b)} x^{a-1} (1 - x)^{b-1},$$

   where $B(a, b)$ is defined as:

   $$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)},$$

   where Γ is the Gamma function defined for every a > 0 as :

   $$\Gamma(a) := \int_0^\infty x^{a-1} e^{-x} dx$$

We recall two properties of the Gamma function, which you don't need to prove:

$$\forall n \in \{1, 2, 3, \dots\},\ \Gamma(n) = (n-1)!$$
$$\forall a > 0,\ \Gamma(a+1) = a\Gamma(a)$$

(a) Verify that if $a = b = 1$, $\mathcal{B}(a, b)$ is the uniform distribution on $[0, 1]$.

(b) Show that for a random variable $X \sim \mathcal{B}(a, b)$, $\mathbb{E}[X] = \frac{a}{a+b}$. **Hint** remember that $x \in [0, 1] \mapsto \mathcal{B}(x \mid a, b)$ is a valid pdf, meaning that $\int_0^1 \mathcal{B}(x \mid a, b)dx = 1$ for every $a, b > 0$.

(c) **Bonus question:** Evaluate $Var(X)$ for $X \sim \mathcal{B}(a, b)$.

5. **Bonus question:** Provide an example of a **discrete** random variable X that does not have a finite expectation.

Exercise 5 (Dropout VS L2 regularization)**.**
Dropout is a popular regularization technique used when training neural network, that consists of randomly dropping off some activations (neurons) during training, where the units dropped change randomly at every batch.

Consider a linear regression problem with input data $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ (i.e. we have n examples of dimension d that we represent in the matrix $\boldsymbol{X}$), weights $\boldsymbol{w} \in \mathbb{R}^d$ and and targets $\boldsymbol{y} \in \mathbb{R}^n$. Suppose that dropout is applied to the input (with probability $1 - p$ to drop the unit). Let $\boldsymbol{R} \in \mathbb{R}^{n \times d}$ be the dropout mask such that $\boldsymbol{R}_{ij} \sim \text{Bern}(p)$ are sampled i.i.d. from the Bernoulli distribution.

1. Suppose we want to minimize the *expected* sum of squared error (averaging out the randomness from dropout). Express the loss function $L(\boldsymbol{w})$ in matrix form. You should use the notation $\boldsymbol{A} \odot \boldsymbol{B}$ that represents the element-wise multiplication of two matrices $\boldsymbol{A}, \boldsymbol{B}$ of the same shape, i.e. $(\boldsymbol{A} \odot \boldsymbol{B})_{i,j} = \boldsymbol{A}_{i,j}\boldsymbol{B}_{i,j}$

2. Let Γ be a diagonal matrix with $\Gamma_{ii} = (\boldsymbol{X}^\top \boldsymbol{X})_{ii}^{1/2}$. Show that the loss function can be rewritten as $L(\boldsymbol{w}) = ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||\Gamma \boldsymbol{w}||^2$

3. **Bonus question:** Show that the solution to the minimization problem $\boldsymbol{w}^*$ satisfies

$$p\boldsymbol{w}^* = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \Gamma^2)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

where λ is a regularization coefficient depending on p. Compare with L_2 regularization.

Exercise 6 (Log determinant gradient)**.**
The goal of this exercise is to find the gradient of the transformation $f : \boldsymbol{A} \in S_n^+(\mathbb{R}) \mapsto \log(\det \boldsymbol{A}) \in \mathbb{R}$, where $S_n^+(\mathbb{R})$ is the the set of symmetric matrices in $\mathbb{R}^{n \times n}$ that are positive definite. This gradient is useful in many areas of statistics and machine learning, such as the Maximum Likelihood estimates of Gaussian parameters. Let $\boldsymbol{A}$ be a matrix in $S_n^+(\mathbb{R})$. We consider the following matrix norm, called the **spectral norm**:

$$||\boldsymbol{H}|| = \sup_{x \in \mathbb{R}^n, ||\boldsymbol{x}||_2 = 1} ||\boldsymbol{H}\boldsymbol{x}||_2,$$

where $\|\boldsymbol{y}\|_2^2 = \boldsymbol{y}^\top \boldsymbol{y} = \sum_{i=1}^n \boldsymbol{y}_i^2$. We recall that $||\boldsymbol{H}||$ is also the largest singular value of the matrix $\boldsymbol{H}$. **When $\boldsymbol{H}$ is symmetric**, it coincides with the largest absolute value of the eigenvalues of $\boldsymbol{H}$, i.e. $\|\boldsymbol{H}\|_2 = \max_{\lambda \in Sp(\boldsymbol{H})} |\lambda|$. We recall that it is a sub-multiplicative norm, meaning that: $\|\boldsymbol{H}\boldsymbol{H}'\| \leq \|\boldsymbol{H}\|\|\boldsymbol{H}'\|$.

1. Show that there exists $r > 0$ such that for all symmetric matrices $\boldsymbol{H}$ satisfying $||\boldsymbol{H}|| \leq r$, $\boldsymbol{A} + \boldsymbol{H}$ is positive definite. This results is known as the openness of the set of positive definite matrices. It is necessary to prove the openness of this set to be able to talk about the gradient of f

2. Show that for every symmetric $\boldsymbol{H}$ satisfying $||\boldsymbol{H}|| \leq r$, we have $f(\boldsymbol{A} + \boldsymbol{H}) - f(\boldsymbol{A}) = \sum_i \log(1 + \lambda_i(\boldsymbol{A}^{-1}\boldsymbol{H}))$, where $\lambda_1(\boldsymbol{H}'), \dots \lambda_n(\boldsymbol{H}')$ are the n eigenvalues of a symmetric positive definite matrix $\boldsymbol{H}'$

3. **Bonus question:** Show that the previous sum can be written as $f(\boldsymbol{A} + \boldsymbol{H}) - f(\boldsymbol{A}) = \text{Tr}(\boldsymbol{A}^{-1}\boldsymbol{H}) + g(\boldsymbol{H})$, where g is a function satisfying $\lim_{\boldsymbol{H} \to \boldsymbol{0}} \frac{g(\boldsymbol{H})}{||\boldsymbol{H}||} = 0$. (Hint: Use Taylor expansion and the fact that the matrix norm is sub-multiplicative, i.e. $||\boldsymbol{A}\boldsymbol{B}|| \leq ||\boldsymbol{A}||||\boldsymbol{B}||$.)

4. Conclude by finding the gradient of f at $\boldsymbol{A}$