UGE: M2 Maths-Info
Computational Foundations of Data Sciences
T. Lacombe 2023-2024

Exam - Computational Foundations of Data Sciences.

January 18, 2024

Note: This exam is quite long on purpose. You do not need to treat every question perfectly to have a good grade.

Formula cheat sheet

We provide below some formula and definitions seen during the lectures and that can be helpful (or not) to solve some of the questions presented in this exam.

In the following cheat sheet[1], $(x_i, y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ refers to a list of n couples of observations and labels belonging to spaces $\mathcal{X}$ and $\mathcal{Y}$. Whenever needed, we may assume that $\mathcal{X} = \mathbb{R}^d$ and either that $\mathcal{Y} = \mathbb{R}^k$ (regression task) or $\mathcal{Y} = \{1, \dots, k\}$ (classification task). In the regression case, we may introduce the observation matrix $X \in \mathbb{R}^{n \times d}$ with $X_{ij} = x_i[j]$ and the label matrix $Y \in \mathbb{R}^{n \times k}$ with $Y_{ij} = y_i[j]$. The notation F refers to a model, i.e. a map $F : \mathcal{X} \to \mathcal{Y}$.

- The *mean squared error* achieved by $F : \mathbb{R}^d \to \mathbb{R}^k$ on the dataset $(x_i, y_i)_{i=1}^n$ is defined as

$$\mathrm{MSE}(F, (x_i, y_i)_{i=1}^n) := \frac{1}{n} \sum_{i=1}^n |F(x_i) - y_i|^2 \tag{MSE}$$

- The *softmax* function is defined as

$$\mathrm{smax} : \mathbb{R}^k \ni (f_1, \dots, f_k) \mapsto \left(\frac{e^{f_i}}{\sum_{j=1}^k e^{f_j}} \right)_{i=1}^k . \tag{smax}$$

- The *Cross-entropy* loss of a model $F : \mathbb{R}^d \to \{1, \dots, k\}$ is defined as

$$\mathrm{CE}(F, (x_i, y_i)_{i=1}^n) := -\frac{1}{n} \sum_{i=1}^n \mathbf{y}_i \cdot \log[\mathrm{smax}(F(x_i))] \tag{CE}$$

  where $\mathbf{y}_i \in \mathbb{R}^k$ is the *one-hot-encoding* of y_i, that is $\mathbf{y}_i[j] = (1_{\{y_i = j\}})_{j=1}^k$.

- A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be convex if, for all $t \in (0, 1)$, one has

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y). \tag{cvx}$$

  It is α-strongly convex for some $\alpha > 0$ if

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y) - \frac{\alpha}{2} t(1-t)|x-y|^2. \tag{scvx}$$

  Assuming f is of class $\mathcal{C}^1$, it is β-smooth (for some $\beta > 0$) if its gradient $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is β-Lipschitz:

$$\forall x, y, \ |\nabla f(x) - \nabla f(y)| \leq \beta |x - y|. \tag{smooth}$$

[1] Beware, the notation can slightly differ in the exercises.

- A *PSD kernel* is a map $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that is symmetric and such that

$$\forall n \in \mathbb{N},\ \forall (\lambda_i)_{i=1}^n \in \mathbb{R}^n, \forall (x_i)_{i=1}^n \in \mathcal{X}^n,\quad \sum_{1 \le i,j \le n} \lambda_i \lambda_j K(x_i, x_j) \ge 0. \tag{PSD}$$

Exercise 1 : Cross-entropy. Explain, in your own words, what is the role of the cross-entropy (CE) and in which situations is it used as a loss (and why).

Exercise 2 : Multi-dimensional scaling (MDS). Let $X \in \mathbb{R}^{n \times D}$ be a set of n points in dimension D, and pick $d \ll D$. We assume that $n > D, d$. We denote by $x_i \in \mathbb{R}^D$ the i-th row of X. The goal of the multi-dimensional scaling problem is to find a point cloud $\hat{X} \in \mathbb{R}^{n \times d}$ such that the Gram matrix $\hat{X}\hat{X}^T \in \mathbb{R}^{n \times n}$ of $\hat{X}$ is as close as possible to the one of X that is given by $G := XX^T \in \mathbb{R}^{n \times n}$.

Formally, it means that $\hat{X}$ should minimize the objective function

$$L : \hat{X} \mapsto \|\hat{X}\hat{X}^T - G\|_2^2,$$

where we recall that $\|A\|_2^2 := \mathrm{Tr}(A^T A)$ for a matrix $A \in \mathbb{R}^{k \times l}$ $(k, l \in \mathbb{N})$.

Q1. Is this problem supervised or unsupervised? To which subclass (regression, clustering, etc.) does it belong?

Q2. Justify that G can be diagonalized in an orthonormal basis, and that it has non-negative eigenvalues, denoted by $\lambda_1 \ge \cdots \ge \lambda_n \ge 0$.

Let $Q \in \mathbb{R}^{n \times n}$ denote the transition matrix, that is $G = Q^T \Delta Q$ with $\Delta = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$.

Q3. Introduce the matrix $U = Q^T \underbrace{\begin{pmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_d} \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix}}_{\in \mathbb{R}^{n \times d}}$. Determine $L(U)$.

Q4. Prove that U is a minimizer of L.

Q5. Assume that we are not given the $(x_i)_{i=1}^n$, but only the matrix of their squared pairwise distances $D \in \mathbb{R}^{n \times n}$, i.e. $D_{ij} := \|x_i - x_j\|_2^2$. Show that we can retrieve G by setting $G = -\frac{1}{2} CDC$ where C is a matrix that will be determined.

Q6. Propose a comparison of PCA and MDS.

Exercise 3 : The (Kernel) Ridge Regression. Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}^k$, and consider a set of n observations $X \in \mathbb{R}^{n \times d}$ and labels $Y \in \mathbb{R}^{n \times k}$. Recall that the Ridge regression consists of considering the linear model $x \mapsto xM^*$ where $M^* \in \mathbb{R}^{d \times k}$, where M^* is chosen as a minimizer of

$$M \mapsto \|XM - Y\|_2^2 + \lambda \|M\|_2^2, \tag{1}$$

for some parameter $\lambda > 0$.

Q1. Recall briefly why introducing the penalty $\lambda\|M\|_2^2$ in (1) can be useful.

Q2. Prove that for all but finitely many $\lambda > 0$, (1) admits a unique minimizer M^* for which an explicit expression is accessible (that you must determine).

Q3. We now assume that our observations belong to an abstract set $\mathcal{X}$ equipped with a PSD kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, with corresponding RKHS $\mathcal{H}$ and feature map φ.

(a) Recall briefly what the "kernel trick" refers to.

(b) Show that the kernel trick can be applied to the Ridge regression model.

Exercise 4: The implicit bias of Gradient Descent in Linear Regression. Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$. Consider a dataset with n observations $X \in \mathbb{R}^{n \times d}$ and labels $Y \in \mathbb{R}^n$. We assume that $n < d$ and that X is of rank n, so that $XX^T \in \mathbb{R}^{n \times n}$ is non-singular (note: in the course we usually made the assumption that $X^T X \in \mathbb{R}^{d \times d}$ was non-singular). We consider the linear regression problem:

$$\min_{w \in \mathbb{R}^d} f(w), \quad \text{where} \quad f(w) = \|Xw - Y\|_2^2.$$

Q1. Let $w^* = X^T(XX^T)^{-1}y$. Justify that w^* is a minimizer of f.

Q2. Determine the set of minimizers of f.

Q3. Show that w^* is the element of minimal norm in this set.

Q4. We now consider the gradient descent on f with parameter $\eta > 0$ chosen small enough to ensure its convergence, starting from $w_0 = 0$, and thus defined for $t \in \mathbb{N}$ as

$$w_{t+1} = w_t - \eta \nabla f(w_t).$$

(a) Show that for all t, there exists a unique $u_t \in \mathbb{R}^n$ such that $w_t = X^T u_t$.

(b) Determine the recursion equation satisfied by u_t.

(c) Show that it can be interpreted as a gradient descent performed on some map $g : \mathbb{R}^n \to \mathbb{R}$ that will be determined.

(d) Determine the limit of $(u_t)_t$.

(e) Deduce that $w_t \to w^*$ as $t \to \infty$.

(f) Comment the sentence "there is an implicit bias when running a gradient descent to minimize a linear regression".

Exercise 5 : (Problem: Sliced Wasserstein Kernel between probability measures) Let $\mathcal{P}_2(\mathbb{R}^d)$ denote the set of probability measures supported on $\mathbb{R}^d$ with finite second order moment, i.e. $\int |x|^2 \mathrm{d}\mu(x) < \infty$ for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, and that admit a density ρ (with respect to the Lebesgue measure on $\mathbb{R}^d$), i.e. $\mathrm{d}\mu(x) = \rho(x)\mathrm{d}x$. The goal of this exercise is to define a positive definite (PD) kernel on this space. The problem decomposes in two parts: the first one defines a notion of Gaussian Kernel on arbitrary metric spaces (under some condition), and the second part apply this theory to $\mathcal{P}_2(\mathbb{R}^d)$ equipped with a suitable metric.

Part I: Gaussian kernel in metric spaces. We first introduce some notation and definitions.

- Let $(\mathcal{X}, d)$ be a (non-empty) metric space. We say that it is Hilbertian if there exists a Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $d(x, x') = \|\phi(x) - \phi(x')\|_{\mathcal{H}}$ for all $x, x' \in \mathcal{X}$, where $\|\cdot\|_{\mathcal{H}} := \langle \cdot, \cdot \rangle_{\mathcal{H}}^{\frac{1}{2}}$.
- The Gaussian kernel on $(\mathcal{X}, d)$ with parameter $t > 0$ is the map $K_t : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined as

$$K_t(x, x') := \exp\left(-t d(x, x')^2\right). \tag{2}$$

- A function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be *conditionally negative definite* (CND) if it is symmetric and if, for any $n \in \mathbb{N}$, $a_1, \dots, a_n \in \mathbb{R}$ such that $\sum_{i=1}^n a_i = 0$, and $x_1, \dots, x_n \in \mathcal{X}$, we have

$$\sum_{1 \le i,j \le n} a_i a_j f(x_i, x_j) \le 0. \tag{3}$$

We also say that $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is *positive definite* (PD) if for any $n \in \mathbb{N}$, $a_1, \dots, a_n \in \mathbb{R}$, we have

$$\sum_{1 \le i,j \le n} a_i a_j f(x_i, x_j) \ge 0. \tag{4}$$

Recall that if f is PD, then there exists an Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $f(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$ for all $x, y \in \mathcal{X}$.

Note: we omit the terminology "semi" for the sake of conciseness.

The goal of this first part is to prove the following result:

Theorem: Let $(\mathcal{X}, d)$ be a metric space. The Gaussian kernel $K_t : (x, x') \mapsto \exp(-t d(x, x')^2)$ defines a PD kernel on $\mathcal{X}$ for all $t > 0$ if and only if $(\mathcal{X}, d)$ is Hilbertian.

Q1. Let $x_0 \in \mathcal{X}$ be any arbitrary point. Define $\varphi(x, x') := d(x, x_0)^2 + d(x', x_0)^2 - d(x, x')^2$. Show that φ is positive definite if and only if $(x, x') \mapsto d(x, x')^2$ is conditionally negative definite.

Q2.
- (a) Show that if f_1 and f_2 are PD, so are $f_1 + f_2$ and $f_1 \cdot f_2$.
- (b) Let $(f_n)_n$ be a sequence of function from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$, all being PD (resp. CND), and converging pointwise toward some limit f (that is, for all x, x', $f_n(x, x') \to f(x, x')$). Show that f is also PD (resp. CND).
- (c) Deduce that if f is PD, so is the function $(x, x') \mapsto \exp(f(x, x'))$.

Q3.
- (a) Assume that K_t is PD for all $t > 0$. Show that $(x, x') \mapsto d(x, x')^2$ must be CND.
- (b) Conversely, assume that $(x, x') \mapsto d(x, x')^2$ is CND, and deduce that K_t is PD for all $t > 0$.

Q4.
- (a) Assume that $(\mathcal{X}, d)$ is Hilbertian. Show that $(x, x') \mapsto d(x, x')^2$ is CND.
- (b) Conversely, assume that $(x, x') \mapsto d(x, x')^2$ is CND, and show that $(\mathcal{X}, d)$ is Hilbertian.

  Hint: You may consider the function φ defined above and the associated Reproducing Kernel Hilbert Space.

Q5. Conclude.

Conditioning

Part II: The Sliced-Wasserstein kernel between probability measures. Let us start with some definitions as well.

- The *Wasserstein distance* on $\mathcal{P}_2(\mathbb{R}^d)$ is defined for $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ as

$$W_2(\mu,\nu) := \inf_{\pi \in \Pi(\mu,\nu)} \left(\iint_{\mathbb{R}^d \times \mathbb{R}^d} |x-y|^2 \mathrm{d}\pi(x,y) \right)^{\frac{1}{2}}, \tag{5}$$

where $\Pi(\mu,\nu)$ denote the set of probability measures supported on $\mathbb{R}^d \times \mathbb{R}^d$ that have μ and ν as first and second marginal, respectively. An optimal π in (5) is called an *optimal transport plan* between μ and ν. We admit that it defines a distance on $\mathcal{P}_2(\mathbb{R}^d)$.

We also admit the following theorem, due to Brenier.

Brenier's Theorem: Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. There exists a unique optimal transport plan between μ and ν. It is *deterministic*, that is, there exists a map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $\mathrm{spt}(\pi) \subset \{(x, T(x)), x \in \mathbb{R}^d\}$. Furthermore, $T = \nabla\phi$, where $\phi : \mathbb{R}^d \to \mathbb{R}$ is a convex function.

- For $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\theta \in S^{d-1}$ (the unit sphere in $\mathbb{R}^d$), let $p_\theta : x \mapsto \langle x, \theta \rangle$ denote the projection on the line span by θ, and $p_\theta \# \mu$ denote the *pushforward* of μ by p_θ, i.e. the measure in $\mathcal{P}_2(\mathbb{R})$ defined as $p_\theta \# \mu(A) = \mu(p_\theta^{-1}(A))$ for all Borel $A \subset \mathbb{R}$.
- For $\mu \in \mathcal{P}_2(\mathbb{R})$ a probability measure supported on the one-dimensional line, let $F_\mu(t) := \mu((-\infty, t]) \in [0,1]$ denote its cumulative distribution function defined for $t \in \mathbb{R}$. Let Q_μ be the corresponding *quantile function*, defined for $q \in [0,1]$ as $Q_\mu(q) := \inf\{t,\ F_\mu(t) \geq q\} \in \overline{\mathbb{R}}$ (its pseudo-inverse).
- The *Sliced Wasserstein distance* on $\mathcal{P}_2(\mathbb{R}^d)$ is defined as

$$SW_2(\mu,\nu) = \left(\int_{\theta \in S^{d-1}} W_2(p_\theta \# \mu, p_\theta \# \nu)^2 \mathrm{d}\theta \right)^{\frac{1}{2}}. \tag{6}$$

We also admit for the sake of conciseness that it defines a distance on $\mathcal{P}_2(\mathbb{R}^d)$.

- Eventually, the Sliced Wasserstein Kernel on $\mathcal{P}_2(\mathbb{R}^d)$ is the Gaussian kernel associated to SW_2, defined for $t > 0$ as

$$\mathrm{SWK}_t(\mu,\nu) = \exp\left(-tSW_2(\mu,\nu)^2\right). \tag{7}$$

The goal of this part is to prove that SWK_t defines a proper PD kernel for all $t > 0$.

Q1. Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$. Show that there exists a unique non-decreasing map T such that $T\#\mu = \nu$, where $T\#\mu$ denotes the measure on $\mathbb{R}$ defined as $T\#\mu(A) = \mu(T^{-1}(A))$ for all Borel $A \subset \mathbb{R}$.

Hint: You may use the cumulative distribution functions and / or the quantile functions.

Q2. Deduce that for $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$, one has

$$W_2(\mu,\nu)^2 = \int_0^1 |Q_\mu(q) - Q_\nu(q)|^2 \mathrm{d}q. \tag{8}$$

Q3. Deduce that the metric space $(\mathcal{P}_2(\mathbb{R}), W_2)$ is Hilbertian.

Q4. Using Part I and the previous questions, deduce that SWK_t defines a PD kernel on $\mathcal{P}_2(\mathbb{R}^d)$ for all $t > 0$.

Q5. Why couldn't we simply consider the kernel $(\mu,\nu) \mapsto \exp(-tW_2(\mu,\nu)^2)$?

UGE: M2 Maths-Info
Computational Foundations of Data Sciences
T. Lacombe 2024-2025

Exam - Computational Foundations of Data Sciences.

Note: This exam is quite long on purpose. You do not need to treat every question perfectly to have a good grade. The current scale is over 100 points and will be (non-linearly) rescaled to 12 points for the final grade of the course.

Formula cheat sheet

We provide below some formula and definitions seen during the lectures and that can be helpful (or not) to solve some of the questions presented in this exam.

In the following cheat sheet[1], $(x_i, y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ refers to a list of n couples of observations and labels.

- The *softmax* function is defined as

$$\text{smax} : \mathbb{R}^k \ni (f_1, \dots, f_k) \mapsto \left(\frac{e^{f_i}}{\sum_{j=1}^k e^{f_j}} \right)_{i=1}^k . \tag{smax}$$

- The *Cross-entropy* loss of a model $F : \mathbb{R}^d \to \{1, \dots, k\}$ is defined as

$$\text{CE}(F, (x_i, y_i)_{i=1}^n) := -\frac{1}{n} \sum_{i=1}^n \mathrm{y}_i \cdot \log[\text{smax}(F(x_i))] \tag{CE}$$

  where $\mathrm{y}_i \in \mathbb{R}^k$ is the *one-hot-encoding* of y_i, that is $\mathrm{y}_i[j] = (1_{\{y_i = j\}})_{j=1}^k$.

- A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be convex if, for all $t \in (0,1)$, one has

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y). \tag{cvx}$$

  It is α-strongly convex for some $\alpha > 0$ if

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y) - \frac{\alpha}{2} t(1-t)|x-y|^2. \tag{scvx}$$

  Assuming f is of class $\mathcal{C}^1$, it is β-smooth ($\beta > 0$) if $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is β-Lipschitz:

$$\forall x, y, \ |\nabla f(x) - \nabla f(y)| \leq \beta |x - y|. \tag{smooth}$$

- A *PSD kernel* is a map $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that is symmetric and such that

$$\forall n \in \mathbb{N}, \ \forall (\lambda_i)_{i=1}^n \in \mathbb{R}^n, \forall (x_i)_{i=1}^n \in \mathcal{X}^n, \quad \sum_{1 \leq i,j \leq n} \lambda_i \lambda_j K(x_i, x_j) \geq 0. \tag{PSD}$$

[1] Beware, the notation can slightly differ in the exercises.

UGE: M2 Maths-Info
Computational Foundations of Data Sciences
T. Lacombe

2024-2025

Exam - Computational Foundations of Data Sciences.

Note: This exam is quite long on purpose. You do not need to treat every question perfectly to have a good grade. The current scale is over 100 points and will be (non-linearly) rescaled to 12 points for the final grade of the course.

Formula cheat sheet

We provide below some formula and definitions seen during the lectures and that can be helpful (or not) to solve some of the questions presented in this exam.

In the following cheat sheet[1], $(x_i, y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ refers to a list of n couples of observations and labels.

- The *softmax* function is defined as

$$\text{smax} : \mathbb{R}^k \ni (f_1, \dots, f_k) \mapsto \left(\frac{e^{f_i}}{\sum_{j=1}^k e^{f_j}} \right)_{i=1}^k . \tag{smax}$$

- The *Cross-entropy* loss of a model $F : \mathbb{R}^d \to \{1, \dots, k\}$ is defined as

$$\text{CE}(F, (x_i, y_i)_{i=1}^n) := -\frac{1}{n} \sum_{i=1}^n \mathbf{y}_i \cdot \log[\text{smax}(F(x_i))] \tag{CE}$$

  where $\mathbf{y}_i \in \mathbb{R}^k$ is the *one-hot-encoding* of y_i, that is $\mathbf{y}_i[j] = (1_{\{y_i = j\}})_{j=1}^k$.

- A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be convex if, for all $t \in (0,1)$, one has

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y). \tag{cvx}$$

  It is α-strongly convex for some $\alpha > 0$ if

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y) - \frac{\alpha}{2} t(1-t)|x-y|^2. \tag{scvx}$$

  Assuming f is of class $\mathcal{C}^1$, it is β-smooth ($\beta > 0$) if $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is β-Lipschitz:

$$\forall x, y, \ |\nabla f(x) - \nabla f(y)| \leq \beta |x - y|. \tag{smooth}$$

- A *PSD kernel* is a map $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that is symmetric and such that

$$\forall n \in \mathbb{N}, \ \forall (\lambda_i)_{i=1}^n \in \mathbb{R}^n, \forall (x_i)_{i=1}^n \in \mathcal{X}^n, \quad \sum_{1 \leq i,j \leq n} \lambda_i \lambda_j K(x_i, x_j) \geq 0. \tag{PSD}$$

[1] Beware, the notation can slightly differ in the exercises.

Exercise 1 : Overfitting (10 pts). Explain, in your own words, what "overfitting" refers to in machine learning, and how this phenomenon may be mitigated.

Exercise 2 : Multi-dimensional scaling (MDS) (15 pts). Let $X \in \mathbb{R}^{n \times D}$ be a set of n points in dimension D, and pick $d \ll D$. We assume that $n > D, d$. We denote by $x_i \in \mathbb{R}^D$ the i-th row of X. The goal of the multi-dimensional scaling problem is to find a point cloud $\hat{X} \in \mathbb{R}^{n \times d}$ such that the Gram matrix $\hat{X}\hat{X}^T \in \mathbb{R}^{n \times n}$ of $\hat{X}$ is as close as possible to the one of X that is given by $G := XX^T \in \mathbb{R}^{n \times n}$.

Formally, it means that $\hat{X}$ should minimize the objective function

$$L : \hat{X} \mapsto \|\hat{X}\hat{X}^T - G\|_2^2,$$

where we recall that $\|A\|_2^2 := \mathrm{Tr}(A^T A)$ for a matrix $A \in \mathbb{R}^{k \times l}$ $(k, l \in \mathbb{N})$.

1. Is this problem supervised or unsupervised? To which subclass (regression, clustering, etc.) does it belong?
2. Justify that G can be diagonalized in an orthonormal basis, and that it has non-negative eigenvalues, denoted by $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$.

   Let $Q \in \mathbb{R}^{n \times n}$ denote the transition matrix, that is $G = Q^T \Delta Q$ with $\Delta = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$.
3. Introduce the matrix $U = Q^T \underbrace{\begin{pmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_d} \\ 0 & \ldots & 0 \\ \vdots & & \vdots \\ 0 & \ldots & 0 \end{pmatrix}}_{\in \mathbb{R}^{n \times d}}$. Determine $L(U)$.
4. Prove that U is a minimizer of L.
5. Assume that we are not given the $(x_i)_{i=1}^n$, but only the matrix of their squared pairwise distances $D \in \mathbb{R}^{n \times n}$, i.e. $D_{ij} := \|x_i - x_j\|_2^2$. Show that we can retrieve G by setting $G = -\frac{1}{2} CDC$ where C is a matrix that will be determined.
6. Propose a comparison of PCA and MDS.

Exercise 3 : The (Kernel) Ridge Regression (15 pts). Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}^k$, and consider a set of n observations $X \in \mathbb{R}^{n \times d}$ and labels $Y \in \mathbb{R}^{n \times k}$. Recall that the Ridge regression consists of considering the linear model $x \mapsto xM^*$ where $M^* \in \mathbb{R}^{d \times k}$, where M^* is chosen as a minimizer of

$$M \mapsto \|XM - Y\|_2^2 + \lambda \|M\|_2^2, \tag{1}$$

for some parameter $\lambda > 0$.

1. Recall briefly why introducing the penalty $\lambda\|M\|_2^2$ in (1) can be useful.
2. Prove that for all $\lambda > 0$, (1) admits a unique minimizer M^* for which an explicit expression is accessible (that you must determine).
3. We now assume that our observations belong to an abstract set $\mathcal{X}$ equipped with a PSD kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, with corresponding RKHS $\mathcal{H}$ and feature map φ.

f(x)

(a) Recall briefly what the "kernel trick" refers to.

(b) Show that the kernel trick can be applied to the Ridge regression model.

Exercise 4 : Properties of convex functions (15 pts) Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, assumed to be of class $\mathcal{C}^2$.

1. Prove that for all $x, y \in D(f)$ (the domain of f), one has
$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle . \tag{2}$$
2. Deduce that the gradient of f should be monotone, that is for all $x, y \in D(f)$,
$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0. \tag{3}$$
3. Deduce also that if $\nabla f(x) = 0$, then x should be a global minimizer of f.
4. Prove that for all $x \in D(f)$, the Hessian matrix of f at x should be positive semi-definite (i.e. all its eigenvalues should be ≥ 0).

Now, assume furthermore that $f : \mathbb{R}^d \to \mathbb{R}$ is α-strongly convex, for some $\alpha > 0$.

5. Show that for all $x, y \in D(f)$, one has
$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|^2. \tag{4}$$
6. Show that the function $x \mapsto f(x) - \frac{\alpha}{2}\|x\|^2$ is convex.
7. Show that the eigenvalues of the Hessian matrix of f at any $x \in D(f)$ should be larger than α.
8. Show that f admits a unique minimizer x^*.
9. Prove that f satisfies the PL-inequality
$$0 \leq f(x) - f(x^*) \leq \frac{1}{2\alpha}\|\nabla f(x)\|^2. \tag{5}$$

Exercise 5 : Convexity of the Logistic Regression (15 pts) The goal of this exercise is to show the convexity of the cross-entropy loss for the logistic regression. More precisely, given $(x_i, y_i)_{i=1}^n$, with $x_i \in \mathbb{R}^d$ and $\mathbf{y}_i = (0, \dots, 0, 1, 0, \dots, 0) \in \mathbb{R}^K$ (K being the number of classes), we want to prove that
$$L : \mathbb{R}^{K \times d} \ni \theta \mapsto -\sum_{i=1}^{n} \mathbf{y}_i \cdot \log(\mathrm{smax}(\theta x_i)) \tag{6}$$
is convex, where $\mathrm{smax}(a) = \left(\frac{e^{a[j]}}{\sum_{j'=1}^{K} e^{a[j']}}\right)_{j=1}^{K}$ and the log is applied term-wise.

1. Show that if a function $f : \mathbb{R}^K \to \mathbb{R}$ is convex and $x \in \mathbb{R}^d$, then $\mathbb{R}^{K \times d} \ni \theta \mapsto f(\theta x)$ is convex.

2. Show that the problem boils down to showing that the *log-sum-exp* function $\varphi : a \mapsto \log\left(\sum_{j=1}^{K} e^{a[j]}\right)$ is convex.

3. Using Hölder inequality, prove that φ is convex.

Exercise 6 : Problem: Gaussian kernel on measure spaces (30 pts) The goal of this problem is to provide a condition under which one can define a "Gaussian kernel" on a metric space $(\mathcal{X}, d)$ using the formula $K_t : (x, x') \mapsto e^{-td(x,x')^2}$ for some parameter $t > 0$. We start by introducing some definitions.

- Let $(\mathcal{X}, d)$ be a (non-empty) metric space. We say that it is Hilbertian if there exists a Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $d(x, x') = \|\phi(x) - \phi(x')\|_{\mathcal{H}}$ for all $x, x' \in \mathcal{X}$, where $\|\cdot\|_{\mathcal{H}} \coloneqq \langle \cdot, \cdot \rangle_{\mathcal{H}}^{\frac{1}{2}}$.
- The Gaussian kernel on $(\mathcal{X}, d)$ with parameter $t > 0$ is the map $K_t : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined as

$$K_t(x, x') \coloneqq \exp\left(-td(x, x')^2\right). \tag{7}$$

- A function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be *conditionally negative definite* (CND) if it is symmetric and if, for any $n \in \mathbb{N}$, $a_1, \ldots, a_n \in \mathbb{R}$ such that $\sum_{i=1}^{n} a_i = 0$, and $x_1, \ldots, x_n \in \mathcal{X}$, we have

$$\sum_{1 \le i,j \le n} a_i a_j f(x_i, x_j) \le 0. \tag{8}$$

We also say that $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is *positive definite* (PD) if for any $n \in \mathbb{N}$, $a_1, \ldots, a_n \in \mathbb{R}$, we have

$$\sum_{1 \le i,j \le n} a_i a_j f(x_i, x_j) \ge 0. \tag{9}$$

Recall that if f is PD, then there exists an Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $f(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$ for all $x, y \in \mathcal{X}$.

Note: we omit the terminology "semi" for the sake of conciseness.

The goal of this problem is to prove the following result:

Theorem: Let $(\mathcal{X}, d)$ be a metric space. The Gaussian kernel $K_t : (x, x') \mapsto \exp(-td(x, x')^2)$ defines a PD kernel on $\mathcal{X}$ for all $t > 0$ if and only if $(\mathcal{X}, d)$ is Hilbertian.

Questions:

1. Let $x_0 \in \mathcal{X}$ be any arbitrary point. Define $\varphi(x, x') \coloneqq d(x, x_0)^2 + d(x', x_0)^2 - d(x, x')^2$. Show that φ is positive definite if and only if $(x, x') \mapsto d(x, x')^2$ is conditionally negative definite.

2. - (a) Show that if f_1 and f_2 are PD, so are $f_1 + f_2$ and $f_1 \cdot f_2$.
   - (b) Let $(f_n)_n$ be a sequence of function from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$, all being PD (resp. CND), and converging pointwise toward some limit f (that is, for all x, x', $f_n(x, x') \to f(x, x')$). Show that f is also PD (resp. CND).
   - (c) Deduce that if f is PD, so is the function $(x, x') \mapsto \exp(f(x, x'))$.

3. - (a) Assume that K_t is PD for all $t > 0$. Show that $(x, x') \mapsto d(x, x')^2$ must be CND.

- (b) Conversely, assume that $(x, x') \mapsto d(x, x')^2$ is CND, and deduce that K_t is PD for all $t > 0$.

4. - (a) Assume that $(\mathcal{X}, d)$ is Hilbertian. Show that $(x, x') \mapsto d(x, x')^2$ is CND.
   - (b) Conversely, assume that $(x, x') \mapsto d(x, x')^2$ is CND, and show that $(\mathcal{X}, d)$ is Hilbertian.

     Hint: You may consider the function φ defined above and the associated Reproducing Kernel Hilbert Space.

5. Conclude.