

LINMA2472 Algorithms in Data Science

August 2026

Instructions

1. Write your name, surname, NOMA on each sheet.
2. Write in the frames, not outside (your exam will be scanned and only the text within the frame will appear on-screen for the grader)
3. The exam lasts 3 hours.
4. The back face (verso) of each sheet is not read, you can use it as draft paper.
5. You can use your own blank paper as draft paper.
6. Calculators and other electronic devices are forbidden.
7. You must hand in everything at the end of the exam : questions, answers, drafts.
8. Please hand in the questions in the correct order.

Question 1 : Kernels (cosine kernel)

Suppose that you have N datapoints $x_i \in \mathcal{X}$ with labels $y_i \in \{-1, 1\}$ that you want to classify, where $1 \leq i \leq N$. For the whole exercise, we introduce the following kernel-related definitions :

- $\mathcal{H}$ is an RKHS,
- $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$ is the inner product associated with $\mathcal{H}$,
- $\| \cdot \|_{\mathcal{H}} : \mathcal{H} \rightarrow \mathbb{R}^+$ is the induced norm in $\mathcal{H}$ such that $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}$ for any $f \in \mathcal{H}$,
- $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is the (symmetric positive-definite) kernel,
- $\phi : \mathcal{X} \rightarrow \mathcal{H}$ is a feature map such that $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$ for any $x, y \in \mathcal{X}$,
- $K \in \mathbb{R}^{N \times N}$ is the kernel matrix such that $K_{ij} = k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$.

A) Cosine kernel. Consider the *cosine kernel* :

$$k(x, y) = \cos((x - y)^{\top} t),$$

where $x, y \in \mathbb{R}^n$ and $t \in \mathbb{R}^n$ is a fixed hyperparameter associated with the *direction* and the *frequency* of the kernel.

1. Show that $k(x, y)$ is a symmetric kernel.

Solution: We need to show $k(x, y) = k(y, x)$:

$$k(y, x) = \cos((y - x)^{\top} t) = \cos(-(x - y)^{\top} t) = \cos((x - y)^{\top} t) = k(x, y),$$

since cosine is an even function : $\cos(-\theta) = \cos(\theta)$.

2. Find the expression of a simple feature map $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^h$ such that $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathbb{R}^h}$, where $\langle \cdot, \cdot \rangle_{\mathbb{R}^h}$ is the usual inner product in $\mathbb{R}^h$.
Moreover, give the value of h .

Hint : Recall the trigonometric identity $\cos(a - b) = \cos(a) \cos(b) + \sin(a) \sin(b)$.

Solution: Using the hint with $a = x^{\top} t$ and $b = y^{\top} t$:

$$k(x, y) = \cos(x^{\top} t) \cos(y^{\top} t) + \sin(x^{\top} t) \sin(y^{\top} t) = \langle \phi(x), \phi(y) \rangle_{\mathbb{R}^2},$$

$$\text{with } \phi(x) = \begin{bmatrix} \cos(x^{\top} t) \\ \sin(x^{\top} t) \end{bmatrix} \in \mathbb{R}^2.$$

Therefore, $h = 2$.

3. Using the previous subquestion, show that $k(x, y)$ is a positive-definite kernel.
In other words, show that for any set of M points $x_i \in \mathbb{R}^n$ with $1 \leq i \leq M$ and for any $c \in \mathbb{R}^M$, we have $c^{\top} K c \geq 0$, where $K_{ij} = k(x_i, x_j)$.

Solution: Since $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathbb{R}^2}$, for any $c \in \mathbb{R}^M$:

$$\begin{aligned} c^\top K c &= \sum_{i,j=1}^M c_i c_j K_{ij} = \sum_{i,j=1}^M c_i c_j k(x_i, x_j) \\ &= \sum_{i,j=1}^M c_i c_j \langle \phi(x_i), \phi(x_j) \rangle_{\mathbb{R}^2} \\ &= \left\langle \sum_{i=1}^M c_i \phi(x_i), \sum_{j=1}^M c_j \phi(x_j) \right\rangle_{\mathbb{R}^2} \\ &= \left\| \sum_{i=1}^M c_i \phi(x_i) \right\|_{\mathbb{R}^2}^2 \geq 0. \end{aligned}$$

Therefore, k is positive-definite. Note that this argument holds for any kernel admitting a feature map.

4. Compute $k(x, x)$ for any $x \in \mathbb{R}^n$. What does this imply about where the embedded datapoints $\phi(x)$ live in the feature space $\mathbb{R}^h$?

Solution: $k(x, x) = \cos((x - x)^\top t) = \cos(0) = 1$. Since $k(x, x) = \|\phi(x)\|_{\mathbb{R}^2}^2$, every embedded point satisfies $\|\phi(x)\|_{\mathbb{R}^2} = 1$: all datapoints are mapped onto the unit circle in $\mathbb{R}^2$.

5. Now consider m frequency vectors $t_1, \dots, t_m \in \mathbb{R}^n$ and define the kernel

$$k_m(x, y) = \frac{1}{m} \sum_{j=1}^m \cos((x - y)^\top t_j).$$

Show that $k_m(x, y)$ is a symmetric positive-definite kernel. Moreover, give a feature map $\phi_m : \mathbb{R}^n \rightarrow \mathbb{R}^{h_m}$ such that $k_m(x, y) = \langle \phi_m(x), \phi_m(y) \rangle_{\mathbb{R}^{h_m}}$, together with the value of h_m .

Solution: Symmetry : k_m is a sum of symmetric kernels (by subquestion 1, each term is symmetric), hence $k_m(x, y) = k_m(y, x)$.

Positive-definiteness : let $K^{(j)}$ be the kernel matrix associated with the single-frequency kernel of frequency t_j . For any $c \in \mathbb{R}^M$:

$$c^\top K_m c = \frac{1}{m} \sum_{j=1}^m c^\top K^{(j)} c \geq 0,$$

since each single-frequency cosine kernel is positive-definite (subquestion 3).

Feature map : stack the single-frequency feature maps, scaled by $1/\sqrt{m}$:

$$\phi_m(x) = \frac{1}{\sqrt{m}} \begin{bmatrix} \cos(x^\top t_1) \\ \sin(x^\top t_1) \\ \vdots \\ \cos(x^\top t_m) \\ \sin(x^\top t_m) \end{bmatrix} \in \mathbb{R}^{2m}.$$

Then

$$\begin{aligned} \langle \phi_m(x), \phi_m(y) \rangle_{\mathbb{R}^{2m}} &= \frac{1}{m} \sum_{j=1}^m [\cos(x^\top t_j) \cos(y^\top t_j) + \sin(x^\top t_j) \sin(y^\top t_j)] \\ &= k_m(x, y). \end{aligned}$$

Therefore, $h_m = 2m$.

B) Classification methods with the cosine kernel. For this second part, assume that the data are real vectors in $\mathcal{X} = \mathbb{R}^n$ and that the embedding space $\mathcal{H} = \mathbb{R}^h$ is of finite dimension h , thus it can be understood as the Euclidean space equipped with the usual inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}} = \langle \cdot, \cdot \rangle_{\mathbb{R}^h}$.

Under these assumptions, you choose to use the cosine kernel for your classification task, and you consider two classical classification methods. Both predict the class of a new point $x \in \mathcal{X}$ as $\hat{y}(x) = \text{sign}(d(x))$, where d is the decision function of the method and the sign function is defined as $\text{sign}(u) = -1$ if $u < 0$, $\text{sign}(0) = 0$ and $\text{sign}(u) = 1$ if $u > 0$.

— **Method I : hard-margin kernel support-vector machine (k -SVM).** Its decision function reads

$$d_I(x) = \sum_{i=1}^N \gamma_i^* k(x_i, x) + b^*,$$

where $\gamma^* \in \mathbb{R}^N$ and $b^* \in \mathbb{R}$ are the solutions of the k -SVM optimization problem (we assume that they exist).

— **Method II : (squared) distance to the (lifted) mean.** Its decision function reads

$$d_{II}(x) = \|\phi(x) - \mu_-\|_{\mathcal{H}}^2 - \|\phi(x) - \mu_+\|_{\mathcal{H}}^2,$$

where the centers of each class are defined as

$$\mu_+ = \frac{1}{n_+} \sum_{\substack{i=1 \\ y_i=1}}^N \phi(x_i), \quad \mu_- = \frac{1}{n_-} \sum_{\substack{i=1 \\ y_i=-1}}^N \phi(x_i),$$

and n_+ (resp. n_-) is the number of points labeled $+1$ (resp. -1).

Focus on the decision functions $d_I(x)$ and $d_{II}(x)$ using the cosine kernel.

1. If all datapoints x_i were shifted by the same vector $s \in \mathbb{R}^n$, would the decision functions $d_I(x)$ and $d_{II}(x)$ qualitatively "look the same" (simply be shifted by s)? Justify briefly.

Moreover, give the name of kernels satisfying this property.

Solution: Yes, the decision functions would simply be shifted by s .

The cosine kernel depends only on the difference $x - y$, so $k(x_i + s, x_j + s) = k(x_i, x_j)$: the kernel matrix K is unchanged, and so are the coefficients defining the decision functions (γ^* and b^* for Method I, the means for Method II). Denoting $\tilde{d}$ a decision function trained on the shifted datapoints, every kernel evaluation satisfies $k(x + s, x_i + s) = \cos((x - x_i)^\top t) = k(x, x_i)$, so $\tilde{d}(x + s) = d(x)$ for all x : $\tilde{d}$ is exactly d translated by s .

Kernels depending only on the difference $(x - y)$ are called **translation-invariant** (or **stationary**) kernels.

2. Show that if $k(x, y) = 1$ for some pair of points $x, y \in \mathbb{R}^n$, then x and y are *indistinguishable* for both methods, meaning that $d_I(x) = d_I(y)$ and $d_{II}(x) = d_{II}(y)$, so they always receive the same prediction. Moreover, characterize all such pairs for the cosine kernel, and briefly discuss the practical limitation that this reveals.

Hint : Use the result of subquestion 4 of Part A.

Solution: By subquestion 4 of Part A, $\|\phi(x)\|_{\mathcal{H}} = 1$ for any $x \in \mathbb{R}^n$. Hence, if $k(x, y) = 1$:

$$\begin{aligned}\|\phi(x) - \phi(y)\|_{\mathcal{H}}^2 &= \|\phi(x)\|_{\mathcal{H}}^2 - 2\langle \phi(x), \phi(y) \rangle_{\mathcal{H}} + \|\phi(y)\|_{\mathcal{H}}^2 \\ &= 1 - 2k(x, y) + 1 = 0,\end{aligned}$$

so $\phi(x) = \phi(y)$. Since both decision functions depend on the input only through its embedding $\phi(x)$, it follows that $d_I(x) = d_I(y)$ and $d_{II}(x) = d_{II}(y)$.

Characterization : $\cos((x - y)^\top t) = 1$ if and only if $(x - y)^\top t = 2\pi z$ with $z \in \mathbb{Z}$; in particular, whenever $x - y$ is orthogonal to t .

Limitation : the methods only see the projection $x^\top t$ (modulo 2π); all information orthogonal to t is lost. A remedy is to combine several frequencies, e.g. the kernel k_m of Part A.

Question 2 : Implicit differentiation of a PDE

Consider the one-dimensional Poisson equation

$$-u''(x) = \alpha, \quad x \in (0, 1), \quad u(0) = u(1) = 1,$$

where $\alpha \in \mathbb{R}$ is a parameter. We discretize the interval at $x_i = i/10$, for $i = 0, \dots, 10$, and denote by $\mathbf{u}(\alpha) = (u_1(\alpha), \dots, u_9(\alpha))^\top$ the values at the nine interior points. The standard centered finite-difference discretization gives the implicit equation

$$A\mathbf{u}(\alpha) = h^2\alpha\mathbf{1} + \mathbf{b}, \quad h = \frac{1}{10},$$

where $\mathbf{1} \in \mathbb{R}^9$ is the vector of ones and

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 2 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

The vector $\mathbf{b}$ contains the contribution of the two boundary values. You may assume that A is invertible. Find the derivative of the value at the center of the interval with respect to α , that is, compute $\frac{du_5}{d\alpha}$.

Solution: Let

$$F(\mathbf{u}, \alpha) = A\mathbf{u} - h^2\alpha\mathbf{1} - \mathbf{b}.$$

Since $F(\mathbf{u}(\alpha), \alpha) = \mathbf{0}$, implicit differentiation gives

$$A \frac{d\mathbf{u}}{d\alpha} - h^2\mathbf{1} = \mathbf{0}.$$

Hence the derivative vector $\mathbf{v} = d\mathbf{u}/d\alpha$ is obtained by solving the linear system $A\mathbf{v} = h^2\mathbf{1}$.

For this particular system, it is easy to obtain a closed-form solution, even though this was not required to get the points. The entries

$$v_i = \frac{x_i(1 - x_i)}{2} = \frac{i(10 - i)}{200}, \quad i = 1, \dots, 9,$$

satisfy this linear system. Indeed, $2v_i - v_{i-1} - v_{i+1} = h^2$, with $v_0 = v_{10} = 0$. Therefore

$$\frac{du_5}{d\alpha} = v_5 = \frac{5(10 - 5)}{200} = \frac{1}{8}.$$

Question 3 : Sparse Jacobian recovery

Let $n \geq 3$ and consider the function

$$F : \mathbb{R}^n \rightarrow \mathbb{R}^n, \quad F(x) = \begin{pmatrix} x_1^2 + \sum_{j=2}^n x_j^2 \\ c_2 \exp(x_1) \\ \vdots \\ c_n \exp(x_1) \end{pmatrix},$$

where $c_2, \dots, c_n$ are nonzero constants.

Assume that the sparsity pattern of the Jacobian is known and that $x_i \neq 0$ for every i . A Jacobian-vector product (JVP) with seed v is $J_F(x)v$, and a vector-Jacobian product (VJP) with seed w is $w^\top J_F(x)$. Each product may use an arbitrary seed vector.

1. If only JVPs may be used, what is the minimum number of products required to recover the complete Jacobian? Justify your answer.

Solution: Every pair of columns has a structural nonzero in the first row. Hence no two columns are structurally orthogonal and they must have different colors. Therefore, $\boxed{n}$ JVPs are required.

2. If only VJPs may be used, what is the minimum number of products required to recover the complete Jacobian? Justify your answer.

Solution: Every pair of rows has a structural nonzero in the first column. Hence no two rows are structurally orthogonal and they must have different colors. Therefore, $\boxed{n}$ VJPs are required.

3. Can fewer Jacobian products be used if JVPs and VJP are combined? If so, give suitable seed vectors and explain how the complete Jacobian is recovered.

Solution: Yes. Let $e_1 = (1, 0, \dots, 0)^\top$. One JVP gives

$$J_F(x)e_1 = J_F(x)_{:,1},$$

which is the complete first column. One VJP gives

$$e_1^\top J_F(x) = J_F(x)_{1,:},$$

which is the complete first row. All entries outside this row and column are known to be zero, so these two products recover the whole Jacobian :

$$J_F(x) = \begin{pmatrix} 2x_1 & 2x_2 & \cdots & 2x_n \\ c_2 \exp(x_1) & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ c_n \exp(x_1) & 0 & \cdots & 0 \end{pmatrix}.$$

Thus only $\boxed{2}$ Jacobian products are needed : one JVP and one VJP.

One product cannot suffice : a single JVP does not distinguish the $n - 1$ entries in the first row, while a single VJP does not distinguish the $n - 1$ entries in the first column.

Question 4 : Automatic differentiation

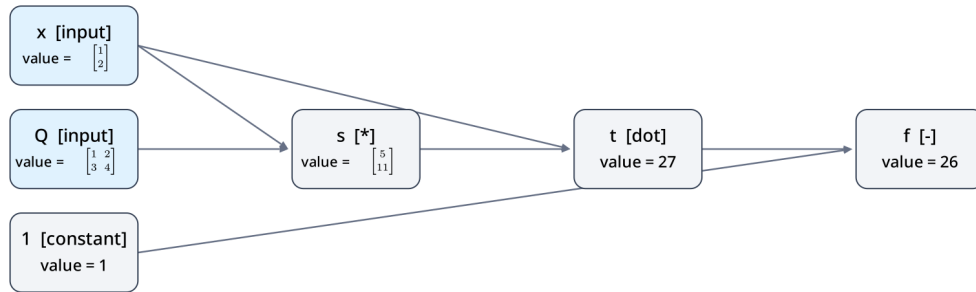
Let

$$f(x, Q) = x^\top Q x - 1, \quad x = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad Q = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix},$$

where Q is not assumed to be symmetric. We evaluate this function as

$$s = Qx, \quad t = x^\top s, \quad f = t - 1.$$

The corresponding expression graph and forward values are shown below. An arrow points from each argument to the operation that uses it.



- Starting with $\bar{f} = \partial f / \partial f = 1$, perform the reverse pass. Give the order in which you process the operation nodes and show how the adjoint of x accumulates its two contributions.

Solution: A valid reverse topological order is f , t , then s . From $f = t - 1$, we obtain

$$\bar{t} += \bar{f} = 1.$$

From $t = x^\top s$, we obtain

$$\bar{x} += \bar{t} s = \begin{bmatrix} 5 \\ 11 \end{bmatrix}, \quad \bar{s} += \bar{t} x = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

Finally, from $s = Qx$,

$$\bar{Q} += \bar{s} x^\top = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}, \quad \bar{x} += Q^\top \bar{s} = \begin{bmatrix} 7 \\ 10 \end{bmatrix}.$$

Hence the two contributions to $\bar{x}$ accumulate to

$$\bar{x} = \begin{bmatrix} 5 \\ 11 \end{bmatrix} + \begin{bmatrix} 7 \\ 10 \end{bmatrix} = \begin{bmatrix} 12 \\ 21 \end{bmatrix}.$$

- Give $\nabla_x f$ and $\nabla_Q f$, first as formulas valid for arbitrary x and Q , and then at the values above. Why does $\nabla_Q f$ have rank one whenever $x \neq 0$?

Solution: The input adjoints give

$$\nabla_x f = (Q + Q^\top)x, \quad \nabla_Q f = xx^\top.$$

At the specified values,

$$\nabla_x f = \begin{bmatrix} 12 \\ 21 \end{bmatrix}, \quad \nabla_Q f = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}.$$

For every vector v , $(xx^\top)v = x(x^\top v)$ belongs to $\text{span}(x)$. Thus $xx^\top$ has rank at most one, and its rank is exactly one when $x \neq 0$.

- More generally, suppose that a scalar-valued function has many inputs. To compute its full gradient, which mode is usually preferable : forward or reverse ? Briefly justify your answer.

Solution: Reverse mode is usually preferable. A single reverse pass computes the derivative of the scalar output with respect to every input, whereas forward mode generally requires one pass per input direction.