Theoretical Exercises: Computational Foundations of Data Sciences

January 13, 2026

Supervised vs. Unsupervised Learning

- **Supervised Learning:** In this paradigm, the model is provided with a labeled dataset $(x_i, y_i)_{i=1}^n$. The objective is to learn a mapping $F : \mathcal{X} \to \mathcal{Y}$ that can predict labels for unseen data .
- **Unsupervised Learning:** The model receives only observations $(x_i)_{i=1}^n$ without target labels. The goal is to discover intrinsic structures, such as clusters or low-dimensional manifolds, within the data.
- **Pros and Cons:**
  - **Supervised:** Offers high precision and clear evaluation metrics but requires expensive, manually labeled data .
  - **Unsupervised:** Can scale easily to massive raw datasets and find hidden patterns, but results can be subjective and harder to evaluate.

The Overfitting Phenomenon

- **Definition:** Overfitting occurs when a model learns the "noise" or specific fluctuations in the training data rather than the underlying general distribution . This results in excellent performance on training data but poor generalization to new data .
- **Mitigation:** Overfitting can be mitigated by:
  - Increasing the size of the dataset .
  - Using regularization techniques, such as Ridge (ℓ_2) or Lasso (ℓ_1) penalties, to restrict model complexity.
  - Simplifying the model architecture (reducing parameters) .

Exercise 4: Linear Regression

Let $(x_i, y_i)_{i=1}^n$ be couples of observations and labels in $\mathbb{R}^d \times \mathbb{R}^k$. We consider the linear model with parameter $M \in \mathbb{R}^{d\times k}$ defined as $F_M(x) = xM$ for $x \in \mathbb{R}^d$. We aim to minimize the objective function:

$$M \mapsto \sum_{i=1}^{n} \|F_M(x_i) - y_i\|^2 \tag{1}$$

where $\|\cdot\|$ denotes a loss function, specifically the squared Frobenius norm in this context.

1. Encompassing Affine and Polynomial Regression

Linear regression models are "linear" with respect to the parameters M, not necessarily the input features x.

- **Affine Regression ($x \mapsto ax+b$):** By augmenting the input vector x with a constant 1 (creating $\tilde{x} = [x, 1]$), the bias term b becomes a parameter within the matrix M, making the model $F_M(\tilde{x}) = \tilde{x}M$ equivalent to an affine map.
- **Polynomial Regression ($x \mapsto \sum a_j x^j$):** By transforming the input x into a feature vector of powers $\Phi(x) = [1, x, x^2, \ldots, x^d]$, the prediction becomes a linear combination of these powers, which is still a linear model in terms of the coefficients a_j.

2. Matrix Formulation

Let $X \in \mathbb{R}^{n\times d}$ be the matrix where each row i is the observation x_i, and $Y \in \mathbb{R}^{n\times k}$ be the matrix where each row i is the label y_i. Using the definition of the squared Frobenius norm $\|A\|_F^2 = \mathrm{Tr}(AA^T)$, the objective function is equivalent to :

$$f(M) = \|XM - Y\|_F^2 \tag{2}$$

3. Optimal Solution for M

To find the minimizer, we expand the squared norm:

$$\begin{aligned} f(M) &= \mathrm{Tr}((XM - Y)^T (XM - Y)) \\ &= \mathrm{Tr}(M^T X^T XM - M^T X^T Y - Y^T XM + Y^T Y) \end{aligned}$$

Taking the gradient with respect to M and setting it to zero :

$$\nabla_M f(M) = 2X^T XM - 2X^T Y = 0$$

Assuming $(X^T X)$ is invertible, we obtain the closed-form solution:

$$M^* = (X^T X)^{-1} X^T Y \tag{3}$$

4. Interpretation of the Invertibility Assumption

The assumption that $(X^T X)$ is non-singular implies that X has full column rank d. In practice, this means:

- The number of observations n must be greater than or equal to the number of features d
- There is no perfect multicollinearity among the features (i.e., no feature is a linear combination of others)

5. Singular Case and Pseudo-Inverse

When $(X^T X)$ is singular (e.g., when $n < d$), we use the Moore-Penrose pseudo-inverse $X^\dagger$. The optimal M can still be expressed as :

$$M = X^\dagger Y \tag{4}$$

This solution provides the minimizer with the smallest Frobenius norm $\|M\|_F^2$, effectively providing the unique minimum-norm solution for an underdetermined system.

Exercise 6

Let $L : \mathbb{R}^d \to \mathbb{R}$ be a real-valued function, and assume that it is of class C^1. Explain the claim

"$-\nabla L(\theta)$ indicates the steepest descent direction of L at θ."

Let $L : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function and fix a point $\theta \in \mathbb{R}^d$. We want to understand in which direction u (with $\|u\| = 1$) the function L decreases the fastest when moving away from θ.

For a small step size $\varepsilon > 0$, a first–order Taylor expansion gives

$$L(\theta + \varepsilon u) = L(\theta) + \varepsilon \langle \nabla L(\theta), u \rangle + o(\varepsilon).$$

Therefore, the variation of L in direction u is, at first order,

$$\frac{L(\theta + \varepsilon u) - L(\theta)}{\varepsilon} \approx \langle \nabla L(\theta), u \rangle.$$

This quantity is called the *directional derivative* of L at θ in direction u.

To find the direction of steepest descent, we want to minimize $\langle \nabla L(\theta), u \rangle$ over all unit vectors $\|u\| = 1$. By the Cauchy–Schwarz inequality,

$$\langle \nabla L(\theta), u \rangle \geq -\|\nabla L(\theta)\| \, \|u\| = -\|\nabla L(\theta)\|.$$

Equality is achieved when

$$u = -\frac{\nabla L(\theta)}{\|\nabla L(\theta)\|}.$$

Hence, among all directions with unit norm, the one that makes L decrease the fastest is exactly the direction opposite to the gradient.

Therefore, the vector $-\nabla L(\theta)$ gives the direction of *steepest descent*: moving a little in this direction produces the largest possible decrease of L at first order.

Exercise 7

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, assumed to be of class C^2.

1. First order inequality.
Since f is convex, for all $x, y \in \mathbb{R}^d$ and all $t \in [0, 1]$,

$$f\big(y + t(x - y)\big) \le (1 - t)f(y) + tf(x).$$

Define $\varphi(t) = f(y + t(x - y))$. Then φ is a convex function on $[0, 1]$. Convexity implies

$$\varphi(1) \ge \varphi(0) + \varphi'(0).$$

But

$$\varphi(1) = f(x), \qquad \varphi(0) = f(y), \qquad \varphi'(0) = \langle \nabla f(y), x - y \rangle.$$

Therefore,

$$f(x) \ge f(y) + \langle \nabla f(y), x - y \rangle,$$

which proves (4).

2. Monotonicity of the gradient.
Applying (4) with (x, y) and then with (y, x), we get

$$f(x) \ge f(y) + \langle \nabla f(y), x - y \rangle,$$

$$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle.$$

Adding the two inequalities yields

$$0 \ge \langle \nabla f(y), x - y \rangle + \langle \nabla f(x), y - x \rangle = -\langle \nabla f(x) - \nabla f(y), x - y \rangle.$$

Hence

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \ge 0,$$

which proves (5).

3. Characterization of minimizers.
Assume that $\nabla f(x) = 0$. Applying (4) with $y = x$, we get for all z,

$$f(z) \ge f(x) + \langle \nabla f(x), z - x \rangle = f(x).$$

Thus x is a global minimizer of f.

4. Positivity of the Hessian.

Fix $x \in \mathbb{R}^d$ and $v \in \mathbb{R}^d$. Define

$$\psi(t) = f(x + tv).$$

Since f is convex, ψ is a convex function of $t \in \mathbb{R}$. Therefore, $\psi''(0) \geq 0$.

By the chain rule,

$$\psi'(t) = \langle \nabla f(x + tv), v \rangle, \qquad \psi''(t) = v^\top \nabla^2 f(x + tv)\, v.$$

Hence

$$v^\top \nabla^2 f(x)\, v = \psi''(0) \geq 0.$$

This holds for all v, so the Hessian matrix $\nabla^2 f(x)$ is positive semi-definite.

Exercise 8

Let $f : \mathbb{R}^d \to \mathbb{R}$ be an α-strongly convex function with $\alpha > 0$, assumed to be of class C^2.

1. Proof of inequality (6).
By definition, f is α-strongly convex if the function

$$x \mapsto f(x) - \frac{\alpha}{2}\|x\|^2$$

is convex. Applying the first-order convexity inequality (Exercise 7) to this function, we obtain for all x, y,

$$f(y) - \frac{\alpha}{2}\|y\|^2 \geq f(x) - \frac{\alpha}{2}\|x\|^2 + \langle \nabla f(x) - \alpha x,\, y - x \rangle.$$

Rearranging terms gives

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|^2,$$

which is exactly (6).

2. Convexity of $x \mapsto f(x) - \frac{\alpha}{2}\|x\|^2$.
This is simply the definition of α-strong convexity: subtracting $\frac{\alpha}{2}\|x\|^2$ removes the quadratic curvature α, leaving a convex function.

3. Eigenvalues of the Hessian.
From part 2, the Hessian of $f(x) - \frac{\alpha}{2}\|x\|^2$ is positive semi-definite, hence

$$\nabla^2 f(x) - \alpha I \succeq 0.$$

Therefore all eigenvalues of $\nabla^2 f(x)$ are at least α.

4. Uniqueness of the minimizer.
Strong convexity implies strict convexity, so f can have at most one global minimizer. Since f is continuous and coercive, it admits a minimizer x^*, which must be unique.

5. PL-inequality.

Let x^* be the unique minimizer of f, so $\nabla f(x^*) = 0$. Applying inequality (6) with $y = x^*$, we get

$$f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x \rangle + \frac{\alpha}{2} \|x^* - x\|^2.$$

Rewriting,

$$f(x) - f(x^*) \leq \langle \nabla f(x), x - x^* \rangle - \frac{\alpha}{2} \|x - x^*\|^2.$$

Using Young's inequality,

$$\langle a, b \rangle \leq \frac{1}{2\alpha} \|a\|^2 + \frac{\alpha}{2} \|b\|^2,$$

with $a = \nabla f(x)$ and $b = x - x^*$, we obtain

$$f(x) - f(x^*) \leq \frac{1}{2\alpha} \|\nabla f(x)\|^2.$$

This proves the PL-inequality (7).

Exercise 9 (Convergence rate of the gradient descent).

Let $f : \mathbb{R}^d \to \mathbb{R}$ be an α-strongly convex and β-smooth function, and let x^* be its unique minimizer. We consider the gradient descent iteration

$$x_{t+1} = x_t - \lambda \nabla f(x_t),$$

where $\lambda > 0$.

1. Descent lemma.

Since f is β-smooth, for all x, y,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2} \|y - x\|^2.$$

Apply this with $x = x_t$ and $y = x_{t+1} = x_t - \lambda \nabla f(x_t)$:

$$f(x_{t+1}) \leq f(x_t) - \lambda \|\nabla f(x_t)\|^2 + \frac{\beta}{2} \lambda^2 \|\nabla f(x_t)\|^2.$$

Hence

$$f(x_{t+1}) - f(x_t) \leq -\lambda \Big(1 - \frac{\beta\lambda}{2}\Big) \|\nabla f(x_t)\|^2.$$

2. Linear decrease of the objective.

By the PL-inequality (Exercise 8),

$$\|\nabla f(x_t)\|^2 \geq 2\alpha \big(f(x_t) - f(x^*)\big).$$

Combining with the previous inequality gives

$$f(x_{t+1}) - f(x^*) \le \Big(1 - 2\alpha\lambda\Big(1 - \frac{\beta\lambda}{2}\Big)\Big)\big(f(x_t) - f(x^*)\big).$$

To ensure decrease, we need

$$1 - \frac{\beta\lambda}{2} > 0 \quad \iff \quad \lambda < \frac{2}{\beta}.$$

3. Exponential convergence.
If $0 < \lambda < \frac{2}{\beta}$, then

$$\rho := 1 - 2\alpha\lambda\Big(1 - \frac{\beta\lambda}{2}\Big) \in (0,1),$$

and we obtain the geometric decay

$$f(x_t) - f(x^*) \le \rho^t\big(f(x_0) - f(x^*)\big).$$

Thus the sequence $(f(x_t))_t$ converges exponentially fast to the minimal value $f(x^*)$.

Exercise 11 (Convexity of the logistic regression).

We consider data $(x_i, y_i)_{i=1}^n$ with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}^K$ one–hot encoded vectors. Define

$$L(\theta) = -\sum_{i=1}^n y_i \cdot \log(\mathrm{smax}(\theta x_i)), \qquad \mathrm{smax}(a)_j = \frac{e^{a_j}}{\sum_{\ell=1}^K e^{a_\ell}}.$$

1. Composition with a linear map.
Let $f : \mathbb{R}^K \to \mathbb{R}$ be convex and $x \in \mathbb{R}^d$. For any θ_1, θ_2 and $t \in [0,1]$,

$$f\big((t\theta_1 + (1-t)\theta_2)x\big) = f\big(t(\theta_1 x) + (1-t)(\theta_2 x)\big) \le t f(\theta_1 x) + (1-t) f(\theta_2 x),$$

so $\theta \mapsto f(\theta x)$ is convex.

2. Reduction to the log-sum-exp.
Since

$$-\, y_i \cdot \log(\mathrm{smax}(a)) = \log\Big(\sum_{j=1}^K e^{a_j}\Big) - a_{k(i)},$$

(where $k(i)$ is the index of the nonzero entry of y_i), the loss is a sum of affine functions and the function

$$\varphi(a) = \log\Big(\sum_{j=1}^K e^{a_j}\Big).$$

Hence it suffices to prove that φ is convex.

3. Convexity of φ via Hölder.
For $a, b \in \mathbb{R}^K$ and $t \in [0, 1]$,

$$\sum_j e^{ta_j+(1-t)b_j} = \sum_j (e^{a_j})^t (e^{b_j})^{1-t} \leq \Big(\sum_j e^{a_j}\Big)^t \Big(\sum_j e^{b_j}\Big)^{1-t},$$

by Hölder's inequality. Taking logarithms gives

$$\varphi(ta + (1-t)b) \leq t\varphi(a) + (1-t)\varphi(b),$$

so φ is convex.

4. Hessian proof.
We have

$$\partial_i \varphi(a) = \frac{e^{a_i}}{\sum_j e^{a_j}} = p_i,$$

and

$$\partial^2_{ij} \varphi(a) = p_i(\delta_{ij} - p_j).$$

Thus the Hessian matrix is

$$H = \operatorname{diag}(p) - pp^\top.$$

For any $v \in \mathbb{R}^K$,

$$v^\top H v = \sum_i p_i v_i^2 - \Big(\sum_i p_i v_i\Big)^2 = \operatorname{Var}_p(v) \geq 0,$$

so H is positive semi-definite and φ is convex.

Therefore $L(\theta)$ is convex as a sum of convex functions of θ.

*

Exercise 13 (Quantization problem) — Detailed solution

Let P be a probability distribution on $\mathbb{R}^d$ with finite second moment, i.e. $P \in \mathcal{P}_2(\mathbb{R}^d)$, and let $n \in \mathbb{N}$. Recall that an n-quantizer is a measurable map $f : \mathbb{R}^d \to \mathbb{R}^d$ whose image contains at most n distinct points. Denote by $\mathcal{F}_n$ the set of all n-quantizers. For $f \in \mathcal{F}_n$ define the energy

$$E_P(f) := \mathbb{E}_P\big[\, |X - f(X)|^2 \,\big] = \int_{\mathbb{R}^d} |x - f(x)|^2 \, dP(x).$$

Finally set

$$V_n(P) := \inf_{f \in \mathcal{F}_n} E_P(f).$$

For $(a_1, \ldots, a_n) \in (\mathbb{R}^d)^n$ denote

$$\psi_P(a_1, \ldots, a_n) := \mathbb{E}_P\Big[\min_{1 \leq i \leq n} |X - a_i|^2\Big].$$

1. For all $f \in \mathcal{F}_n$, $E_P(f) < \infty$.

Proof. If $f \in \mathcal{F}_n$ then the image of f is a finite set $\{a_1, \dots, a_m\}$ with $m \le n$. For every $x \in \mathbb{R}^d$,

$$|x - f(x)|^2 \le 2|x|^2 + 2 \max_{1 \le i \le m} |a_i|^2,$$

by $(u - v)^2 \le 2u^2 + 2v^2$. Taking expectation under P and using $\mathbb{E}_P[|X|^2] < \infty$ together with the fact that the finite constant $\max_i |a_i|^2$ is integrable, we obtain $E_P(f) < \infty$. $\square$

2. Proof of the identity (9)

$$V_n(P) = \inf_{\alpha \subset \mathbb{R}^d,\, |\alpha| \le n} \mathbb{E}_P \Big[\min_{a \in \alpha} |X - a|^2 \Big].$$

Proof. Given any $f \in \mathcal{F}_n$, let $\alpha = f(\mathbb{R}^d)$ be its image. Then $|\alpha| \le n$ and for every x we have $f(x) \in \alpha$, hence $|x - f(x)|^2 \ge \min_{a \in \alpha} |x - a|^2$, with equality if $f(x)$ is chosen as a (measurable) minimizer of $a \mapsto |x - a|^2$ over α. Thus

$$E_P(f) \ge \mathbb{E}_P \Big[\min_{a \in \alpha} |X - a|^2 \Big].$$

Taking the infimum over all $f \in \mathcal{F}_n$ shows $V_n(P) \ge \inf_{|\alpha| \le n} \mathbb{E}[\min_{a \in \alpha} |X - a|^2]$.

Conversely, given any finite set $\alpha = \{a_1, \dots, a_m\}$ with $m \le n$, one can define a measurable quantizer f that maps each x to a (measurable) index $a_{i(x)}$ achieving the minimum $\min_{1 \le i \le m} |x - a_i|^2$ (e.g. choose the smallest index achieving the minimum). Then $E_P(f) = \mathbb{E}[\min_{a \in \alpha} |X - a|^2]$. Taking infimum over such α yields the reverse inequality. Combining both directions proves (9). $\square$

3. Relation between minimizers of ψ_P and optimal quantizers.

Proof. If $(a_1, \dots, a_n)$ minimizes ψ_P, define the quantizer f by mapping x to a nearest center among the a_i's (breaking ties measurably). Then $f \in \mathcal{F}_n$ and $E_P(f) = \psi_P(a_1, \dots, a_n)$, so f is optimal for the quantization problem. Conversely, if $f \in \mathcal{F}_n$ is optimal, let $\alpha = f(\mathbb{R}^d) = \{a_1, \dots, a_m\}$. Then $\psi_P(a_1, \dots, a_m) = E_P(f)$ and extending this m-tuple with repetition (to get n centers) does not increase ψ_P. Thus minimizing ψ_P and minimizing E_P are equivalent; there is a natural one-to-(many) correspondence: each optimal n-tuple of centers yields an optimal quantizer (via Voronoi tie-breaking), and each optimal quantizer has image an optimal set of centers. $\square$

4. Is ψ_P convex in general?

No. The function $\psi_P : (\mathbb{R}^d)^n \to \mathbb{R}$ is in general *not convex.* Intuitively this is because the operation "take the minimum over i" is non-smooth and non-convex. A simple 1D counterexample: take P supported equally on two points $x = -1$ and $x = +1$ and $n = 2$. Consider two center configurations $a = (-2, 0)$ and $b = (0, 2)$. The average of these two configurations is $((-2 + 0)/2, (0 + 2)/2) =$

$(-1,1)$. One checks that $\psi_P(a) = \psi_P(b) = 0$ (each configuration can place a center at -1 and $+1$), but $\psi_P((-1,1)) > 0$. Hence ψ_P is not convex.

5. Continuity of ψ_P.

Proof. Fix $(a_1, \ldots, a_n)$. For each $x \in \mathbb{R}^d$, the map $(a_1, \ldots, a_n) \mapsto \min_{1 \le i \le n} |x - a_i|^2$ is continuous (minimum of finitely many continuous functions). By the growth bound

$$\min_i |x - a_i|^2 \le 2|x|^2 + 2 \max_i |a_i|^2,$$

and since $\mathbb{E}[|X|^2] < \infty$, the dominated convergence theorem applies. Therefore ψ_P is continuous on $(\mathbb{R}^d)^n$. □

6. Strict decrease $V_n(P) < V_{n-1}(P)$.

Proof. Since P is assumed to be supported on at least n distinct points, one can place n centers exactly on n distinct support points to obtain zero error on those points; intuitively allowing one extra center cannot increase the optimal value, so $V_n(P) \le V_{n-1}(P)$. We must show strict inequality.

Assume by contradiction that $V_n(P) = V_{n-1}(P)$. Then an optimal choice of n centers achieves the same error as the best choice of $n-1$ centers. But then one of the n centers must be superfluous: its removal yields a configuration of $n-1$ centers with the same energy. Repeating this argument would show that $V_1(P) = V_{n-1}(P)$, which is impossible when the support of P contains at least n distinct points (since one center cannot exactly represent n distinct points with strictly smaller squared distance sum than using n centers). More concretely, there exists a measurable set A of positive P-mass containing at least two distinct support points that are represented by different centers in the n-center optimal configuration; merging those two centers increases the error, so equality cannot hold. Hence $V_n(P) < V_{n-1}(P)$. □

7. (Hard) For $c \in (0, V_{n-1}(P))$ the sublevel set $\{(a_1, \ldots, a_n) : \psi_P(a_1, \ldots, a_n) \le c\}$ is compact.

Proof. We prove compactness by showing that the sublevel set $S_c := \{a \in (\mathbb{R}^d)^n : \psi_P(a) \le c\}$ is closed and bounded. Closedness follows from continuity of ψ_P (proved above).

It remains to show boundedness. Suppose by contradiction that there exists a sequence $(a^{(m)})_{m \ge 1} \subset S_c$ with $\|a^{(m)}\| \to \infty$ as $m \to \infty$ (where $\|a\| := \max_{1 \le i \le n} |a_i|$). We consider two cases.

Case 1: For some subsequence, all n centers go to infinity, i.e. for this subsequence $\min_i |a_i^{(m)}| \to \infty$. Fix $R > 0$. For m large enough, every center $a_i^{(m)}$ lies outside the ball $B(0, R)$. Then for any $x \in B(0, R/2)$ and any i,

$$|x - a_i^{(m)}| \ge |a_i^{(m)}| - |x| \ge R - \tfrac{R}{2} = \tfrac{R}{2},$$

hence $\min_i |x - a_i^{(m)}|^2 \geq (R/2)^2$. Therefore

$$\psi_P(a^{(m)}) \geq P\big(B(0, R/2)\big) \cdot (R/2)^2.$$

Choose R so large that $P(B(0, R/2)) \cdot (R/2)^2 > c$ (possible because $\lim_{R\to\infty} P(B(0, R/2)) = 1$ and $(R/2)^2 \to \infty$). This contradicts $\psi_P(a^{(m)}) \leq c$ for large m. Thus Case 1 is impossible.

Case 2: There exists an index j and a subsequence (still denoted m) such that $|a_j^{(m)}| \to \infty$ while the remaining $n-1$ centers stay bounded. By compactness of closed balls in $\mathbb{R}^d$, we can extract a further subsequence so that the remaining $n-1$ centers converge to some $(b_1, \ldots, b_{n-1})$. For a fixed $x \in \mathbb{R}^d$, for large m we have $|x - a_j^{(m)}|^2$ arbitrarily large, hence $\min_{1\leq i\leq n} |x - a_i^{(m)}|^2 = \min_{1\leq i\leq n-1} |x - a_i^{(m)}|^2$ eventually. Because the other centers converge, the pointwise limit of the integrands is $\min_{1\leq i\leq n-1} |x - b_i|^2$. By the dominated convergence theorem (using the uniform bound $\min_i |x - a_i^{(m)}|^2 \leq 2|x|^2 + 2\max_i |a_i^{(m)}|^2$ and the fact that the bounded centers keep the dominating constants integrable), we get

$$\lim_{m\to\infty} \psi_P(a^{(m)}) = \psi_P(b_1, \ldots, b_{n-1}, \infty) = \mathbb{E}_P\Big[\min_{1\leq i\leq n-1} |X - b_i|^2\Big] \geq V_{n-1}(P).$$

Since each $\psi_P(a^{(m)}) \leq c$ and $c < V_{n-1}(P)$, this is a contradiction. Hence Case 2 cannot occur either.

Thus every sequence in S_c is bounded, so S_c is bounded. Being closed and bounded in $(\mathbb{R}^d)^n \cong \mathbb{R}^{nd}$, S_c is compact. □

8. Deduce that the set of n-optimal quantizers of P is non-empty.

Proof. Let $c_* := \inf_{a\in(\mathbb{R}^d)^n} \psi_P(a) = V_n(P)$. Choose a sequence $(a^{(m)})$ such that $\psi_P(a^{(m)}) \downarrow c_*$. Pick any c with $0 < c - c_* < V_{n-1}(P) - c_*$ (possible because $c_* < V_{n-1}(P)$ by part 6). For large m we have $\psi_P(a^{(m)}) \leq c$, so all but finitely many $a^{(m)}$ belong to the compact sublevel set S_c. Extract a convergent subsequence $a^{(m_k)} \to a^*$. By continuity of ψ_P, $\psi_P(a^*) = \lim_k \psi_P(a^{(m_k)}) = c_* = V_n(P)$. Hence a^* is a minimizer of ψ_P. Mapping a^* to the corresponding quantizer (via nearest-center rule) yields an optimal quantizer. Therefore the set of n-optimal quantizers is non-empty. □

9. Finite sample case: $P = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$. Is quantization a supervised or unsupervised problem?

It is an *unsupervised* learning problem. There are no target labels to predict; the goal is to approximate the distribution (or the dataset) by a finite set of representative points (centers) minimizing a reconstruction error. In particular the k-means problem is the empirical quantization for the empirical measure above.

10. Observe that the k-means problem is a quantization problem.

Proof. For the empirical measure $P = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}$, the quantity $\psi_P(a_1, \dots, a_n) = \frac{1}{N}\sum_{i=1}^{N}\min_{1\le j\le n}|x_i - a_j|^2$ is exactly the k-means objective (up to the factor $1/N$). Thus minimizing the k-means objective over centers is equivalent to finding minimizers of ψ_P, i.e. empirical optimal quantizers. $\square$

11. (Bonus in statement) Relation with Wasserstein distance:

$$V_n(P) = \inf_{f\in\mathcal{F}_n} W_2(P, f\#P)^2,$$

where W_2 denotes the 2-Wasserstein distance and $f\#P$ is the pushforward of P by f.

Sketch of proof. For a measurable map f, consider the coupling π of P and $f\#P$ given by the pushforward of P under the map $x \mapsto (x, f(x))$. Then

$$\int |x-y|^2\, d\pi(x,y) = \mathbb{E}_P[|X - f(X)|^2] = E_P(f),$$

so $W_2(P, f\#P)^2 \le E_P(f)$. Conversely, any coupling π between P and a discrete measure supported on at most n points induces a measurable map f (by disintegrating π) with $f\#P$ equal to that discrete measure and with $E_P(f)$ equal to the transport cost of π. Taking infimum over π and over f gives the equality. A fully rigorous proof requires disintegration/selection arguments but the identity expresses that approximating P by pushforwards with at most n atoms in Wasserstein-2 sense is equivalent to minimizing the mean squared reconstruction error. $\square$

Using the definition of the W_2 distance

$$W_2(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)}\left(\int \|x-y\|^2 d\pi(x,y)\right)^{1/2} \quad [?]$$

The quantization energy can be rewritten as

$$V_n(P) = \inf_{f\in\mathcal{F}_n} W_2(P, f\#P)^2$$

This shows that quantization is the problem of finding the best discrete approximation of a measure P in the Wasserstein-2 metric.

Exercise 16 (Linear Autoencoder).

We consider a linear autoencoder with encoder $x \mapsto Ax$ and decoder $z \mapsto Bz$, where $A \in \mathbb{R}^{k\times d}$ and $B \in \mathbb{R}^{d\times k}$. Given a data matrix $X \in \mathbb{R}^{d\times n}$, training consists in minimizing

$$\mathcal{L}(A,B) = \|BAX - X\|_F^2.$$

1. Optimal A for fixed B.
Assume that $XX^\top$ and $B^\top B$ are invertible. We minimize

$$\|BAX - X\|_F^2 = \mathrm{Tr}\big((BAX - X)(BAX - X)^\top\big).$$

Differentiating with respect to A and setting the gradient to zero gives

$$B^\top(BAX - X)X^\top = 0.$$

Hence

$$B^\top B\,A\,XX^\top = B^\top XX^\top.$$

Since $B^\top B$ and $XX^\top$ are invertible,

$$A = (B^\top B)^{-1}B^\top.$$

This is the Moore–Penrose pseudo-inverse of B.

2. Relation with PCA.
Plugging $A = (B^\top B)^{-1}B^\top$ into the loss gives

$$\|B(B^\top B)^{-1}B^\top X - X\|_F^2.$$

The matrix $P_B := B(B^\top B)^{-1}B^\top$ is the orthogonal projector onto $\mathrm{Im}(B)$. Hence training the autoencoder amounts to choosing a k-dimensional subspace $\mathrm{Im}(B)$ that best approximates the columns of X in least–squares sense. This is exactly the PCA problem: the optimal $\mathrm{Im}(B)$ is spanned by the k leading eigenvectors of the covariance matrix $XX^\top$.

3. Affine autoencoder.
If we allow affine maps $x \mapsto Ax + a$ and $z \mapsto Bz + b$, the bias terms a, b allow the model to fit the mean of the data. Optimizing over a, b leads to centering the data matrix X. After centering, the optimal linear part again corresponds to PCA. Thus affine autoencoders perform PCA on centered data.

Exercise 17 (Examples of kernels).

Let $X = \mathbb{R}^d$.

1. The linear kernel.
We define

$$K(x, y) = \langle x, y\rangle.$$

(a) K is PSD. For any points $x_1, \dots, x_n \in \mathbb{R}^d$ and any coefficients $c_1, \dots, c_n \in \mathbb{R}$,

$$\sum_{i,j=1}^n c_i c_j K(x_i, x_j) = \sum_{i,j} c_i c_j \langle x_i, x_j\rangle = \Big\langle \sum_i c_i x_i, \sum_j c_j x_j \Big\rangle = \Big\| \sum_i c_i x_i \Big\|^2 \geq 0.$$

Hence K is positive semidefinite.

(b) RKHS associated to K. The feature map is $\phi(x) = x$. The corresponding RKHS is $\mathcal{H} = \mathbb{R}^d$ equipped with the usual inner product $\langle \cdot, \cdot \rangle$. Indeed,

$$K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}.$$

2. The quadratic kernel.
We define

$$K(x, y) = \langle x, y \rangle^2.$$

(a) K is PSD. Let $x, y \in \mathbb{R}^d$. Then

$$\langle x, y \rangle^2 = \langle x \otimes x,\, y \otimes y \rangle,$$

where $x \otimes x \in \mathbb{R}^{d \times d}$ is the rank–one matrix with entries $(x \otimes x)_{ij} = x_i x_j$, and the inner product is the Frobenius product $\langle A, B \rangle = \mathrm{Tr}(A^\top B)$. Hence the feature map $\phi(x) = x \otimes x$ satisfies

$$K(x, y) = \langle \phi(x), \phi(y) \rangle,$$

so K is a PSD kernel.

(b) RKHS associated to K. The RKHS is the space of symmetric matrices generated by rank–one tensors $x \otimes x$, equipped with the Frobenius inner product. Functions in this RKHS are quadratic forms $f(x) = x^\top M x$.

Exercise 19 (Kernel and Fourier transform).

Let $K(x, y) = h(x - y)$ with $h : \mathbb{R}^d \to \mathbb{R}$ satisfying $h(u) = h(-u)$. Assume that the Fourier transform

$$\widehat{h}(\omega) = \int_{\mathbb{R}^d} e^{-i\langle \omega, u \rangle} h(u)\, du$$

is non–negative.

1. K is a PSD kernel.
Take any $x_1, \ldots, x_n \in \mathbb{R}^d$ and $c_1, \ldots, c_n \in \mathbb{R}$. We compute

$$\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) = \sum_{i,j} c_i c_j h(x_i - x_j).$$

Using the inverse Fourier transform,

$$h(u) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{i\langle \omega, u \rangle} \widehat{h}(\omega)\, d\omega,$$

we get

$$\sum_{i,j} c_i c_j h(x_i - x_j) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \sum_{i,j} c_i c_j e^{i\langle \omega, x_i - x_j \rangle} \widehat{h}(\omega)\, d\omega.$$

But

$$\sum_{i,j} c_i c_j e^{i\langle \omega, x_i - x_j\rangle} = \Big| \sum_i c_i e^{i\langle \omega, x_i\rangle} \Big|^2 \geq 0.$$

Since $\widehat{h}(\omega) \geq 0$, the whole integral is non–negative. Hence K is positive semidefinite.

2. Gaussian kernel.
Let

$$h(u) = \exp\Big(- \frac{\|u\|^2}{2\sigma^2} \Big), \qquad u \in \mathbb{R}^d.$$

We compute its Fourier transform

$$\widehat{h}(\omega) = \int_{\mathbb{R}^d} e^{-i\langle \omega, u\rangle} \exp\Big(- \frac{\|u\|^2}{2\sigma^2} \Big)\, du.$$

Step 1: write the exponent

We combine the two exponential terms:

$$-\frac{\|u\|^2}{2\sigma^2} - i\langle \omega, u\rangle.$$

We now complete the square. Observe that

$$\|u + i\sigma^2\omega\|^2 = \|u\|^2 + 2i\sigma^2\langle \omega, u\rangle - \sigma^4\|\omega\|^2.$$

Therefore

$$-\frac{\|u\|^2}{2\sigma^2} - i\langle \omega, u\rangle = -\frac{1}{2\sigma^2}\|u + i\sigma^2\omega\|^2 - \frac{\sigma^2}{2}\|\omega\|^2.$$

Hence

$$e^{-i\langle \omega, u\rangle} e^{-\frac{\|u\|^2}{2\sigma^2}} = \exp\Big(- \frac{\sigma^2}{2}\|\omega\|^2 \Big) \exp\Big(- \frac{\|u + i\sigma^2\omega\|^2}{2\sigma^2} \Big).$$

Step 2: change of variables

We get

$$\widehat{h}(\omega) = \exp\Big(- \frac{\sigma^2}{2}\|\omega\|^2 \Big) \int_{\mathbb{R}^d} \exp\Big(- \frac{\|u + i\sigma^2\omega\|^2}{2\sigma^2} \Big)\, du.$$

Now perform the change of variable

$$v = u + i\sigma^2\omega.$$

Since this is just a translation in $\mathbb{C}^d$, the integral remains the same:

$$\int_{\mathbb{R}^d} \exp\Big(- \frac{\|u + i\sigma^2\omega\|^2}{2\sigma^2} \Big)\, du = \int_{\mathbb{R}^d} \exp\Big(- \frac{\|v\|^2}{2\sigma^2} \Big)\, dv.$$

Step 3: evaluate the Gaussian integral

The Gaussian integral in d dimensions is well known:

$$\int_{\mathbb{R}^d} \exp\Big(-\frac{\|v\|^2}{2\sigma^2}\Big)\, dv = (2\pi\sigma^2)^{d/2}.$$

Final result

Putting everything together,

$$\widehat{h}(\omega) = (2\pi\sigma^2)^{d/2} \exp\Big(-\frac{\sigma^2}{2}\|\omega\|^2\Big).$$

Exercise 20 (Kernelization of k-means).

Let K be a PSD kernel on a set X, with RKHS $\mathcal{H}$ and feature map $\phi : X \to \mathcal{H}$. We observe data $x_1, \dots, x_n \in X$ and consider the k-means objective in $\mathcal{H}$:

$$L(c_1, \dots, c_k) = \sum_{i=1}^{n} \min_{1 \le j \le k} \|\phi(x_i) - c_j\|_{\mathcal{H}}^2, \qquad c_j \in \mathcal{H}.$$

1. Fixing the cluster assignments.
For each i, define

$$s_i = \arg\min_{1 \le j \le k} \|\phi(x_i) - c_j\|_{\mathcal{H}}^2.$$

Then by definition of the minimum,

$$L(c_1, \dots, c_k) = \sum_{i=1}^{n} \|\phi(x_i) - c_{s_i}\|_{\mathcal{H}}^2.$$

2. Optimal centroid in a Hilbert space.
Let $z_1, \dots, z_m \in \mathcal{H}$ and consider

$$F(c) = \sum_{i=1}^{m} \|z_i - c\|_{\mathcal{H}}^2.$$

Expand:

$$F(c) = \sum_i \big(\|z_i\|^2 - 2\langle z_i, c\rangle + \|c\|^2\big) = \sum_i \|z_i\|^2 - 2\Big\langle \sum_i z_i, c\Big\rangle + m\|c\|^2.$$

This is a strictly convex quadratic in c. Its unique minimizer satisfies

$$\nabla F(c) = -2\sum_i z_i + 2mc = 0,$$

hence

$$c = \frac{1}{m}\sum_{i=1}^{m} z_i.$$

Therefore the optimal centroid of points z_i is their average. This remains true in any Hilbert space.

3. Expression of the optimal centers.
Let

$$C_j = \{i : \; s_i = j\}$$

be the index set of points in cluster j. Applying the previous result with $z_i = \phi(x_i)$, the optimal center for cluster j is

$$c_j = \frac{1}{|C_j|}\sum_{i\in C_j} \phi(x_i).$$

Plugging this back into the objective yields

$$L = \sum_{i=1}^{n} \left\| \phi(x_i) - \frac{1}{|C_{s_i}|}\sum_{j\in C_{s_i}} \phi(x_j) \right\|_{\mathcal{H}}^2.$$

4. Expanding the squared norm.
For each i, expand using the Hilbert norm:

$$\left\|\phi(x_i) - \frac{1}{|C|}\sum_{j\in C}\phi(x_j)\right\|^2 = \langle \phi(x_i), \phi(x_i)\rangle - \frac{2}{|C|}\sum_{j\in C}\langle \phi(x_i), \phi(x_j)\rangle + \frac{1}{|C|^2}\sum_{j,\ell\in C}\langle \phi(x_j), \phi(x_\ell)\rangle.$$

Using the reproducing property,

$$\langle \phi(x), \phi(y)\rangle = K(x,y).$$

Therefore

$$\left\|\phi(x_i) - \frac{1}{|C|}\sum_{j\in C}\phi(x_j)\right\|^2 = K(x_i,x_i) - \frac{2}{|C|}\sum_{j\in C}K(x_i,x_j) + \frac{1}{|C|^2}\sum_{j,\ell\in C}K(x_j,x_\ell).$$

5. Final kernelized k-means objective.
Hence the k-means problem in $\mathcal{H}$ is equivalent to minimizing over assignments $s_1, \ldots, s_n \in \{1, \ldots, k\}$:

$$\boxed{\sum_{i=1}^{n}\left[K(x_i,x_i) - \frac{2}{|C_{s_i}|}\sum_{j\in C_{s_i}}K(x_i,x_j) + \frac{1}{|C_{s_i}|^2}\sum_{j,\ell\in C_{s_i}}K(x_j,x_\ell)\right]}$$

where $C_j = \{i : s_i = j\}$.

All terms depend only on the Gram matrix

$$G_{ij} = K(x_i, x_j).$$

Conclusion.
Therefore, performing k-means in the (possibly infinite-dimensional) RKHS $\mathcal{H}$ is equivalent to solving a combinatorial optimization problem over the cluster labels (s_i) using only the kernel matrix G. This is called *kernel k-means.*

Exercise 21 (Heat kernel on graphs).

Let $G = (V, E)$ be a finite, connected, undirected graph with vertices $V = \{v_1, \ldots, v_n\}$. Let A be its adjacency matrix, D its degree matrix and $L = D - A$ its graph Laplacian.

For a vector $f(t) \in \mathbb{R}^n$ we consider the diffusion equation

$$\partial_t f(t) = -Lf(t), \qquad f(0) = f_0.$$

1. Solution of the diffusion equation.
We claim that

$$f(t) = e^{-tL} f_0$$

is a solution.

Recall that for any square matrix M,

$$e^M = \sum_{k=0}^{\infty} \frac{M^k}{k!}.$$

Then

$$\frac{d}{dt} e^{-tL} = \sum_{k=1}^{\infty} \frac{(-1)^k t^{k-1}}{(k-1)!} L^k = -L \sum_{k=0}^{\infty} \frac{(-t)^k}{k!} L^k = -Le^{-tL}.$$

Hence

$$\partial_t f(t) = \partial_t (e^{-tL} f_0) = -Le^{-tL} f_0 = -Lf(t),$$

and clearly $f(0) = e^0 f_0 = f_0$.

2. L is symmetric and positive semi-definite.
Since A is symmetric and D is diagonal, $L = D - A$ is symmetric.
For any $x \in \mathbb{R}^n$,

$$x^\top Lx = \sum_i D_{ii} x_i^2 - \sum_{i,j} A_{ij} x_i x_j = \frac{1}{2} \sum_{i,j} A_{ij} (x_i - x_j)^2 \geq 0.$$

Hence L is positive semi-definite.

3. The heat kernel is a PSD kernel.

Since L is symmetric positive semi-definite, it admits an orthonormal eigen-decomposition

$$L = U\Lambda U^\top,$$

with $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$, $\lambda_i \geq 0$. Then

$$e^{-tL} = Ue^{-t\Lambda}U^\top, \qquad e^{-t\Lambda} = \mathrm{diag}(e^{-t\lambda_1}, \dots, e^{-t\lambda_n}),$$

which has nonnegative eigenvalues. Thus e^{-tL} is symmetric positive semi-definite. Therefore

$$K_t(v_i, v_j) = (e^{-tL})_{ij}$$

defines a PSD kernel on V.

4. Heat kernel on the complete graph.

Let G be the complete graph on n vertices. Then

$$A = \mathbf{1}\mathbf{1}^\top - I, \qquad D = (n-1)I,$$

so

$$L = nI - \mathbf{1}\mathbf{1}^\top.$$

The vector $\mathbf{1}$ is an eigenvector with eigenvalue 0, and any vector orthogonal to $\mathbf{1}$ has eigenvalue n. Hence

$$e^{-tL} = P_0 + e^{-nt}P_\perp,$$

where $P_0 = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ and $P_\perp = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. Thus

$$e^{-tL} = e^{-nt}I + (1 - e^{-nt})\frac{1}{n}\mathbf{1}\mathbf{1}^\top.$$

5. Why shortest-path kernels are not PSD.

Define

$$K_{ij} = e^{-d(v_i, v_j)^2},$$

where d is the graph shortest-path distance. In general this matrix need not be PSD. For example, take a simple path graph with three vertices $v_1 - v_2 - v_3$. Then

$$d = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{pmatrix}, \qquad K = \begin{pmatrix} 1 & e^{-1} & e^{-4} \\ e^{-1} & 1 & e^{-1} \\ e^{-4} & e^{-1} & 1 \end{pmatrix}.$$

One can check that this matrix has a negative eigenvalue, hence is not PSD. Therefore $e^{-d(v_i, v_j)^2}$ does not in general define a kernel on graphs.