Université de Toulouse
M2 RI

Robust Optimization and Statistical Learning

Final Exam

2024/2025

Rules: *All answers should be as developed as possible. Some questions are more guided or open than others; for the latter, there can be more than one admissible answer, the thorough treatment of the question will be privileged over one particular answer. I you think you need to add an extra assumption, say it explicitly in your answer. Duration: 3 hours.*

In the whole exam, unless explicitly specified, we let $(\Omega, \mathcal{A})$ be a measurable space. We denote by $\mathcal{P}(\Omega)$ the set of probability distributions over Ω, and we let μ and ν be two probability measures on $(\Omega, \mathcal{A})$. In addition, we assume that τ is a σ-finite measure on $(\Omega, \mathcal{A})$ satisfying $\mu \ll \tau$ and $\nu \ll \tau$ and define $p = d\mu/d\tau, q = d\nu/d\tau$.

Exercise 1. — The *Hellinger distance* between μ and ν is defined as follows:

$$H(\mu,\nu) := \sqrt{\int \left(\sqrt{p} - \sqrt{q}\right)^2 d\tau} = \sqrt{\int \left(\sqrt{d\mu} - \sqrt{d\nu}\right)^2}. \tag{1}$$

Show the following properties:

1. $H(\mu,\nu) \geq 0$ with equality if and only if $\mu = \nu$.
2. H is symmetric and verifies the triangle inequality.
3. $H^2(\mu,\nu) = 2\left(1 - \int \sqrt{pq} d\tau\right) = 2\left(1 - \int \sqrt{d\mu d\nu}\right)$.
4. $\frac{1}{2}H^2(\mu,\nu) \leq \|\mu - \nu\|_{TV}$. *Hint: we recall that* $\|\mu - \nu\|_{TV} = 1 - \int \min(p,q) d\tau$.
5. $\|\mu - \nu\|_{TV} \leq H(\mu,\nu)\sqrt{1 - \frac{H^2(\mu,\nu)}{4}}$. *Hint: we can first show that* $\int \min(p,q)d\tau + \int \max(p,q)d\tau = 2$

Exercise 2 (IPM on quadratics)**.** — In this exercise, we restrict ourselves to distributions on $\mathbb{R}$. Let us consider the following Integral Probability Metric (IPM) candidate:

$$d(\mu,\nu) := \sup_{f \in \mathcal{F}_q} \left| \int f \, d\mu - \int f \, d\nu \right|.$$

where the function class $\mathcal{F}_q := \{x \mapsto ax^2 + bx + c : a, b, c \in [-1,1]^3\}$ is a set of quadratic functions on $\mathbb{R}$.

1. Can we get rid of the absolute value in the definition of d?
2. How can you express d using the moments of μ and ν?
   (We recall that that the k-th moment of μ is $m_k(\mu) = \mathbb{E}_{X\sim\mu}[X^k]$, $k \in \mathbb{N}$)
3. Is d a metric on probability measures on $\mathbb{R}$?
4. Is d a metric on Gaussian distributions? More generally, for what families of distributions is d a metric?
5. Can we compare d with W_1 ? With the total variation ? What if we restricted ourselves to probability measures on $[0,1]$? *Hint: we recall that* W_1 *and the total variation are IPMs with* $\mathcal{F}$ *the set of 1-Lipchitz continuous functions and the set of functions with* ∞*-norm bounded by* 1 *respectively.*
6. Is the condition $\{a, b, c \in [-1,1]^3\}$ in $\mathcal{F}_q$ necessary to define a metric? If not, by what can we replace it?

Exercise 3 (Nash Equilibria)**.** — The parts of this exercise are independent.

A– *2×2 Coordination Game.* Consider the following two-player game in normal form. Both players have two strategies: A and B. The payoff matrix is given by:

	A	B
A	$(3,3)$	$(0,2)$
B	$(2,0)$	$(1,1)$

1. Identify all *pure strategy* Nash equilibria.
2. Find the *mixed strategy* Nash equilibrium and verify that in equilibrium each player is strongly indifferent among the pure strategies played with positive probability.

B– *Iterated Elimination of Strictly Dominated Strategies in a 3×3 Game.* Consider the following game with Player 1 choosing among rows T, M, and B, and Player 2 choosing among columns L, C, and R:

	L	C	R
T	$(2,0)$	$(1,1)$	$(4,2)$
M	$(3,4)$	$(1,2)$	$(2,3)$
B	$(1,3)$	$(0,2)$	$(3,0)$

3. Identify a strategy for Player 1 that is strictly dominated and eliminate it.
4. Then, for the reduced game, identify a strictly dominated strategy for Player 2 and eliminate it.
5. Find the pure strategy Nash equilibria in the resulting 2×2 game.
6. Find the mixed Nash equilibrium of the full 3×3 game.

Exercise 4. — Let $X_1, \dots, X_n$ be i.i.d. samples from distribution μ on $\mathbb{R}$ and define the empirical measure

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}\,.$$

We give ourselves a family $(q_\lambda)_{\lambda\in\Lambda}$ of probability distributions on $\mathbb{R}$ parametrized by $\lambda \in \Lambda \subset \mathbb{R}$ (eg. the family of exponential distributions). The goal of the exercise is to explore different ways of tuning λ and their consequences in terms of variational inference.

1. Show that

$$\mathrm{KL}(\mu||q_\lambda) = \mathbb{E}_{X\sim\mu}[-\log(q_\lambda(X))] + C \tag{2}$$

where C is a quantity independent from λ to precise. *Hint: we recall that* $KL(\mu||\nu) = \int \log\left(\frac{d\mu}{d\nu}\right) d\mu$ *if* $\mu \ll \nu$ *and* $+\infty$ *otherwise.*

2. Deduce that with probability one,

$$\arg\min_{\lambda\in\Lambda} \mathrm{KL}(\mu||q_\lambda) = \arg\max_{\lambda\in\Lambda} \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \log(q_\lambda(X_i)) \tag{3}$$

what is the interpretation of this result?

Now, we note $g = \log \circ\, q_\lambda$ and consider the distributionally robust problem

$$\begin{aligned} \sup_{\nu\in\mathcal{P}(\mathbb{R})} \mathbb{E}_\nu[g] &= \int g(x) d\nu(x) \\ \text{s.t. } \mathrm{KL}(\nu||\mu_n) &\le \rho \end{aligned} \tag{4}$$

where $\rho \ge 0$ is the robustness radius.

3. If $\rho = 0$, what is the closed form expression of problem (4)? What if $\rho \to +\infty$?
4. For some fixed $\rho \in [0, +\infty)$, what is the support of the optimal $\nu^\star$ (that we will assume to exist)? Then, how can problem (4) be reformulated as finite-dimensional optimization problem?
5. Let $n = 2, \rho = 0.1$, what is the optimal value of problem (4)? *Hint: we give* $x\log(2x) + (1-x)\log(2(1-x))) = 0.1$ *for* $x = 0.28$ *and* $x = 0.72$.
6. In fact, one can show that the value of problem (4) is equal to $\inf_{\gamma>0}\{\gamma\rho + \gamma\log\mathbb{E}_{\mu_n}\exp(g/\gamma)\}$. How can you relate this to question 3.?
7. Why is the expression of 6. more adapted than equation (4) to study the problem as $n \to \infty$?

Exercise 5 (f-Divergences)**.** — Let $\mu, \nu \in \mathcal{P}(\Omega)$ be such that $\mu \ll \nu$. Then, for a convex function $f : X := [0, +\infty) \to \mathbb{R} \cup \{+\infty\}$ such that $f(x)$ is finite for all $x > 0$, $f(1) = 0$, and $f(0) = \lim_{t \to 0^+} f(t)$, the f-divergence of μ from ν is defined as

$$D_f(\mu \parallel \nu) \equiv \int_\Omega f\left(\frac{d\mu}{d\nu}\right) d\nu\,. \tag{5}$$

Let f^* be the convex conjugate of f, ie $f^*(y) := \sup\{\langle y, x\rangle - f(x) \,:\, x \in X\}$ for all $y \in \mathbb{R}$. Let $\mathrm{effdom}(f^*)$ be the effective domain of f^*, that is, $\mathrm{effdom}(f^*) = \{y : f^*(y) < \infty\}$.

1. Show that

$$\mathbb{E}_\mu[g] \leq D_f(\mu \parallel \nu) + \mathbb{E}_\nu[f^* \circ g] \tag{6}$$

for all $g : \Omega \to \mathrm{effdom}(f^*)$. *(Hint: you can use that $f^*(y) + f(x) \geq xy$ for all x, y by convex conjugation)*

In fact, we have the following variational principle:

$$D_f(\mu \parallel \nu) = \sup_{g:\Omega \to \mathrm{effdom}(f^*)} \mathbb{E}_\mu[g] - \mathbb{E}_\nu[f^* \circ g] \tag{7}$$

that we will take as granted for the remainder of the exercise.

2. For $f(x) = x \log x$, verify that D_f is the Kullback-Liebler divergence. What is f^*? Instantiate (7) in that case.
3. For $f(x) = \frac{1}{2}|x - 1|$, show that D_f is an integral probability metric. What common distance does this correspond to?

In addition, if $g : \Omega \to \mathrm{effdom}(f^*)$ is upper-bounded and $\lim_{t \to +\infty} f(t) = +\infty$, convex duality tells us similarly that:

$$\sup_{\mu \in \mathcal{P}(\Omega) : \mu \ll \nu} \{\mathbb{E}_\mu[g] - D_f(\mu \parallel \nu)\} = \inf_{\lambda \in \mathbb{R}} \{\mathbb{E}_\nu[f^*(g + \lambda)] - \lambda\} \tag{8}$$

that we will also take as granted for the remainder of the exercise.

4. Instantiate (8) with $f(x) = x \log x$. What kind of result of the course do we recover?
5. Instantiate (8) with $f(x) = (x-1)^2$. Is D_f a distance? What are the striking differences with the previous question?
6. Let $\Omega = \{a_1, a_2, ..., a_n\}$ be a set of n points. What is $\sup_{\mu \in \mathcal{P}(\Omega)} \{\mathbb{E}_\mu[g]\}$?
7. In the setup of question 6., instantiate (8) with $f(x) = x \log x$ and ν the uniform distribution over these points. Show that we recover a well-known smoothing of the maximum.
8. For a general Ω, how can we use (8) to find an approximation of the supremum of g? What do we have to compute?

Bonus. Why don't we have simply $\sup_{\mu \in \mathcal{P}(\Omega) : \mu \ll \nu} \{\mathbb{E}_\mu[g] - D_f(\mu \parallel \nu)\} = \mathbb{E}_\nu[f^* \circ g]$? Or rather, for which space $\mathcal{Y}$ do we have $\sup_{\mu \in \mathcal{Y} : \mu \ll \nu} \{\mathbb{E}_\mu[g] - D_f(\mu \parallel \nu)\} = \mathbb{E}_\nu[f^* \circ g]$