Lecture 6: Divergences between Probability Measures

Lénaïc Chizat

March 10, 2021

The material of today's lecture is partly adapted from Gabriel Peyré's lecture notes.

1 Motivating problem: density fitting

In statistics, imaging, or machine learning, one of the most fundamental problems is to compare a probability distribution $\nu \in \mathcal{P}(\mathbb{R}^d)$ arising from measurements to a model, namely a parameterized family of distributions $\{\mu_\theta, \theta \in \Theta\}$ where typically $\Theta \subset \mathbb{R}^p$. A suitable parameter can be obtained by minimizing

$$\min_{\theta \in \Theta} F(\theta) \coloneqq D(\mu_\theta, \nu)$$

where $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to [0, +\infty]$ is a *divergence*[1], i.e. a quantity that quantifies the discrepancy between μ and ν.

Example 1.1. One can choose $D(\mu, \nu) = W_p^p(\mu, \nu)$. When ν is an empirical measure and with $p = 2$, this is called the *Minimum Kantorovich estimator*. A drawback of this discrepancy is that it is computationally expensive, compared to other discrepancies that we will see in this lecture.

Example 1.2 (Maximum likelihood)**.** Let $x_1, \dots, x_n \in \mathbb{R}^d$ be independent samples from ν. When μ_θ has a density ρ_θ with respect to a reference measure σ (e.g. the Lebesgue measure), the maximum likelihood estimator (MLE) is obtained by solving

$$\min_{\theta \in \Theta} -\frac{1}{n} \sum_{i=1}^{n} \log(\rho_\theta(x_i)).$$

This corresponds to using an empirical counterpart of the Kullback-Leibler loss since this converges towards $-\int_X \log(\rho_\theta(x) \mathrm{d}\nu(x) = \mathrm{KL}(\mu_\theta, \nu) - \int \log(\mathrm{d}\nu/\mathrm{d}\sigma)\mathrm{d}\nu$ (one can show this equality provided $\mu_\theta \ll \nu$ and $\nu \ll \sigma$ and all the terms are finite).

The MLE is a statistically optimal estimation procedure in certain cases, but fails:

- when there is no natural reference measure σ;
- when the density ρ_θ is difficult to compute;
- the resulting objective F is too complicated to minimize.

[1] In (applied) mathematics, *divergence* generally refers to a nonnegative quantity $D(a, b)$ that satisfies $D(a, b) = 0$ if $a = b$ (such as a distance, a squared distance, the Kullback-Leibler divergence,...)

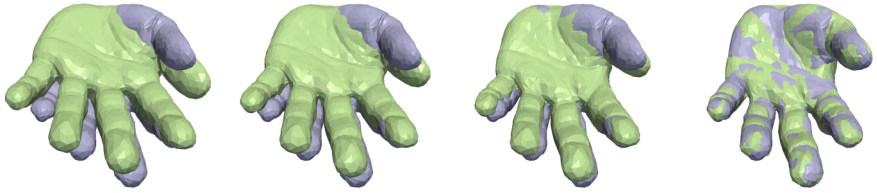

Figure 1: Gradient descent algorithm for density fitting. Here $(h_\theta)_\theta$ is a parameterized set of diffeomorphisms, $(h_{\theta_t})_\#\zeta$ is in green and the target ν is in purple. Image from [3].

Generative models. A typical set-up where all these problems appear is for so-called generative models, where the parametric measure is written as a push-forward of a fixed reference measure $\zeta \in \mathcal{P}(Z)$

$$\mu_\theta = (h_\theta)_\#\zeta \quad \text{where} \quad h_\theta : Z \to \mathbb{R}^d.$$

This leads to the objective function $F(\theta) = D((h_\theta)_\#\zeta, \nu)$.

The typical approach to tackle such problems numerically, is the gradient descent algorithm: initialize $\theta_0 \in \Theta$ and define for $t \geqslant 0$,

$$\theta_{t+1} = \theta_t - \eta \nabla F(\theta_t)$$

where $\eta > 0$ is a step-size (with potentially a projection on Θ if it is not a vector space). See Figure 1 for an example. Since F is non-convex, there is no guarantee that $F(\theta_t)$ converges to the minimum. In practice, the algorithm behaves better when the divergence is "geometrically faithful" (such as W_2^2). Let us give a formula for the gradient under strong regularity assumptions (which could be relaxed). Here $E : \mu \mapsto D(\mu, \nu)$.

Proposition 1.3 (Chain rule for generative models)**.** *Assume that $E : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is such that for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, there exists a function $E'(\mu) \in \mathcal{C}^1(\mathbb{R}^d)$ with $\nabla E'(\mu)$ Lipschitz, and such that for all $\nu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$E(\nu) - E(\mu) = \int_{\mathbb{R}^d} E'(\mu) \mathrm{d}(\nu - \mu) + o(W_2(\mu, \nu)).$$

Assume moreover that $h : \mathbb{R}^p \to L^2(\zeta; \mathbb{R}^d)$ is (Fréchet) differentiable, with partial derivatives at θ denoted by $\partial_i h_\theta \in L^2(\zeta; \mathbb{R}^d)$. Then $F : \theta \mapsto E((h_\theta)_\#\zeta)$ is (Fréchet) differentiable with gradient, for $i = 1, \dots, p$,

$$[\nabla F(\theta)]_i = \int_Z \nabla E'((h_\theta)_\#\zeta)(h_\theta(z))^\top \partial_i h_\theta(z) \mathrm{d}\zeta(z).$$

Proof. We first study $G : f \mapsto E(f_\#\zeta)$ and show that G is Fréchet differentiable with differential $\mathrm{D}G(f)(\delta f) = \int \nabla E'(f_\#\zeta)(f(z))^\top \delta f(z) \mathrm{d}\zeta(z)$. Then the conclusion follows by the usual chain rule for Fréchet differentials.

For $f, \delta f \in L^2(\zeta, \mathbb{R}^d)$, we have that $W_2^2(f_\#\zeta, (f + \delta f)_\#\zeta) \leqslant \|\delta f\|^2_{L^2(\zeta)}$ by taking $(f, f + \delta f)_\#\zeta$ as an admissible transport plan. Thus, by our assumption on E,

$$\begin{aligned} E((f + \delta f)_\#\zeta) - E(f_\#\zeta) &= \int_Z [E'(f_\#\zeta)(f(z) + \delta f(z)) - E'(f_\#\zeta)(f(z))] \mathrm{d}\zeta(z) + o(\|\delta f\|) \\ &= \int_Z \nabla E'(f_\#\zeta)(f(z))^\top \delta f(z) \mathrm{d}\zeta(z) + O(\mathrm{Lip}(\nabla E'(f_\#\zeta)) \|\delta f\|^2) + o(\|\delta f\|) \end{aligned}$$

by a Taylor expansion of $E'(f_{\#}\zeta)$ in the integral. This shows the formula for $\mathrm{D}G(f)$ and concludes the proof. □

Example 1.4. Show that if $W : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is symmetric and differentiable with a Lipschitz gradient, then $E(\mu) \coloneqq \int W(x,y)\mathrm{d}\mu(x)\mathrm{d}\mu(y)$ satisfies the assumptions of Prop 1.3 with $E'(\mu)(x) = \int W(x,y)\mathrm{d}\mu(y)$.

Rest of this lecture We will now introduce various notions of divergences : Csiszár-divergences, dual norms, MMD and Sinkhorn divergences. Much can be said about each of them, but we will focus on discussing (i) the divergence property and (ii) the weak continuity. In the rest of this lecture, we assume that X is a compact metric space.

2 Csiszár divergences

Maybe the most classical way to compare two probability measures are the total variation norm and the Kullback-Leibler divergence. They belong to the family of *Csiszár divergences* – also known as f-divergences – which consist in comparing the relative densities to 1. They are simple to compute between discrete distributions ($O(n)$ operations for distributions with n atoms) but are not weakly continuous.

Definition 2.1 (f-divergence)**.** Let $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be a convex function. For any $\mu, \nu \in \mathcal{P}(X)$, let $\mu = \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\nu + \mu^\perp$ be the Lebesgue decomposition of μ with respect to ν. The divergence is defined by

$$D_f(\mu,\nu) \coloneqq \int_X f\Big(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\Big)\mathrm{d}\nu + f'_\infty(1)\cdot\mu^\perp(X).$$

where $f'_\infty(x) = \lim_{t\to\infty} f(tx)/t \in \mathbb{R}\cup\{\infty\}$ is the asymptotic speed of growth of f in the direction x.

If $f'_\infty(1) = \infty$ then f grows faster than any linear function (f is said *superlinear*).

Proposition 2.2. *Let f be convex such that $\min f = 0$ and $\arg\min f = \{1\}$. Then $D_f(\mu,\nu) \geqslant 0$ with equality if and only if $\mu = \nu$.*

Proof. If $\mu = \nu$, then $\mathrm{d}\mu/\mathrm{d}\nu = 1 \in L^1(\nu)$ and $\mu^\perp = 0$, thus $D_f(\mu,\nu) = \int f(1)\mathrm{d}\nu = 0$. Conversely, if $D_f(\mu,\nu) = 0$ then $\mu^\perp = 0$ (because $f'_\infty(1) \geqslant f(2) - f(1) > 0$) and $\mathrm{d}\mu/\mathrm{d}\nu = 1 \in L^1(\nu)$ so $\mu = \nu$. □

Example 2.3 (Kullback-Leibler divergence/relative entropy)**.** This is the Csiszár divergence associated to the function

$$f(s) = \begin{cases} s\log(s) - s + 1 & \text{if } s > 0 \\ 1 & \text{if } s = 0 \\ +\infty & \text{if } s < 0 \end{cases}$$

which is convex, lsc, with unique minimum $f(1) = 0$. If $\mu \ll \nu$ then

$$D_f(\mu,\nu) = \int_X \Big(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\log\big(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\big) - \frac{\mathrm{d}\mu}{\mathrm{d}\nu} + 1\Big)\mathrm{d}\nu = \int_X \log\Big(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\Big)\mathrm{d}\mu = \mathrm{KL}(\mu,\nu)$$

and $D_f(\mu,\nu) = +\infty$ otherwise since $f'_\infty(1) = +\infty$.

Example 2.4 (Total variation)**.** This is the Csiszár divergence associated to

$$f(s) = \begin{cases} |s-1| & \text{if } s \geqslant 0 \\ +\infty & \text{otherwise} \end{cases}$$

We have $f'_\infty(1) = 1$ thus

$$D_f(\mu,\nu) = \int_X \Big(\big|\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\big|\mathrm{d}\nu + \mathrm{d}\mu^\perp\Big) \overset{(*)}{=} \int_X \mathrm{d}|\mu-\nu| = |\mu-\nu|(X)$$

where (*) comes from the fact that $(\mu-\nu)_+ = \max\{0, \mathrm{d}\mu/\mathrm{d}\nu - 1\}\nu + \mu^\perp$ and $(\mu-\nu)_- = \max\{0, 1 - \mathrm{d}\mu/\mathrm{d}\nu\}\nu$. Beware that in the probability literature there is sometimes a factor $1/2$ in front of the definition (so that it takes values in $[0,1]$ for probability distributions).

In the context of generative models, a drawback of f-divergences is that they are not weakly continuous : for instance $D_f(\delta_x, \delta_y) = f'_\infty(1) \cdot 1_{x=y}$ is not continuous at $x = y$. We have however weak lower-semicontinuity.

Proposition 2.5. *If $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is convex, lsc, and not identically $+\infty$, then $D_f(\mu,\nu)$ is (jointly) convex and weakly lower-semicontinuous and one has*

$$D_f(\mu,\nu) = \sup_{\varphi,\psi \in \mathcal{C}(X)} \int \varphi \mathrm{d}\mu + \int \psi \mathrm{d}\nu \quad s.t. \quad \varphi(x) + f^*(\psi(x)) \leqslant 0, \forall x \in X$$

where $f^* : s \mapsto \sup_{u\in\mathbb{R}} us - \cdot f(u)$ is the convex conjugate of f.

Proof idea. This is a special case of a more general property concerning *integral functionals* of *perspective functions*. Let $\psi_f : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be the perspective of f, defined as

$$\psi_f(t,x) = \begin{cases} t \cdot f(x/t) & \text{if } t > 0 \\ f'_\infty(x) & \text{if } t = 0 \\ +\infty & \text{if } t < 0 \end{cases}$$

By direct computation, it can be seen ψ_f is the convex conjugate of the (convex and lsc) function of $(s,y) \in \mathbb{R}^2$ that is worth 0 if $s + f^*(y) \leqslant 0$ and $+\infty$ otherwise (exercise). Thus ψ_f is convex and lsc. For any $\sigma \in \mathcal{P}(X)$ such that $\mu, \nu \ll \sigma$, it holds

$$D_f(\mu,\nu) = \int \psi_f\big(\frac{\mathrm{d}\nu}{\mathrm{d}\sigma}, \frac{\mathrm{d}\mu}{\mathrm{d}\sigma}\big)\mathrm{d}\sigma = \int \Big(\sup_{\psi + f^*(\varphi) \leqslant 0} \psi\frac{\mathrm{d}\nu}{\mathrm{d}\sigma} + \varphi\frac{\mathrm{d}\mu}{\mathrm{d}\sigma}\Big)\mathrm{d}\sigma \overset{(*)}{=} \sup_{\substack{\varphi,\psi\in\mathcal{C}(X) \\ \psi + f^*\circ\varphi \leqslant 0}} \int \varphi\mathrm{d}\mu + \int \psi\mathrm{d}\nu$$

To exchange $\int$ and sup, we have used a (non-trivial) interversion theorem [5, Thm. 6]. The convexity and weak lsc follow directly by this dual representation. See [1, Sec. 2.6] for a direct proof of lower-semicontinuity. □

Example 2.6 (Total variation)**.** For the total variation, we have by direct computation $f^*(s) = \max\{-1, s\}$ for $s \leqslant 1$ and $f^*(s) = +\infty$ for $s > 1$, so we recover the usual dual characterization

$$|\mu-\nu|(X) = \sup_{\substack{\varphi\in\mathcal{C}(X) \\ \varphi\leqslant 1}} \int \varphi\mathrm{d}\mu - \int \max\{-1,\varphi\}\mathrm{d}\nu = \sup_{\substack{\varphi\in\mathcal{C}(X) \\ \|\varphi\|_\infty \leqslant 1}} \int \varphi\mathrm{d}(\mu-\nu).$$

3 Integral probability metrics (dual norms)

3.1 General case

For a symmetric set B of measurable functions from X to $\mathbb{R}$ and $\alpha \in \mathcal{M}(X)$ a signed measure, let

$$\|\alpha\|_B \coloneqq \sup_{f \in B} \int_X f(x) \mathrm{d}\alpha(x) \tag{3.1}$$

The divergences associated to such dual norms, obtained with $\alpha = \mu - \nu$, for $\mu, \nu \in \mathcal{P}(X)$

$$D_B(\mu, \nu) \coloneqq \|\mu - \nu\|_B = \sup_{f \in B} \int f(x) \mathrm{d}(\mu(x) - \nu(x))$$

are often called "integral probability metrics", see [6] (or also "maximum mean discrepancy" but the latter is sometimes reserved to the special case discussed later).

Proposition 3.1. *If B is symmetric, bounded in sup-norm and contains 0, then $\|\cdot\|_B$ is a seminorm on $\mathcal{M}(X)$ (it is nonnegative, positively homogeneous and subadditive).*

Example 3.2 (Total variation)**.** It is recovered with $B = \{f \in \mathcal{C}(X) \; ; \; \|f\|_\infty \leqslant 1\}$.

Example 3.3 (Wasserstein-1)**.** It is the integral probability metric induced by the set of 1-Lipschitz functions $B = \{f \in \mathcal{C}(X); \mathrm{Lip}(f) \leqslant 1\}$.

Example 3.4 (Flat norm and the Dudley metric)**.** If the set B is bounded in $\|\cdot\|_\infty$, then $\|\cdot\|_B$ is a norm on the whole space $\mathcal{M}(X)$ of signed measures. This is not the case for $\|\cdot\|_{W_1}$, which is only finite for α such that $\int_X \mathrm{d}\alpha = 0$. This can be alleviated by imposing a bound on the value of the potential f, in order to define for instance the *flat norm*

$$B = \{f \; ; \; \mathrm{Lip}(f) \leqslant 1 \quad \text{and} \quad \|f\|_\infty \leqslant 1\}$$

It is similar to the *Dudley* metric, which uses

$$B = \{f \; ; \; \mathrm{Lip}(f) + \|f\|_\infty \leqslant 1\}.$$

The following proposition shows that to metrize the weak convergence, the set B should not be too large nor too small.

Proposition 3.5. *Let $(\alpha_k)_k$ be a bounded (for total variation $\|\cdot\|_{TV}$) sequence in $\mathcal{M}(X)$.*

(i) If $\mathcal{C}(X) \subset \overline{\mathrm{span}(B)}^{\|\cdot\|_\infty}$, i.e. if the span of B is dense in the set of continuous functions endowed with the sup-norm), then $\|\alpha_k - \alpha\|_B \to 0$ implies $\alpha_k \rightharpoonup \alpha$

(ii) If $B \subset \mathcal{C}(X)$ is compact (i.e. if it is closed, uniformly continuous and bounded) then $\alpha_k \rightharpoonup \alpha$ implies $\|\alpha_k - \alpha\|_B \to 0$.

Proof. (i) If $\|\alpha_k - \alpha\|_B \to 0$ then by duality, for any $f \in B$, since $\langle f, \alpha_k - \alpha \rangle| \leqslant \|\alpha_k - \alpha\|_B$ then $\langle f, \alpha_k \rangle \to \langle f, \alpha \rangle$. By linearity, this extends to $\mathrm{span}(B)$ and then to $\overline{\mathrm{span}(B)}^{\|\cdot\|_\infty}$ since $|\langle f, \alpha_k \rangle - \langle f', \alpha_k \rangle| \leqslant \|f - f'\|_\infty \sup_k \|\alpha_k\|_{TV}$.

(ii) We assume that $\alpha_k \rightharpoonup \alpha$ and we consider a subsequence α_{n_k} such that

$$\|\alpha_{n_k} - \alpha\|_B \to \limsup_k \|\alpha_k - \alpha\|_B.$$

Since B is compact, the maximum appearing in the definition of $\|\alpha_{n_k} - \alpha\|_B$ is reached, so there exists some $f \in B$ such that $\langle \alpha_{n_k} - \alpha, f_{n_k} \rangle = \|\alpha_{n_k} - \alpha\|_B$. By compactness, we can again extract a subsequence f_{n_k} (not relabelled for simplicity) that converges to some $f \in B \subset \mathcal{C}(X)$. One has

$$\|\alpha_{n_k} - \alpha\|_B = \langle \alpha_{n_k} - \alpha, f \rangle + \langle \alpha_{n_k}, f_{n_k} - f \rangle - \langle \alpha, f_{n_k} - f \rangle \to 0$$

because $\alpha_{n_k} - \alpha \rightharpoonup 0$ and $\|f_{n_k} - f\|_\infty \to 0$. □

Observe that this proof is a direct generalization of our proof that W_1 metrizes the weak topology on a compact X in Lecture 4.

3.2 Kernel Maximum Mean Discrepancies

We now describe an important class of integral probability metrics.

Definition 3.6 (Positive definite kernel)**.** A symmetric function $k : X \times X \to \mathbb{R}$ is said to be positive definite (p.d.) if for any $n \geqslant 1$, for any family $x_1, \dots, x_n \in X$ the matrix $(k(x_i, x_j))_{i,j}$ is positive semi-definite[2], i.e. for all $r \in \mathbb{R}^n$,

$$\sum_{i,j=1}^{n} r_i r_j k(x_i, x_j) \geqslant 0. \tag{3.2}$$

The kernel is said to be conditionally positive definite if Eq. (3.2) holds for all zero mean vectors r, i.e. such that $\sum_i r_i = 0$.

Definition 3.7 (MMD)**.** Given a continuous and positive semi-definite kernel $k : X \times X \to \mathbb{R}$, we define for $\alpha \in \mathcal{M}(X)$ (finite signed Borel measure)

$$\|\alpha\|_k^2 = \iint_X k(x, y) \mathrm{d}\alpha(x) \mathrm{d}\alpha(y).$$

The squared MMD between $\mu, \nu \in \mathcal{P}(X)$ is then

$$\|\mu - \nu\|_k^2 = \iint k \mathrm{d}\mu \otimes \mu + \iint k \mathrm{d}\nu \otimes \nu - 2 \iint k \mathrm{d}\mu \otimes \nu.$$

This definition as a squared quantity makes sense thanks to the following result.

Proposition 3.8. *If $k \in \mathcal{C}(X^2)$ is conditionally p.d., $\iint k \mathrm{d}\alpha \otimes \alpha \geqslant 0$ if $\int \mathrm{d}\alpha = 0$.*

Proof. Let $\alpha_n \in \mathcal{M}(X)$ be measures with finite support and zero mass such that $\alpha_n \rightharpoonup \alpha$. Since $\alpha_n \otimes \alpha_n \rightharpoonup \alpha \otimes \alpha$ (Lem.2.2 from Lect. 5), we have $0 \leqslant \iint k \mathrm{d}\alpha_n \otimes \alpha_n \to \iint k \mathrm{d}\alpha \otimes \alpha$. □

MMD as a dual norm. To develop a finer understanding of MMD, one needs to develop the theory of *Reproducible Kernel Hilbert Spaces* (RKHS). Since this is beyond the scope of this course, we limit ourselves to showing a link with dual norms via the following result.

Theorem 3.9 (Aronzsajn)**.** *k is a p.d. kernel on the set X if and only if there exists a Hilbert space $\mathcal{H}$ and a mapping $\Phi : X \to \mathcal{H}$ such that $k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$.*

[2] It would be more consistant to call such a kernel "positive semi-definite" but we are using the convention from the literature [7].

As a consequence the MMD consists in embedding $\mathcal{M}(X)$ into a Hilbert space $\mathcal{H}$ via the *kernel mean embedding* $\mu \mapsto \int \Phi \mathrm{d}\mu$, since $\|\mu - \nu\|_k = \|\int \Phi \mathrm{d}\mu - \int \Phi \mathrm{d}\nu\|_{\mathcal{H}}$. We also have

$$\|\alpha\|_k = \sup_{\|h\|_{\mathcal{H}} \leqslant 1} \left\langle h, \int \Phi \mathrm{d}\alpha \right\rangle = \sup_{f \in B} \int f \mathrm{d}\alpha$$

where $B = \{x \mapsto \langle h, \Phi(x)\rangle \ ; \ \|h\|_{\mathcal{H}} \leqslant 1\}$ so it is an integral probability metric.

Metrizing weak convergence. It can be shown that if k is *universal* (i.e. (i) of Prop. 3.5) holds), continuous and conditionally positive definite then $\|\cdot\|_k$ metrizes weak convergence in $\mathcal{P}(X)$, see e.g. [**?**]. Examples of such kernels on $\mathbb{R}^d$ are:

- the Gaussian kernel $k(x,y) = e^{-\frac{\|y-x\|_2^2}{2\sigma^2}}$ with $\sigma > 0$;
- the distance kernel $k(x,y) = -\mathrm{dist}(x,y)$ (its MMD is called the "Energy distance");

Discrete case. In the special case of discrete measures $\mu = \sum_{i=1}^m a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^n b_j \delta_{y_j}$ then we have

$$\|\mu - \nu\|_k^2 = \sum_{i,i'} a_i a_{i'} k(x_i, x_{i'}) + \sum_{j,j'} b_j b_{j'} k(y_j, y_{j'}) - 2 \sum_{i,j} a_i b_j k(x_i, y_j).$$

This requires $O((m+n)^2)$ operations to compute.

4 Sinkhorn divergences

Recall that W_p^p are often good choices of divergence, but are computationally expansive (another downside, not discussed in this course, is that they are difficult to estimate from random samples, in particular in high dimension [8]). Next, we build divergences from entropy regularized optimal transport.

4.1 Reminders on entropy regularized OT

With $c \in \mathcal{C}(X^2)$, the definition of entropy regularized optimal transport is

$$T_{c,\lambda}(\mu,\nu) \coloneqq \min_{\gamma \in \Pi(\mu,\nu)} \int c(x,y) \mathrm{d}\gamma(x,y) + \lambda \, \mathrm{KL}(\gamma; \mu \otimes \nu)$$

Note that this definition differs from the one in Lecture 3 by a constant λ (because the quantity $\mathcal{H}$ from Lecture 3 differs from KL by 1). We recall the following facts from Lecture 3:

- (Duality)

$$T_{c,\lambda}(\mu,\nu) = \sup_{\varphi,\psi \in \mathcal{C}(X)} \int \varphi(x) \mathrm{d}\mu(x) + \int \psi(y) \mathrm{d}\nu(y) + \lambda \Big(1 - \iint e^{(\varphi(x)+\psi(y)-c(x,y))/\lambda} \mathrm{d}\mu(x) \mathrm{d}\nu(y)\Big)$$

- (Optimality conditions) There exists maximizers $(\varphi_\lambda, \psi_\lambda)$ and a unique minimizer γ_λ linked by the optimality condition

$$\mathrm{d}\gamma_\lambda(x,y) = e^{(\varphi_\lambda(x)+\psi_\lambda(y)-c(x,y))/\lambda} \mathrm{d}\mu(x) \mathrm{d}\nu(y)$$

  It follows in particular that

$$T_{c,\lambda}(\mu,\nu) = \int \varphi_\lambda \mathrm{d}\mu + \int \psi_\lambda \mathrm{d}\nu.$$

4.2 Is $T_{c,\lambda}$ a suitable divergence?

Proposition 4.1 (Interpolation properties)**.** *For $\mu, \nu \in \mathcal{P}(X)$ and $c \in \mathcal{C}(X \times X)$, it holds*

$$T_{c,\lambda}(\mu,\nu) \to \begin{cases} T_c(\mu,\nu) \coloneqq T_{c,0}(\mu,\nu) & \text{as } \lambda \to 0 \\ \int c(x,y)\mathrm{d}\mu(x)\mathrm{d}\nu(y) & \text{as } \lambda \to \infty \end{cases}$$

Moreover, denoting γ_λ the unique minimizer for $T_{c,\lambda}$, it holds $\gamma_\lambda \rightharpoonup \mu \otimes \nu$ as $\lambda \to \infty$.

Proof. (i) We first study the limit $\lambda \to 0$. Since $\mathrm{KL} \geqslant 0$, it holds $T_{c,\lambda} \geqslant T_c$. To prove the reverse inequality as $\lambda \to 0$, let $\gamma_0 \in \Pi(\mu,\nu)$ be optimal for $T_c(\mu,\nu)$. For any $\epsilon > 0$, there exists $\gamma_\epsilon \in \Pi(\mu,\nu)$ such that $|\int c\mathrm{d}\gamma - \int c\mathrm{d}\gamma_\epsilon| \leqslant \epsilon$ and $\mathrm{KL}(\gamma_\epsilon, \mu\otimes\nu) < +\infty$. One possible way to build γ_ϵ is to take (Q_i) a partition of X into sets of diameter less than $\mathrm{Lip}(c)/(2\epsilon)$ and to take

$$\gamma_\epsilon = \sum_{i,j} \frac{\gamma_0(Q_i \times Q_j)}{\mu(Q_i)\nu(Q_j)} \mathrm{d}(\mu|_{Q_i} \otimes \nu|_{Q_j})$$

(Exercise: show that γ_ϵ indeed satisfies our requirements). It follows that

$$T_{c,\lambda} \leqslant \epsilon + \lambda\,\mathrm{KL}(\gamma_\epsilon, \mu\otimes\nu) \underset{\lambda\to 0}{\longrightarrow} \epsilon.$$

As $\epsilon > 0$ was arbitrary, this shows that $T_{c,\lambda}(\mu,\nu) \underset{\lambda\to 0}{\longrightarrow} T_c(\mu,\nu)$.

(ii) Now we study the limit $\lambda \to \infty$. First, it is clear that $T_{c,\lambda}(\mu,\nu) \leqslant \int c\mathrm{d}\mu\otimes\nu$ since $\mu\otimes\nu \in \Pi(\mu,\nu)$. Let $(\lambda_k)_k$ be a positive sequence that diverges to $+\infty$, and let γ_k be the corresponding sequence of (unique) minimizers for T_{c,λ_k}. By optimality, we have $\int c\mathrm{d}\gamma_k + \lambda_k\,\mathrm{KL}(\gamma_k, \mu\otimes\nu) \leqslant \int c\mathrm{d}\mu\otimes\nu$ and thus

$$\mathrm{KL}(\gamma_k, \mu\otimes\nu) \leqslant \frac{1}{\lambda_k}\Big(\int c\mathrm{d}\mu\otimes\nu - \int c\mathrm{d}\gamma_k\Big) \to 0.$$

Moreover, by compactness of $\Pi(\mu,\nu)$ we can extract a converging subsequence $\gamma_{n_k} \rightharpoonup \gamma_\infty$. Since KL is weakly lower-semicontinuous (Prop. 2.5), it holds

$$\mathrm{KL}(\gamma_\infty, \mu\otimes\nu) \leqslant \liminf_{k\to\infty} \mathrm{KL}(\gamma_{n_k}, \mu\otimes\nu) = 0$$

Hence $\gamma_\infty = \mu\otimes\nu$ hence the conclusion. □

One could imagine replacing Wasserstein by its entropy regularized version, but the previous result shows that when λ is large, $T_{c,\lambda}$ behaves like an inner product rather than like a divergence. In particular, even for standard costs $c = \mathrm{dist}(x,y)^p$, $\mu \mapsto T_c^\lambda(\mu,\nu)$ is in general not minimized at $\mu = \nu$.

Corollary 4.2. *Let $\nu \in \mathcal{P}(X)$ be such that $\arg\min_{y\in X} \int c(x,y)\mathrm{d}\nu(y)$ is a singleton, denoted x^* and let $\mu_\lambda \in \arg\min_{\mu\in\mathcal{P}(X)} T_{c,\lambda}(\mu,\nu)$. Then as $\lambda \to \infty$, one has $\mu_\lambda \rightharpoonup \delta_{x^*}$.*

Proof. By assumption, the continuous function f defined by $f(x) = \int c(x,y)\mathrm{d}\nu(y)$ admits a unique minimizer $x^* \in X$. It follows that for any $r > 0$, there exists $\epsilon > 0$ such that $|f(x) - f(x^*)| \leqslant \epsilon$ implies $\mathrm{dist}(x,x^*) \leqslant r$ (to see this, observe that $g(r) \coloneqq \min_{x\in X, \mathrm{dist}(x,x^*)\geqslant r} f(x)$ satisfies $g(r) > f(x^*)$ and $\lim_{r\to 0^+} g(r) = f(x^*)$).

Now consider an increasing unbounded sequence $(\lambda_k)_k$ and let $(\mu_k)_k$ be the corresponding set of minimizers. Taking δ_{x^*} as a competitor, the optimality of μ_k together with Prop. 4.1 implies that for any $\epsilon > 0$, there exists k_0 such that for $k \geqslant k_0$,

$$\int f\mathrm{d}\mu_k = \int c\mathrm{d}\mu_k\otimes\nu \leqslant \int f(x^*) + \epsilon$$

Thus for any $r > 0$, there exists k_0' such that for $k \geqslant k_0'$, $\mu_k(B_r(x^*)) \geqslant 1 - r$ from which we deduce that μ_k converges weakly to δ_{x^*}. □

For instance, when $c(x, y) = \frac{1}{2}\|y - x\|_2^2$, then μ_λ converges to a Dirac mass located at the mean $\int x \mathrm{d}\nu(x)$ of ν.

4.3 Debiased quantity: the Sinkhorn divergence

Thinking of $-T_{c,\lambda}$ as an inner product suggests to define

$$S_{c,\lambda}(\mu, \nu) \coloneqq T_{c,\lambda}(\mu, \nu) - \frac{1}{2}T_{c,\lambda}(\mu, \mu) - \frac{1}{2}T_{c,\lambda}(\nu, \nu)$$

From a computational aspect, the debiasing terms add an essentially negligible cost because the Sinkhorn iterations for those problems are well-conditioned. We can already see that these correction terms allow to correct the asymptotic behavior when λ is large.

Proposition 4.3 (Interpolation properties)**.** *For $\mu, \nu \in \mathcal{P}(X)$ and $c \in \mathcal{C}(X \times X)$ it holds*

$$S_{c,\lambda}(\mu, \nu) \longrightarrow \begin{cases} T_c(\mu, \nu) & \text{as } \lambda \to 0 \\ \frac{1}{2}\|\mu - \nu\|_{-c}^2 & \text{as } \lambda \to \infty \end{cases}$$

where $\|\cdot\|_{-c}$ is the MMD associated to the kernel $-c$.

Proof. This is immediate from Proposition 4.1 which deals with $T_{c,\lambda}$. In particular, when $\lambda \to \infty$, it shows that

$$S_{c,\lambda}(\mu, \nu) \to \int c \mathrm{d}\mu \otimes \nu - \frac{1}{2}\int c \mathrm{d}\mu \otimes \mu - \frac{1}{2}\int c \mathrm{d}\nu \otimes \nu$$

which is precisely the definition of $\frac{1}{2}\|\mu - \nu\|_{-c}^2$. □

One can show that under regularity assumptions over μ and ν and for the cost $c(x, y) = \|y - x\|_2^2$ on $\mathbb{R}^d$ that $|S_{c,\lambda} - T_c| = O(\lambda^2)$, see [2]. Finally, under assumptions on the cost, we can show that $\mu \mapsto S_{c,\lambda}(\mu, \nu)$ is minimized at ν.

Proposition 4.4 (Positive semi-definiteness)**.** *If $k(x, y) = e^{-c(x,y)/\lambda}$ is a positive semidefinite kernel, then $S_{c,\lambda}(\mu, \nu) \geqslant 0$ with equality if $\mu = \nu$.*

Proof. In the following, we fix $\lambda > 0$ and denote (φ, φ) a solution for $T_{c,\lambda}(\mu, \mu)$ (there exists a solution of this form by symmetry and concavity of the dual problem) and (ψ, ψ) a solution for $T_{c,\lambda}(\nu, \nu)$. Using the suboptimal function (φ, ψ) in the dual maximization problem, one obtains

$$T_{c,\lambda}(\mu, \nu) \geqslant \int \varphi \mathrm{d}\mu + \int \psi \mathrm{d}\nu + \lambda\Big(1 - \int\int e^{(\varphi \oplus \psi - c)/\lambda} \mathrm{d}\mu \otimes \nu\Big)$$

Using the fact that $T_{c,\lambda}(\mu, \mu) = 2\int \varphi \mathrm{d}\mu$ and $T_{c,\lambda}(\nu, \nu) = 2\int \psi \mathrm{d}\nu$, this implies

$$\frac{1}{\lambda}S_{c,\lambda}(\mu, \nu) \geqslant 1 - \int\int e^{(\varphi \oplus \psi - c)/\lambda} \mathrm{d}\mu \otimes \nu = 1 - \langle \tilde{\mu}, \tilde{\nu} \rangle_k$$

with $\tilde{\mu} = e^{\varphi}/\lambda\mu$, $\tilde{\nu} = e^{\psi}/\lambda\nu$ and $\langle \cdot, \cdot \rangle_k$ is the positive semi-definite inner product associated with the kernel $k \coloneqq e^{-c/\lambda}$. By the optimality condition satisfied by φ, it holds

$$\|\tilde{\mu}\|_k^2 = \int\int e^{(\varphi(x) + \varphi(y) - c(x,y))/\lambda} \mathrm{d}\nu(x) \mathrm{d}\nu(y) = 1$$

and similarly $\|\tilde{\nu}\|_k^2 = 1$. So by Cauchy-Schwartz inequality, one has $\langle \tilde{\mu}, \tilde{\nu} \rangle_k \leqslant 1$ and finally $S_{c,\lambda}(\mu, \nu) \geqslant 0$. □

In case $e^{-c/\lambda}$ is a conditionally positive definite and universal kernel, we can further show that $S_{c,\lambda}(\mu_n, \mu) \to 0$ if and only if $\mu_n \rightharpoonup \mu$, see [4].

5 Practical session

The last practical session (taken from Gabriel Peyré's numerical tours website) is at the following link: https://nbviewer.jupyter.org/github/gpeyre/numerical-tours/blob/master/python/optimaltransp_6_entropic_adv.ipynb. Only exercises 4 and 5 are not already solved, but take time to understand each block of code. For those who do not wish to use Jupyter, you may of course paste the code in a different terminal. Note that there is a sign mistake before cell 11: one should read $+\mathrm{KL}$ instead of $-\mathrm{KL}$ and P is the solution to entropy regularized optimal transport. For those who wish to go further on this topic, you can check out Jean Feydy's great tutorial http://www.math.ens.fr/~feydy/Teaching/DataScience/gradient_flows.html.

References

[1] Luigi Ambrosio, Nicola Fusco, and Diego Pallara, *Functions of bounded variation and free discontinuity problems*, Courier Corporation, 2000.

[2] Lénaïc Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré, *Faster wasserstein distance estimation with the sinkhorn divergence*, Advances in Neural Information Processing Systems (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds.), vol. 33, Curran Associates, Inc., 2020, pp. 2257–2269.

[3] Jean Feydy, Benjamin Charlier, François-Xavier Vialard, and Gabriel Peyré, *Optimal transport for diffeomorphic registration*, International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 291–299.

[4] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré, *Interpolating between optimal transport and MMD using sinkhorn divergences*, The 22nd International Conference on Artificial Intelligence and Statistics, PMLR, 2019, pp. 2681–2690.

[5] Ralph Rockafellar, *Integrals which are convex functionals. ii*, Pacific Journal of Mathematics **39** (1971), no. 2, 439–469.

[6] Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R.G. Lanckriet, *On integral probability metrics,φ-divergences and binary classification*, arXiv preprint arXiv:0901.2698 (2009).

[7] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet, *Universality, characteristic kernels and rkhs embedding of measures.*, Journal of Machine Learning Research **12** (2011), no. 7.

[8] Jonathan Weed, Francis Bach, et al., *Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance*, Bernoulli **25** (2019), no. 4A, 2620–2648.