Lecture 6: Divergences between Probability measures

I Motivating problem: density fitting

- Fundamental problem: compare $\nu \in \mathcal{P}(\mathbb{R}^d)$ arising from measurements to a model which is a parameterized family of distributions $\{\mu_\theta ; \theta \in \Theta\}$ where typically $\Theta \in \mathbb{R}^p$.
- A suitable parameter can be obtained by minimizing:

$$\min_{\theta \in \Theta} F(\theta) \text{ where } F(\theta) = D(\mu_\theta, \nu) \qquad (*)$$

where $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to [0, +\infty]$ is a divergence.

- In this lecture, by divergence we mean $\begin{cases} D(\mu, \nu) \geq 0 \\ D(\mu, \mu) = 0 \end{cases}$.

<u>Example 1</u>: One can choose $D(\mu, \nu) = W_p^p(\mu, \nu)$ for some $p \geq 1$.
When ν is an empical measure, with $p=2$, the solution to $(*)$ is called the Minimum Kantorovich Estimator.

<u>Example 2</u>: Let $x_1, \dots, x_n \in \mathbb{R}^d$ are independent samples from ν. When μ_θ has a density ρ_θ w.r.t a reference measure σ, the maximum likelihood estimator (MLE) is

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

This corresponds to an empirical version of solving $(*)$ with $D(\mu_\theta, \nu) = KL(\mu_\theta, \nu)$ since (1) converges to $-\int \log(\rho_\theta(x))\, d\nu(x) = KL(\mu_\theta, \nu) - \int \log\left(\frac{d\nu}{d\sigma}\right) d\nu$ (provided all the terms are finite).

- Note that the MLE fails:
  - when there is no natural reference measure σ
  - when ρ_θ is difficult to compute
  - when the objective F is too complicated to minimize.

Genetive models

Generative models are when the parametric measure μ_θ is given by

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

and where $\zeta \in \mathcal{P}(\mathcal{Z})$ is a reference measure. This leads to

$$F(\theta) = D((h_\theta)_\# \zeta, \nu).$$

The typical approach to "minimize" F is the gradient descent algorithm:

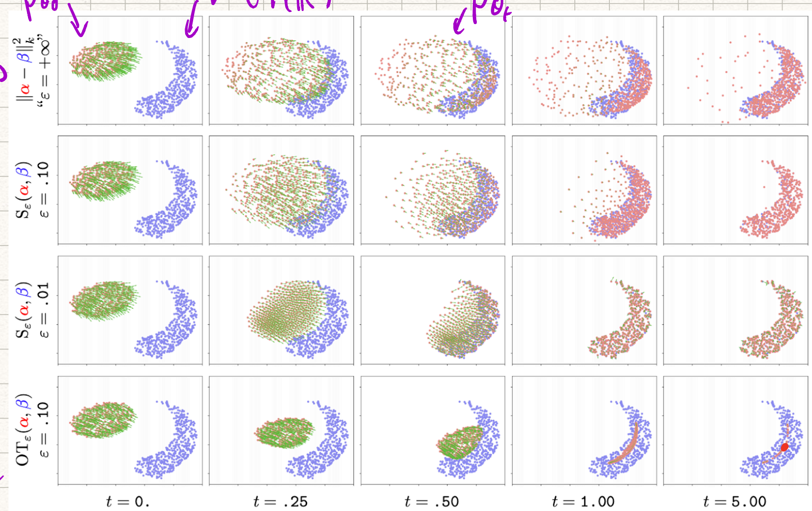
- initialize $\theta_0 \in \Theta$
- for $t = 1, 2, \ldots$ let $\theta_{t+1} = \theta_t - \eta \overbrace{\nabla F(\theta_t)}^{\text{formula ?}}$ where $\eta > 0$ is a step-size

μ_{θ_0} $\nu \in P(\mathbb{R}^2)$ μ_{θ_t}

MMD

Sinkhorn divergence

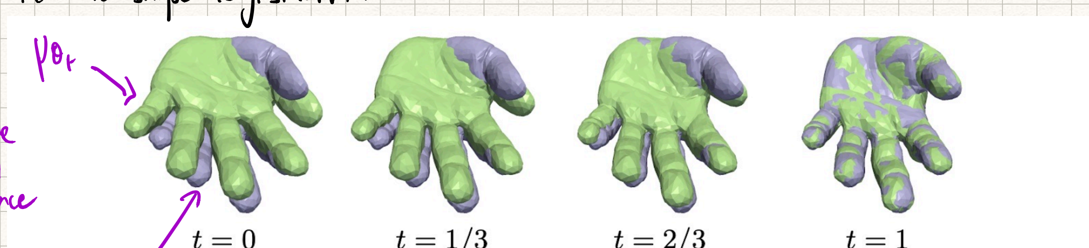
Entropy regularized OT



Application to shape registration:

μ_{θ_t}

D is the Sinkhorn divergence



$\nu \in P(\mathbb{R}^3)$

$(h_\theta)_\theta$ is a parameterized set of diffeomorphisms.

Let us give a formula for $\nabla F(\theta)$ under strong regularity assumptions.

Let us denote $E : \mu \mapsto D(\mu, \nu)$

<u>Proposition</u>. Assume that $E : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is such that $\forall \mu \in \mathcal{P}_2(\mathbb{R}^d)$, there exists a function $E'(\mu) \in \mathcal{C}^1(\mathbb{R}^d)$ with $\nabla E'(\mu)$ Lipschitz, and such that $\forall \nu \in \mathcal{P}(\mathbb{R}^d)$,

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

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

$$[\nabla F(\theta)]_i = \int_Z \nabla E'((h_\theta)_\# \zeta)(h_\theta(z))^T \partial_i h_\theta(z)\, d\zeta(z).$$

to digest in the practical session

<u>Proof</u>: First we study $G : f \overset{\in L^2(\zeta;\mathbb{R}^d)}{\mapsto} E(f_\# \zeta)$ and show that G is (Fréchet) differentiable with differential : $DG(f)(\delta f) = \int \nabla E'(f_\# \zeta)(f(z))^T \delta f(z)\, d\zeta(z)$. 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 $W_2^2(f_\# \zeta, (f+\delta f)_\# \zeta) \leq \|\delta f\|^2_{L^2(\zeta)}$ by taking $(f, f+\delta f)_\# \zeta$ as an admissible transport plan. Thus,

$$E((f+\delta f)_\# \zeta) - E(f_\# \zeta) = \int_Z [E'(f_\# \zeta)(f(z) + \delta f(z)) - E'(f_\# \zeta)(f(z))]\, d\zeta(z) + o(\|\delta f\|)$$

$$= \int_Z \nabla E'(f_\# \zeta)(f(z))^T \delta f(z)\, d\zeta(z) + \underbrace{O(\mathrm{Lip}(\nabla E'(f_\# \zeta)) \|\delta f\|^2) + o(\|\delta f\|)}_{o(\|\delta f\|)}$$

This shows $G(f + \delta f) - G(f) = DG(f)(\delta f) + o(\|\delta f\|)$. Hence the conclusion ∎

<u>Exemple</u>: Show if $W : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is symmetric and differentiable with a Lipschitz gradient, then $E(\mu) := \int W(x,y)\, d\mu(x)\, d\mu(y)$ satisfies the assumptions above with $E'(\mu) : x \mapsto \int W(x,y)\, d\mu(y)$.

Now we will introduce various divergences and study : (i) the "divergence property"

(ii) their weak continuity.

From now, <u>X is a compact metric space</u>.

II. Csiszár divergences (a.k.a. f-divergences)

Definition. Let $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be a convex function. For $\mu, \nu \in \mathcal{P}(X)$, let $\mu = \left(\frac{d\mu}{d\nu}\right)\nu + \mu^\perp$ be the Lebesgue decomposition. We define

$$D_f(\mu,\nu) = \int f\left(\frac{d\mu}{d\nu}\right) d\nu \quad + \quad f'_\infty(1) \cdot \mu^\perp(X)$$

where $\underbrace{f'_\infty(x)}_{\text{↳ recession/horizon of } f} := \lim_{t \to +\infty} f(tx)/t \in \mathbb{R} \cup \{\infty\}$.

Proposition: Let f be convex and such that $\min f = 0$ and $\operatorname{argmin} f = \{1\}$. Then $D_f(\mu,\nu) \geq 0$ with equality if and only if $\mu = \nu$.

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

Example (Kullback-Leibler divergence). Take $f(s) = \begin{cases} s\log s - s + 1 & \text{if } s > 0 \\ 1 & \text{if } s = 0 \\ +\infty & \text{if } s < 0 \end{cases}$

If $\mu \ll \nu$ then

$$D_f(\mu,\nu) = \int_X \left(\frac{d\mu}{d\nu} \log\left(\frac{d\mu}{d\nu}\right) - \frac{d\mu}{d\nu} + 1\right) d\nu = \int_X \log\left(\frac{d\mu}{d\nu}\right) d\mu = KL(\mu,\nu),$$

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

Example (Total variation). Take $f(s) = \begin{cases} |s-1| & \text{if } s \geq 0 \\ +\infty & \text{otherwise} \end{cases}$

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

$$D_f(\mu,\nu) = \int_X \left(\left|\frac{d\mu}{d\nu} - 1\right| d\nu + d\mu^\perp\right) \overset{(*)}{=} \int d|\mu - \nu| = |\mu - \nu|(X) =: \|\mu-\nu\|_{TV} = \sup_{f \in C(X)} \left\{\int f \, d(\mu-\nu) \, ; \, \|f\|_\infty \leq 1\right\}$$

where $(*)$ comes from the fact that $\begin{cases} (\mu-\nu)_+ = \max\{0, \frac{d\mu}{d\nu} - 1\}\nu + \mu^\perp \\ (\mu-\nu)_- = \max\{0, 1 - \frac{d\mu}{d\nu}\}\nu \end{cases}$.

In the context of generative models, a drawback is that D_f is not weakly continuous in general: for instance $D_f(\delta_x, \delta_y) = \begin{cases} 0 & \text{if } x = y \\ f'_\infty(1) & \end{cases}$ is discontinuous in general.

Proposition. If $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is convex, l.s.c and not identically $+\infty$, then $D_f(\mu,\nu)$ is (jointly) convex, weakly l.s.c and one has

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

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

Proof: see the lecture notes.

III Integral Probability Metrics (dual norms)

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

$$\|\alpha\|_B := \sup_{f \in B} \int_X f(x) \, d\alpha(x)$$

For $\mu, \nu \in \mathcal{P}(X)$, with $\alpha = \mu - \nu$, we define

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

This is called an "integral probability metric".

Proposition: If B is symmetric, bounded is sup-norm and contains 0, then $\|\cdot\|_B$ is a semi-norm on $\mathcal{M}(X)$, i.e it is non negative, positively 1-homogeneous and subadditive.

Proof: left as an exercise.

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

Example 2: Wasserstein-1 (W_1) is recovered with $B = \{ f \in \mathcal{C}(X) ; \mathrm{Lip}(f) \leq 1 \}$.

Example 3: The "flat norm" corresponds to

$$B = \{ f \in \mathcal{C}(X) ; \mathrm{Lip}(f) \leq 1 \text{ and } \|f\|_\infty \leq 1 \}$$

To "metrize" the weak convergence, $\mathcal{B}$ should not be too large nor too small.

<u>Proposition 3.5.</u>

(i) If $\mathcal{C}(X) \subset \overline{\mathrm{span}(\mathcal{B})}^{\|\cdot\|_\infty}$, i.e. the span of $\mathcal{B}$ is dense in $(\mathcal{C}(X), \|\cdot\|_\infty)$ then

$$\begin{cases} \|\alpha_k - \alpha\|_{\mathcal{B}} \to 0 \\ (\alpha_k) \text{ bounded for } \|\cdot\|_{TV} \end{cases} \text{ implies } \alpha_k \rightharpoonup \alpha$$

(ii) If $\mathcal{B} \subset \mathcal{C}(X)$ is compact then

$$\begin{cases} \alpha_k \rightharpoonup \alpha \\ \alpha_k \text{ bounded for } \|\cdot\|_{TV} \end{cases} \text{ implies } \|\alpha_k - \alpha\|_{\mathcal{B}} \to 0$$

<u>Proof:</u>

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

(ii) We assume that $\alpha_k \rightharpoonup \alpha$, consider a subsequence $(\alpha_{n_k})_k$ such that

$$\|\alpha_{n_k} - \alpha\|_{\mathcal{B}} \to \limsup \|\alpha_k - \alpha\|_{\mathcal{B}}$$

Since $\mathcal{B}$ is compact, let $f_{n_k} \in \mathcal{B}$ achieve the supremum defining $\|\alpha_{n_k} - \alpha\|_{\mathcal{B}}$. We again extract a subsequence $(f_{n'_k}) \xrightarrow{\|\cdot\|_\infty} f \in \mathcal{C}(X)$. One has:

$$\|\alpha_{n_k} - \alpha\|_{\mathcal{B}} = \langle \alpha_{n_k} - \alpha, f \rangle + \langle \alpha_n, f_{n_k} - f \rangle - \langle \alpha, f_{n_k} - f \rangle \to 0 \; \blacksquare$$

↳ This is a direct generalizat° of our proof of weak continuity of W_1 (in lecture 4).

IV Sinkhorn divergence

IV.1 Entropy Regularized optimal transport

Def (lecture 3). With $c \in \mathcal{C}(X \times X)$, let $\lambda \geq 0$ be the regularization, and

$$T_{c,\lambda}(\mu,\nu) := \min_{\gamma \in \Pi(\mu,\nu)} \int_{X^2} c(x,y)\, d\gamma(x,y) + \lambda \, KL(\gamma, \mu \otimes \nu)$$

differ by 1 from the one in lecture 3.

Reminders:

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

(optimality cdt°). There exists maximizers $(\varphi_\lambda, \psi_\lambda)$ and a unique minimizer γ_λ, linked by:

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

In particular, we have: $T_{c,\lambda}(\mu,\nu) = \int \varphi_\lambda \, d\mu + \int \psi_\lambda \, d\nu$.

IV.2 Is $T_{c,\lambda}$ a suitable divergence?

Proposition: For $\mu,\nu \in \mathcal{P}(X)$, $c \in \mathcal{C}(X \times X)$, it holds:

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

see lecture 3

lecture notes.

Moreover, $\gamma_\lambda \to \mu \otimes \nu$ as $\lambda \to +\infty$.

Proof: see lecture notes.

$x^* = \int y \, d\nu(y)$ if $c(x,y) = \|y-x\|_2^2$

Corollary: Let $\nu \in \mathcal{P}(X)$ be such that $\arg\min_{y \in X} \int c(x,y)\, d\nu(y)$ is a singleton $\{x^*\}$, and let

$$\mu_\lambda \in \arg\min_{\mu} T_{c,\lambda}(\mu,\nu).$$

Then as $\lambda \to +\infty$, one has $\mu_\lambda \to \delta_{x^*}$. (proof see lecture notes).

IV.3 Debiased quantity: the Sinkhorn divergence

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

$$\underbrace{S_{c,\lambda}(\mu,\nu)}_{\text{Sinkhorn divergence}} := T_{c,\lambda}(\mu,\nu) - \frac{1}{2} T_{c,\lambda}(\mu,\mu) - \frac{1}{2} T_{c,\lambda}(\nu,\nu)$$

Proposition (Interpolation properties). It holds, if $c(x,y) = \mathrm{dist}(x,y)^p$ for $p \geq 1$,

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

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

(Proof is immediate from the previous proposition).

Proposition. • If $k(x,y) = e^{-c(x,y)/\lambda}$ is a p.d. kernel, then

$S_\lambda(\mu,\nu) \geq 0$ with equality if $\mu = \nu$.

• If $e^{-c/\lambda}$ is furthermore a universal kernel, then

$S_\lambda(\mu_n,\mu) \to 0$ if and only if $\mu_n \to \mu$.