Lecture 6: Divergences between Probability measures

I Motivating problem: density fitting

- Fundamental problem: compare \( v \in \mathcal{P}(\mathbb{R}^d) \) arising from measurements to a model which is a parameterized family of distributions \( \{\mu_\theta; \theta \in \Theta\} \) where typically \( \Theta \in \mathbb{R}^k \).

- A suitable parameter can be obtained by minimizing:

\[
\min_{\theta \in \Theta} F(\theta) \text{ where } F(\theta) = D(\mu_\theta, v) \quad (\star)
\]

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:

\[
\begin{cases} 
D(\mu, v) \geq 0 \\
D(\mu, v) = 0 
\end{cases}
\]

Example 1: One can choose \( D(\mu, v) = W_p^p(\mu, v) \) for some \( p \geq 1 \).

When \( v \) is an empirical measure, with \( p = 2 \), the solution to (\( \star \)) is called the Minimum Kantorovich Estimator.

Example 2: Let \( x_1, \ldots, x_n \in \mathbb{R}^d \) are independent samples from \( v \). When \( \mu_0 \) has a density \( p_0 \) w.r.t a reference measure \( \sigma \), the maximum likelihood estimator (MLE) is

\[
\min_{\theta \in \Theta} -\frac{1}{n} \sum_{i=1}^{n} \log(p_0(x_i)) \quad (1)
\]

This corresponds to an empirical version of solving (\( \star \)) with \( D(\mu, v) = KL(\mu_\theta, v) \) since (1) converges to \( -\int \log(p_0(x))dv(x) = KL(\mu_0, v) - \int \log(\frac{dv}{d\sigma})dv \) (provided all the terms are finite).

- Note that the MLE fails:
- when there is no natural reference measure \( \sigma \)
- when \( p_0 \) is difficult to compute
- when the objective \( F \) is too complicated to minimize.
Generative models are when the parametric measure $\nu_0$ is given by

$$\nu_0 = (h_0) \# \nu$$

where $h_0 : \mathbb{R} \to \mathbb{R}$.

and where $\nu \in \mathcal{P}(\mathbb{R})$ is a reference measure. This leads to

$$F(\theta) = D((h_0) \# \nu, \nu).$$

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

- Initialize $\theta_0 \in \Theta$
- For $t = 1, 2, \ldots$ let $\theta_{t+1} = \theta_t - y \nabla F(\theta_t)$ where $y > 0$ is a step-size

Application to shape registration:

$D$ is the Sinkhorn divergence

$(h_0)_t$ is a parameterized set of diffeomorphisms.
Let us give a formula for $\nabla F(\theta)$ under strong regularity assumption.

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

**Proposition.** 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 $E'(\mu) \in C(\mathbb{R}^d)$ with $\nabla E'(\mu)$ Lipschitz, and such that $\forall \nu \in \mathcal{P}(\mathbb{R}^d)$

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

Assume moreover that $h : \mathbb{R}^d \to L^2(S_1; \mathbb{R}^d)$ is (Fréchet) differentiable, with partial derivatives of order 0 denoted $\partial_i h_0 \in L^2(S_1; \mathbb{R}^d)$. Then $F : \theta \mapsto E(h_0(\theta))$ is differentiable with gradient, for $i = 1, \ldots, p$,

$$[\nabla F(\theta)]_i = \int_{\mathbb{R}^d} \nabla E'(h_0(\theta)) (h_0(\theta))_{i} \partial_i h_0(\theta) \, d \mathcal{L}(\theta).$$

**Proof:** First we study $G : f \mapsto E(f \# \mathbf{1})$ and show that $G$ is (Fréchet) differentiable with differential:

$$DG(f)(g) = \int \nabla E'(f \# \mathbf{1}) (f \# \mathbf{1})_i g_i \, d \mathcal{L}(z).$$

Then the conclusion follows by the usual chain rule for Fréchet differentials.

For $f, g \in L^2(S_1; \mathbb{R}^d)$, we have that $W_2(f \# \mathbf{1}, (f + s g) \# \mathbf{1}) < \|s\|_{L^2(S_1)}$ by taking $(f, (f + s g))$ as an admissible transport plan. Thus,

$$E((f + s g) \# \mathbf{1}) - E(f \# \mathbf{1}) = \int \left[ E'(f \# \mathbf{1}) (f \# \mathbf{1})_i g_i + E'(f \# \mathbf{1}) (f \# \mathbf{1})_i s_i \right] \, d \mathcal{L}(z) + O(\|s\|_{L^2(S_1)}).$$

This shows $G(f + s g) - G(f) = DG(f)(g) + o(\|s\|_{L^2(S_1)})$. Hence the conclusion.

**Example:** Show if $W : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is symmetric and differentiable with a Lipschitz gradient, then $E(\mathbf{1}) = \int W(x, \cdot) \, d \mathbf{1}(x) \, dp(y)$ satisfies the assumption above with $E'(\mathbf{1}) : x \mapsto \int W(x, y) \, dp(y)$.

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

From now, $X$ is a compact metric space.

(ii) their weak continuity.
II. Csiszar divergences (a.k.a. f-divergences)

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

\[
D_f(\mu, \nu) = \int f\left(\frac{d\mu}{d\nu}\right) \, d\nu + f'(1) \cdot \mu^\perp(X)
\]

where \( f'_\infty(x) := \lim_{x \to \infty} f'(x)/x \in \mathbb{R}_+ \).

Proposition. Let \( f \) be convex and such that \( \min f = 0 \) and \( \max f = 1/2 \).

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 \( f'_\infty(1) \geq f'(2) > 0 \)) and \( \frac{d\mu}{d\nu} = 1 \in L^1(\nu) \) so \( \mu = \nu \).

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

If \( \mu \ll \nu \) then

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

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

\[
D_f(\mu, \nu) = \int_X \left( |\frac{d\mu}{d\nu} - 1| \, d\nu + 0 \right) = \int X |\mu - \nu| = 1_{\mu \neq \nu}(X) = \|\mu - \nu\|_1.
\]

Where \((*)\) comes from the fact that \( (\mu - \nu)_+ = \max\{0, \frac{d\mu}{d\nu} - 1\} \nu + \mu^\perp \)

\[
(\mu - \nu)_- = \max\{0, 1 - \frac{d\mu}{d\nu}\} \nu.
\]

In the context of generative models, a drawback is that \( D_f \) is not weakly continuous in general; for instance \( D_f(S_x, S_y) = \begin{cases} 0 & \text{if } x = y \\ f'_\infty(1) & \text{otherwise} \end{cases} \).
Proposition. If \( f : \mathbb{R} \to \mathbb{R} \cup \{\pm \infty\} \) is convex, l.s.c and not identically \(+\infty\), then \( D^*_g(\mu, \nu) \) is (jointly) convex, weakly l.s.c and one has
\[
D^*_g(\mu, \nu) = \sup_{\psi \in \mathcal{P}(\mathcal{X})} \int \psi \, d\mu + \int \psi \, d\nu \; \text{s.t.} \; \psi(x) + g^*(\psi(x)) < 0, \; \forall x \in \mathcal{X},
\]
where \( g^* : s \mapsto \sup_{u \in \mathbb{R}} u \cdot s - g(u) \) is the convex conjugate of \( g \).

Proof: see the lecture notes.

### III. Integral Probability Metrics (dual norms)

**Definition.** For a symmetric set \( \mathcal{B} \) of measurable functions \( X \to \mathbb{R} \) and \( \alpha \in C_b(X) \) a signed finite measure, let
\[
\| \alpha \|_{\mathcal{B}} := \sup_{f \in \mathcal{B}} \int f(x) \, d\alpha(x).
\]
For \( \mu, \nu \in \mathcal{P}(\mathcal{X}) \), with \( \alpha = \mu - \nu \), we define
\[
D_B(\mu, \nu) := \| \mu - \nu \|_{\mathcal{B}} = \sup_{f \in \mathcal{B}} \int f(x) \, d(\mu(x) - \nu(x)).
\]
This is called an “integral probability metric”.

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

Proof: left as an exercise.

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

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

**Example 3:** The “\( f \)-norm” corresponds to
\[
\mathcal{B} = \{ f \in C(X) : \text{Lip}(f) \leq 1 \text{ and } \| f \|_{\infty} \leq 1 \}.
\]
To "measurize" the weak convergence, \( B \) should not be too large nor too small.

**Proposition 3.5.**

(i) If \( \ell^1(B) \leq \text{span}(B) \), i.e. the span of \( B \) is dense in \( \ell^1 \), then

\[
\| \alpha_k - \alpha \|_B \to 0 \quad \text{implies} \quad \alpha_k \rightharpoonup \alpha
\]

(ii) If \( B \subset \ell^1 \) is compact then

\[
\| \alpha_k - \alpha \|_B \quad \text{implied} \quad \| \alpha_k - \alpha \|_B \to 0
\]

**Proof:**

(i) If \( \| \alpha_k - \alpha \|_B \to 0 \), then \( \forall \ f \in B \), since \( \langle f, \alpha_k - \alpha \rangle \leq \| \alpha_k - \alpha \|_B \)

so \( \langle f, \alpha_k \rangle \to \langle f, \alpha \rangle \). By linearity, this extends to \( \text{span}(B) \) and then to \( \text{span}(B) \).

(ii) We assume that \( \alpha_k \rightharpoonup \alpha \), consider a subsequence \( (\alpha_{k_n})_n \) such that

\[
\| \alpha_{k_n} - \alpha \|_B \to \lim \sup \| \alpha_{k_n} - \alpha \|_B
\]

Since \( B \) is compact, let \( f_{k_n} \in B \) achieve the supremum defining \( \| \alpha_{k_n} - \alpha \|_B \).

We again extract a subsequence \( (f_{k_n}) \rightharpoonup f \in \ell^1 \). One has:

\[
\| \alpha_{k_n} - \alpha \|_B = \langle \alpha_{k_n} - \alpha, f \rangle \to \langle \alpha, f \rangle - \langle \alpha, f_{k_n} - f \rangle \to 0
\]

L* This is a direct generalization of our proof of weak continuity of \( W_t \) in Lecture 4.1
Sinkhorn divergence

1. Entropy Regularized optimal transport

**Def (lecture 3).** With \( c \in E(X \times X) \), let \( \lambda > 0 \) be the regularization, and

\[
\mathcal{T}_{\lambda}(\mu, \nu) := \min_{\pi \in \Pi(\mu, \nu)} \int c(x, y) d\pi(x, y) + \lambda \; KL(\pi, \mu \otimes \nu)
\]

\( \pi \) differ by 1 from the one in Lecture 3.

**Reminders:**

(duality) \( \mathcal{T}_{\lambda}(\mu, \nu) = \sup_{\Psi, \Psi \in E(X)} \int \Psi d\mu + \int \Psi d\nu + \lambda \left( 1 - \int \int c(x, y) d\mu(x) d\nu(y) \right) \)

(optimality \( \mathcal{T}_{\lambda}(\mu, \nu) \) - There exists a maximizer \((\Psi_{\lambda}, \Psi_{\lambda})\) and a unique minimizer \( \pi_{\lambda} \), linked by:

\[
d\pi_{\lambda}(x, y) = e^{-\frac{\Psi_{\lambda}(x) + \Psi_{\lambda}(y) - c(x, y)}{\lambda}} \; d\mu(x) d\nu(y)
\]

In particular, we have: \( \mathcal{T}_{\lambda}(\mu, \nu) = \int \Psi_{\lambda} d\mu + \int \Psi_{\lambda} d\nu \).  

2. Is \( \mathcal{T}_{\lambda}(\mu, \nu) \) a suitable divergence?

**Proposition:** For \( \mu, \nu \in S(X) \), \( c \in E(X \times X) \), it holds:

\[
\mathcal{T}_{\lambda}(\mu, \nu) \to \begin{cases} 
\mathcal{T}(\mu, \nu) = \mathcal{T}_{0}(\mu, \nu) \text{ as } \lambda \to 0 \\
\int c(x, y) d\mu(x) d\nu(y) \text{ as } \lambda \to +\infty
\end{cases}
\]

Moreover, \( \pi_{\lambda} \to \mu \otimes \nu \) as \( \lambda \to +\infty \).

**Proof:** see lecture notes.

**Corollary:** Let \( \nu \in S(X) \) be such that \( \arg \min_{\nu \in S(X)} \int c(x, y) d\nu(y) \) is a singleton \( \{x^*\} \), and let \( \mu_\lambda = \arg \min_{\mu} \mathcal{T}_{\lambda}(\mu, \nu) \).

Then as \( \lambda \to +\infty \), one has \( \mu_\lambda \to \delta_{x^*} \). (Proof see lecture notes.)
III. 3 Debiased quantity: the Sinkhorn divergence

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

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

Sinkhorn divergence

**Proposition (Interpolation property):** It holds, if $c(x, y) = \text{dist}(x, y)^p$, for $p \geq 1$,

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

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

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

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

$$S_{k,\lambda}(\mu_n, \mu) \rightarrow 0 \text{ if and only if } \mu_n \rightarrow \mu.$$