# Optimal transport

## VI Divergences between Probability Measures

Seongho, Joo

Seoul National University

# Motivating problem: density fitting

• In statistics, imaging, or machine learning, one of the most fundamental problems is to compare a probability distribution $\nu \in \mathcal{P}(\mathsf{r}^d)$ 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

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

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

**Example.** One can choose $D(\mu, \nu) = W_p^p(\mu, \nu)$. When $\nu$ is an empirical measure and with $p = 2$, this is called *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.** (Maximum Likelihood) Let $x_1, \dots x_n \in \mathbb{R}^d$ be independent samples from $\nu$. When $\nu_\theta$ has a density $\rho_\theta$ with respect to a reference measure $\rho$ (e.g. the Lebesgue measure), the MLE is obtained by solving

$$\min -\frac{1}{n} \log(\rho_\theta(x_i)).$$

This corresponds to using an empirical counterpart of the Kullback-Leibler loss.

# Cont.

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

- when there is no natural reference measure $\sigma$;
- when the density $\rho_\theta$ is difficult to compute;
- the resulting objective $F$ is too complicated to minimize;

# 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. 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 $E : \mu \mapsto D(\mu, \nu)$.

# Cont.

**Proposition 1 (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 $E'(\mu) \in C^1(\mathsf{r}^d)$ with $\nabla E'(\mu)$ is Lipschiz, and such that for all $\nu \in \mathcal{P}_2(\mathsf{r}^d)$,

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

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

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

**Example.** If $W : \mathsf{r}^d \times \mathsf{r}^d \to \mathsf{r}$ is symmetric and diffrentiable with a Lipchitz gradient, then $E(\mu) := \int W(x, y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)$ satisfies the assumptions the proposition with $E'(\mu)(x) = \int W(x, y)\,\mathrm{d}\mu(y)$.

# Csiszár divergences

# 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.

**Definition 1 ($f$-divergence)**

Let $f : \mathsf{r} \to \mathsf{r} \cup \{+\infty\}$ be a convex funciton. 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 $\mu$ with respect to $\nu$. The divergence is defined by

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

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 $f$ is superlinear.

# Csiszár divergences

**Proposition 2**

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

**Example.** (Relative entropy). This is the $f$-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 $\mu << \nu$ then

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

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

# Csiszár divergences

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

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

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

$$D_f(\mu,\nu) = \int_X \left( |\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1| \, \mathrm{d}\nu + \mathrm{d}\mu^\perp \right) \overset{(\dagger)}{=} \int_X \mathrm{d}|\mu-\nu| = |\mu-\nu|(X)$$

where (†) comes from the fact that

$$(\mu-\nu)_+ = \max\{0, \mathrm{d}\mu/\mathrm{d}\nu - 1\}\nu + \mu^\perp, \quad (\mu-\nu)_- = \max\{0, 1 - \mathrm{d}\mu/\mathrm{d}\nu\}\nu$$

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

# Csiszár divergences

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

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

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

$$D_f(\mu,\nu) = \int_X \left( |\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1| \, \mathrm{d}\nu + \mathrm{d}\mu^\perp \right) \overset{(\dagger)}{=} \int_X \mathrm{d}|\mu-\nu| = |\mu-\nu|(X)$$

where (†) comes from the fact that

$$(\mu-\nu)_+ = \max\{0, \mathrm{d}\mu/\mathrm{d}\nu - 1\}\nu + \mu^\perp, \quad (\mu-\nu)_- = \max\{0, 1 - \mathrm{d}\mu/\mathrm{d}\nu\}\nu$$

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

**Proposition 3**

If $f : \mathsf{r} \to \mathsf{r} \cup \{+\infty\}$ is convex, lsc, and not indentically $+\infty$, then $D_f(\mu,\nu)$ is jointly convex and weakly lower-semicontinous and one has

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

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

# Integral probability metrics (dual norms)

# General case

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

$$\|\alpha\|_B := \sup_{f \in B} \int_X f(x) \, \mathrm{d}\alpha(x) \tag{2.1}$$

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

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

are often called "integral probability metrics" (or also "maximum mean discrepancy").

**Proposition 4**

If $B$ is symmetric, bounded in sup-norm contains 0, then $\|\cdot\|_B$ is a semi-norm on $\mathcal{M}(X)$.

**Example.** (Total variation) It is recovered with $B = \left\{ f \in C(X) \mid \|f\|_\infty \leq 1 \right\}$.

**Example.** (Wasserstein-1) It is the integral probability metric induced by the set of 1-Lipschitz function $B = \{f \in C(X) \,|\, \mathrm{Lip}(f) \leq 1\}$.

# Genreal case

**Example.** (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 note the case of $\|\cdot\|_{W_1}$, which is only finite for $\alpha$ 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.

$$B = \left\{ f \,|\, \mathsf{Lip}(f) \leq 1 \text{ and } \|f\|_\infty \leq 1 \right\}$$

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

$$B = \left\{ f \,|\, \mathsf{Lip}(f) + \|f\|_\infty \leq 1 \right\}$$

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

# Genreal case

**Example.** (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 note the case of $\|\cdot\|_{W_1}$, which is only finite for $\alpha$ 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.

$$B = \left\{ f \,|\, \mathrm{Lip}(f) \leq 1 \text{ and } \|f\|_\infty \leq 1 \right\}$$

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

$$B = \left\{ f \,|\, \mathrm{Lip}(f) + \|f\|_\infty \leq 1 \right\}$$

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

**Proposition 5**

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

1. If $C(X) \subset \overline{\mathrm{span}(B)}^{\|\cdot\|_\infty}$ (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 $\alpha_k \rightharpoonup \alpha$.
2. If $B \subset C(X)$ is compact (i.e. if it is closed, uniformly continuous and bounded) then $\alpha_k \rightharpoonup \alpha$ implies $\|\alpha_k - \alpha\|_B \to 0$.

# Kernel Maximum Mean Discrepancies

We now describe an important class of integral probability metrics.

**Definition 2 (Positive definite kernel)**

A symmteric function $k : X \times X \to \mathsf{r}$ is said to be positive definite (p.d.) if for any $n \geq 1$, for any family $x_1, \ldots, x_n \in X$ the matrix $(k(x_i, x_j))_{i,j}$ is positive semi-definite, i.e. for all $r \in \mathsf{r}^n$,

$$\sum_{i,j=1}^{n} r_i r_j k(x_i, x_j) \geq 0 \tag{2.2}$$

The kernel is said to be conditionally positive definite if Eq. 2.2 holds for all zero mean vector $r$, i.e. such that $\sum_i r_i = 0$.

# Kernel Maximum Mean Discrepancies

We now describe an important class of integral probability metrics.

**Definition 2 (Positive definite kernel)**

A symmteric function $k : X \times X \to \mathsf{r}$ is said to be positive definite (p.d.) if for any $n \geq 1$, for any family $x_1, \ldots, x_n \in X$ the matrix $(k(x_i, x_j))_{i,j}$ is positive semi-definite, i.e. for all $r \in \mathsf{r}^n$,

$$\sum_{i,j=1}^{n} r_i r_j k(x_i, x_j) \geq 0 \tag{2.2}$$

The kernel is said to be conditionally positive definite if Eq. 2.2 holds for all zero mean vector $r$, i.e. such that $\sum_i r_i = 0$.

**Definition 3 (MMD)**

Given a continuous and positive semi-definite kernel $k : X \times X \to \mathsf{r}$, we define for $\alpha \in \mathcal{M}(X)$ (finite signed Borel measure)

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

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

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

# Kernel Maximum Mean Discrepancies

The definition as a squared quantity makes snes thanks to the following result.

**Proposition 6**

If $k \in C(X^2)$ is conditionally p.d., $\int\int k \, \mathrm{d}\alpha \geq 0$ 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 $\alpha_n \otimes \alpha_n \rightharpoonup \alpha \otimes \alpha$, we have $0 \leq \int\int k \, \mathrm{d}\alpha_n \otimes \alpha_n \to \int\int k \, \mathrm{d}\alpha \otimes \alpha$. We show a link with dual norms via the following results.

# Kernel Maximum Mean Discrepancies

The definition as a squared quantity makes snes thanks to the following result.

**Proposition 6**

If $k \in C(X^2)$ is conditionally p.d., $\int\int k \, \mathrm{d}\alpha \geq 0$ 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 $\alpha_n \otimes \alpha_n \rightharpoonup \alpha \otimes \alpha$, we have $0 \leq \int\int k \, \mathrm{d}\alpha_n \otimes \alpha_n \to \int\int k \, \mathrm{d}\alpha \otimes \alpha$. We show a link with dual norms via the following results.

**Theorem 1 (Aronzsajn)**

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

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

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

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

# Kernel Maximum Mean Discrepancies

- It can be shown that if $k$ is universal (i.e. the first condition of Prop 5 holds), continuous and conditionally positive definite then $\|\cdot\|_k$ metrizes weak convergence in $\mathcal{P}(X)$. Examples of such kernels on $\mathrm{r}^d$ are:
  - the Gaussian kernel $k(x,y) = e^{-\frac{\|y-x\|}{2\sigma^2}}$ with $\sigma > 0$;
  - the distance kernel $k(x,y) = -\mathrm{dist}(x,y)$ (its MMD is called the "Energy distance");
- In the special case of discrete measure $\mu = \sum_{i=1}^m a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^n b_j \delta_{y_j}$ then we have

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

This requires $\mathcal{O}((m+n)^2)$ operations to compute.

# Sinkhorm divergences

# Reminders on entropy regularized OT

- Recall that $W_p^p$ are often good choices of divergence, but are computationally expansive. Next, we build divergences from entropy regularized optimal transport.

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

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

We recall the following facts from Lecture 3:

# Reminders on entropy regularized OT

- Recall that $W_p^p$ are often good choices of divergence, but are computationally expansive. Next, we build divergences from entropy regularized optimal transport.

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

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

We recall the following facts from Lecture 3:

- (Duality)

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

- (Optimality conditions) There exists maximizes $(\varphi_\lambda, \psi_\lambda)$ and a unique minimizer $\gamma_\lambda$ linked by the optimlaity 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

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

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

**Proposition 7 (Interpolation properties)**

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

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

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

One could imagine replacing Wasserstein by its entropy regularized version, but the previous result shows that when $\lambda$ 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 $\mu = \nu$.

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

**Proposition 7 (Interpolation properties)**

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

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

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

One could imagine replacing Wasserstein by its entropy regularized version, but the previous result shows that when $\lambda$ 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 1**

Let $\nu \in \mathcal{P}(X)$ be such that $\arg\min_{y \in X} \int c(x,y)\,\mathrm{d}\nu(y)$ is a signleton, 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 $\mu_\lambda \rightharpoonup \delta_{x^*}$.

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

# Debiased quantity: the Sinkhorn divergence

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

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

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 $\lambda$ is large.

# Debiased quantity: the Sinkhorn divergence

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

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

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 $\lambda$ is large.

**Proposition 8 (Interpolation properties)**

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

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

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

**Remark.** One can show that under the regularity assumptions over $\mu$ and $\nu$ and for the cost $c(x,y) = \|x-y\|_2^2$ on $\mathsf{r}^d$ that $\left|S_{c,\lambda} - T_c\right| = \mathcal{O}(\lambda^2)$. Finally, under assumptions on the cost, we can show that $\mu \mapsto S_{c,\lambda}(\mu,\nu)$ is minimized at $\nu$.

# Debiased quantity: the Sinkhorn divergence

**Proposition 9 (Positive semi-definiteness)**

If $k(x, y) = e^{-c(x,y)/\lambda}$ is a positive semi-definite kernel, then $S_{c,\lambda}(\mu, \nu) \geq 0$ with equality if $\mu = \nu$.