Exercises on statistical distances

Gabriel V. Cardoso
STIM
Ecole des Mines Paris(PSL)
Fontaineableau, France
gabriel.victorino_cardoso@minesparis.psl.eu

September 14, 2026

Notations

- We denote a sequence of vectors $(\mathbf{x}_1, \cdots, \mathbf{x}_n)$ by $\mathbf{x}_{1:n}$.
- We denote by $\mathrm{F}(\mathbb{R}^m, \mathbb{R}^n)$ the set of measurable functions from $\mathbb{R}^m$ to $\mathbb{R}^n$.
- Cartesian projection $\pi_i : \mathbb{R}^d \ni \mathbf{x} = (x_0, \cdots, x_1) \to \mathbf{x}_i \in \mathbb{R}$ and $\pi_{i:j}(\mathbf{x}) = (\pi_i(\mathbf{x}), \cdots, \pi_j(\mathbf{x}))$.
- $\mathcal{P}(\mathbb{R}^d)$ the set of probability measures over the Borel sigma algebra on $\mathbb{R}^d$, which we denote $\mathcal{B}(\mathbb{R}^d)$.
- Given $\mathbf{x} \in \mathbb{R}^d$ we define the dirac measure as

$$\delta_{\mathbf{x}} : \mathcal{B}(\mathbb{R}^d) \ni \mathrm{A} \to \begin{cases} 1 & \text{if} \quad \mathbf{x} \in \mathrm{A} \\ 0 & \text{if} \quad \mathbf{x} \notin \mathrm{A} \end{cases} . \tag{1}$$

- Given $\mathbf{x}_{1:n}$, we define the empirical (occupational) measure as

$$\hat{\mu}[\mathbf{x}_{1:n}] : \mathcal{B}(\mathbb{R}^d) \ni \mathrm{A} \to n^{-1} \sum_{i=1}^{n} \delta_{\mathbf{x}_i}(\mathrm{A}) . \tag{2}$$

- For $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $f \in \mathrm{F}(\mathbb{R}^d, \mathbb{R})$ we denote $\mu f = \int f(\mathbf{x}) \mu(\mathrm{d}\mathbf{x})$.
- For $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $T \in \mathrm{F}(\mathbb{R}^d, \mathbb{R}^m)$ we denote $T_{\#}\mu : \mathcal{B}(\mathbb{R}^m) \ni \mathrm{A} \to \mu(T^{-1}(\mathrm{A}))$.
- We say that $\mu \ll \nu$ if for every $\mathrm{A} \in \mathcal{B}(\mathbb{R}^d)$, $\nu(\mathrm{A}) = 0 \Rightarrow \mu(\mathrm{A}) = 0$
- For $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)^2$, we define $\mathcal{C}(\mu, \nu) = \{\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) | \pi_{1:d\#}\gamma = \mu \, , \pi_{d+1:2d\#}\gamma = \nu\}$.
- We say that the function

$$\mathsf{Q} : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to \mathbb{R}_+ \tag{3}$$

  is a Markov kernel if
  - For all $x \in \mathbb{R}^d$, $\mathsf{Q}(x, \cdot) : \mathcal{B}(\mathbb{R}^d) \to \mathbb{R}_+$ is a probability measure,
  - For all $\mathrm{A} \in \mathcal{B}(\mathbb{R}^d)$, $\mathsf{Q}(\cdot, \mathrm{A}) : \mathbb{R}^d \to \mathbb{R}_+$ is measurable.

  We define for $\mu \in \mathcal{P}(\mathbb{R}^d)$, the probability measure $\mu\mathsf{Q}(\mathrm{A}) = \int \mathsf{Q}(x, \mathrm{A}) \mu(\mathrm{d}x)$.
- We define $\Theta_{\mathcal{N}} = \mathbb{R}^d \times \mathrm{M}_{\geq 0}(\mathbb{R}^d)$, where $\mathrm{M}_{\geq 0}(\mathbb{R}^d)$ is the set of positive definite d square matrices.
- For m, Σ we define

$$\mathcal{N}(\mathbf{x}; \mathbf{m}, \Sigma) := \frac{1}{(2\pi|\Sigma|)^{d/2}} \exp\left(-\frac{(\mathbf{x} - \mathbf{m})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x} - \mathbf{m})}{2} \right) .$$

  We also define the measure

$$\mathcal{N}(\mathbf{m}, \Sigma) : \mathrm{A} \in \mathcal{B}(\mathbb{R}^d) \to \int 1_{\mathrm{A}}(\mathbf{x}) \mathcal{N}(\mathbf{x}; \mathbf{m}, \Sigma) \, \mathrm{d}\mathbf{x} .$$

1 Total variation distance

There exists two possible definitions of the total variation distance:

$$\|\mu-\nu\|_{\mathrm{TV}} := \sup_{\mathrm{A}_{n\,1:\infty}\in \mathrm{P}_{\mathbb{N}}} \sum_{n=1}^{\infty} |\mu(\mathrm{A}_n)-\nu(\mathrm{A}_n)|\,, \tag{4}$$

$$\mathrm{d}_{\mathrm{TV}}(\mu,\nu) := \sup_{\mathrm{A}\in\mathcal{B}(\mathbb{R}^d)} |\mu(\mathrm{A})-\nu(\mathrm{A})|\,, \tag{5}$$

where $\mathrm{P}_{\mathbb{N}} \subset \mathcal{B}(\mathbb{R}^d)$ is the set of countable partitions of $\mathbb{R}^d$. Prove the following assertions

1. Show that $\mathrm{d}_{\mathrm{TV}}(\mu,\nu) = \frac{1}{2}\|\mu-\nu\|_{\mathrm{TV}}$.
2. $\mathrm{d}_{\mathrm{TV}}(\mu,\nu)$ induces a metric over $\mathcal{P}(\mathbb{R}^d)$.
3. $\mathcal{P}(\mathbb{R}^d)$ is complete over this metric.

Solution: Consider $(\mu_n)_{n\in\mathbb{N}}$ a TV-Cauchy sequence. We define

$$\lambda : \mathcal{B}(\mathbb{R}^d) \ni \mathrm{A} \to \frac{1}{2}\sum_{n\in\mathbb{N}} 2^{-n}\mu_n(\mathrm{A})\,.$$

Then $\mu_n \ll \lambda$ for all n, and we denote $f_n = \frac{\mathrm{d}\mu_n}{\mathrm{d}\lambda}$. Then, $f_n \in \mathcal{L}_1(\lambda)$. Furthermore, it is a Cauchy sequence

$$\begin{aligned}\int |f_n - f_m|(\mathbf{x})\lambda(\mathrm{d}x) &= \int 1_{f_n\geq f_m}(f_n-f_m)(\mathbf{x})\lambda(\mathrm{d}x) + \int 1_{f_m\geq f_n}(f_m-f_n)(\mathbf{x})\lambda(\mathrm{d}x)\\ &= \mu_n(\{f_n\geq f_m\}) - \mu_m(\{f_n\geq f_m\}) + \mu_m(\{f_m\geq f_n\}) - \mu_n(\{f_m\geq f_n\}) \leq \|\mu_n-\mu_m\|_{\mathrm{TV}}\,.\end{aligned}$$

Thus, there exists $f_\infty \in \mathcal{L}_1(\lambda)$ such that $\lim_{n\to 0}\int |f_n - f_\infty|(\mathbf{x})\lambda(\mathrm{d}\mathbf{x}) = 0$. It is easy to see that

$$\mu_\infty : \mathcal{B}(\mathbb{R}^d) \ni \mathrm{A} \to \int 1_{\mathrm{A}}(\mathbf{x})\, f_\infty(\mathbf{x})\lambda(\mathrm{d}\mathbf{x})$$

is a *probability* measure. Since

$$\|\mu_\infty - \mu_n\|_{\mathrm{TV}} \leq \int |f_\infty - f_n|(\mathbf{x})\lambda(\mathrm{d}\mathbf{x})\,,$$

then $\lim_{n\to\infty}\|\mu_n-\mu_\infty\|_{\mathrm{TV}} = 0$.

4. Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space. For $n\in\mathbb{N}\cup\{\infty\}$, let $\mathbf{X}_n : (\Omega,\mathcal{F}) \to (\mathbb{R},\mathcal{B}(\mathbb{R}))$. If $\lim_{n\to\infty}\|\mathcal{L}(\mathbf{X}_n) - \mathcal{L}(\mathbf{X}_\infty)\|_{\mathrm{TV}} = 0$ then $\lim_{n\to\infty}\mathbf{X}_n \overset{w}{=} \mathbf{X}_\infty$.
5. Let $F_1 = \{f\in\mathcal{L}_\infty(\mathbb{R}^d) \mid \|f\|_\infty \leq 1\}$. Then

$$\|\mu-\nu\|_{\mathrm{TV}} = \sup_{f\in F_1} |\mu f - \nu f|\,. \tag{6}$$

Solution: Consider a measure λ that dominates both μ and ν. (For example, $\lambda = (\mu+\nu)/2$). Then, by the same reason as above,

$$|\mu(\mathrm{A})-\nu(\mathrm{A})| = \int 1_{\mathrm{A}}(\mathbf{x})(1_{\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}>\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}}(\mathbf{x}) - 1_{\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}<\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}}(\mathbf{x}))\left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}-\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\right)(\mathbf{x})\lambda(\mathrm{d}\mathbf{x})\,.$$

But

$$(1_{\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}>\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}}(\mathbf{x}) - 1_{\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}<\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}}(\mathbf{x}))\left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}-\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\right)(\mathbf{x}) = \left|\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}-\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\right|(\mathbf{x})\,,$$

which shows that

$$|\mu(\mathrm{A})-\nu(\mathrm{A})| = \int 1_{\mathrm{A}}(\mathbf{x})\left|\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}-\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\right|(\mathbf{x})\lambda(\mathrm{d}\mathbf{x})\,.$$

Now note that if $\|f\|_\infty \leq 1$,

$$|\mu f - \nu f| \leq \int \left| \frac{\mathrm{d}\mu}{\mathrm{d}\lambda} - \frac{\mathrm{d}\nu}{\mathrm{d}\lambda} \right| (\mathbf{x}) \lambda(\mathrm{d}\mathbf{x}) = \|\mu - \nu\|_{\mathrm{TV}} \,. \tag{7}$$

6. It holds that

$$\|\mu - \nu\|_{\mathrm{TV}} \leq \inf_{\gamma \in \mathcal{C}(\mu,\nu)} \int 1_{\mathbf{y} \neq \mathbf{x}}(\mathbf{x}, \mathbf{y}) \gamma(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}) \,. \tag{8}$$

There is an actual equality in (8). The proof is done by a constructive argument and the interested reader can look at [1, Theorem 19.1.6]

2 Kullback Leibler

We define

$$\mathrm{KL}\left(\mu \| \nu\right) := \begin{cases} \int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)(x) \mu(\mathrm{d}x) & \text{if } \mu \ll \nu \,, \\ \infty & . \end{cases} \tag{9}$$

Prove the following assertions

1. Prove that $\mathrm{KL}\left(\mu \| \nu\right) \geq 0$ and $\mathrm{KL}\left(\mu \| \nu\right) = 0 \iff \mu = \nu$.

Solution: We suppose that $\mu \ll \nu$. Otherwise, there is nothing to prove. Consider the function $x \to -\log(x)$, which is strictly convex on $\mathbb{R}_+$. By Jensen's inequality,

$$\begin{aligned} \mathrm{KL}\left(\mu \| \nu\right) = \int \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\mathbf{x}) \mu(\mathrm{d}\mathbf{x}) &= \int -\log \frac{1}{\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\mathbf{x})} \mu(\mathrm{d}\mathbf{x}) \geq -\log\left(\int \frac{1}{\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\mathbf{x})} \mu(\mathrm{d}\mathbf{x}) \right) \\ &= -\log\left(\int \frac{1}{\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\mathbf{x})} \mu(\mathrm{d}\mathbf{x}) \right) = -\log\left(\int \frac{1}{\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\mathbf{x})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\mathbf{x}) \nu(\mathrm{d}\mathbf{x}) \right) = -\log(1) = 0 \,. \end{aligned}$$

Jensen inequality for strictly convex functions yield an egality if and only if $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\mathbf{x}) = c$ almost surely. It is then easy to see that $c = 1$ and thus $\mu = \nu$.

1. If $\mu = \mathcal{N}\left(\mathbf{m}_0, \Sigma_0\right)$ and $\nu = \mathcal{N}\left(\mathbf{m}_1, \Sigma_1\right)$ then

$$\mathrm{KL}\left(\mu \| \nu\right) = \frac{1}{2}\left(\log\left(\frac{|\Sigma_1|}{|\Sigma_0|} \right) + \mathrm{Tr}\left(\Sigma_1^{-1}\Sigma_0\right) - d + (\mathbf{m}_1 - \mathbf{m}_0)^{\mathrm{T}} \Sigma_1^{-1} (\mathbf{m}_1 - \mathbf{m}_0) \right) .$$

Solution: Let $Z \sim \mathcal{N}\left(0, \mathrm{I}\right)$. Then,

$$\mathrm{KL}\left(\mu \| \nu\right) = \mathbb{E}\left[\log \mathcal{N}\left(\mathbf{m}_0 + \Sigma_0^{1/2} Z; \mathbf{m}_0, \Sigma_0 \right) \right] - \mathbb{E}\left[\log \mathcal{N}\left(\mathbf{m}_0 + \Sigma_0^{1/2} Z; \mathbf{m}_1, \Sigma_1 \right) \right] .$$

But we have that

$$\begin{aligned} \mathbb{E}\left[\log \mathcal{N}\left(\mathbf{m}_0 + \Sigma_0^{1/2} Z; \mathbf{m}_0, \Sigma_0 \right) \right] &= -\frac{1}{2}\left(\mathbb{E}\left[Z^{\mathrm{T}} \Sigma_0^{1/2} \Sigma_0^{-1} \Sigma_0^{1/2} Z \right] + d \log(2\pi) + \log\det \Sigma_0 \right) \\ &= -\frac{1}{2}\left(\mathbb{E}\left[Z^{\mathrm{T}} Z \right] + d \log(2\pi) + \log\det \Sigma_0 \right) . \end{aligned}$$

But $\mathbb{E}\left[Z^{\mathrm{T}} Z \right] = \mathbb{E}\left[\mathrm{Tr}\left(Z^{\mathrm{T}} Z \right) \right] = \mathbb{E}\left[\mathrm{Tr}\left(Z Z^{\mathrm{T}} \right) \right] = \mathrm{Tr}\left(\mathbb{E}\left[Z Z^{\mathrm{T}} \right] \right) = d$.
Similarly,

$$\begin{aligned} \mathbb{E}\left[\log \mathcal{N}\left(\mathbf{m}_0 + \Sigma_0^{1/2} Z; \mathbf{m}_1, \Sigma_1 \right) \right] = -\frac{1}{2}\Bigg(& \mathbb{E}\left[(\mathbf{m}_0 - \mathbf{m}_1 + Z\Sigma_0^{1/2})^{\mathrm{T}} \Sigma_1^{-1} (\mathbf{m}_0 - \mathbf{m}_1 + Z\Sigma_0^{1/2}) \right] \\ & + d \log(2\pi) + \log\det \Sigma_1 \Bigg) . \end{aligned}$$

But,

$$\mathbb{E}\left[(\mathbf{m}_0-\mathbf{m}_1+Z\Sigma_0^{1/2})^{\mathrm{T}}\Sigma_1^{-1}(\mathbf{m}_0-\mathbf{m}_1+Z\Sigma_0^{1/2})\right]=(\mathbf{m}_0-\mathbf{m}_1)^{\mathrm{T}}\Sigma_1^{-1}(\mathbf{m}_0-\mathbf{m}_1)+\mathbb{E}\left[(Z\Sigma_0^{1/2})^{\mathrm{T}}\Sigma_1^{-1}(Z\Sigma_0^{1/2})\right]$$
$$=(\mathbf{m}_0-\mathbf{m}_1)^{\mathrm{T}}\Sigma_1^{-1}(\mathbf{m}_0-\mathbf{m}_1)+\mathrm{Tr}\left(\Sigma_1^{-1}\Sigma_0\right)\,.$$

2. Let $\mathbf{m}_\mu=\int \mathbf{x}\mu(\mathrm{d}\mathbf{x})$ and $\Sigma_\mu=\int \mathbf{x}\mathbf{x}^{\mathrm{T}}\mu(\mathrm{d}\mathbf{x})-\mathbf{m}_\mu\mathbf{m}_\mu{}^{\mathrm{T}}$.

$$(\mathbf{m}_\mu,\Sigma_\mu)\in\arg\min_{m,\Sigma}\mathrm{KL}\left(\mu\|\mathcal{N}\left(\mathbf{m},\Sigma\right)\right)$$

3. (Data processing inequality) if $(\mu,\nu)\in\mathcal{P}(\mathbb{R}^d)^2$ and Q is a Markov Kernel, then

$$\mathrm{KL}\left(\mu\mathsf{Q}\|\nu\mathsf{Q}\right)\leq\mathrm{KL}\left(\mu\|\nu\right)\,. \tag{10}$$

(You can admit everything has a density with respect to the Lebesgue measure.)

4. Pinsker inequality:

$$2\,\mathrm{d_{TV}}(\mu,\nu)^2\leq\mathrm{KL}\left(\mu\|\nu\right)\,.$$

Tips: Use the data processing inequality with Markov Kernel $\mathsf{Q}(\mathbf{x},\mathrm{A})=\delta_{\mathbf{x}}\left(\mathrm{A}\right)$ and the fact that $2(p-q)^2\leq p\log(p/q)+(1-p)\log((1-p)/(1-q))$.

3 Wasserstein distance

The Wasserstein distance is defined by

$$\mathcal{W}_p^p(\mu,\nu):=\inf_{\gamma\in\mathcal{C}(\mu,\nu)}\int\|\mathbf{x}-\mathbf{y}\|^p\gamma(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{y})\,, \tag{11}$$

and $\mathcal{W}_p(\mu,\nu)=\left(\mathcal{W}_p^p(\mu,\nu)\right)^{1/p}$. We will admit the following theorems

Theorem 3.1: Existence of optimal transport

There exists $\gamma^\star\in\mathcal{C}(\mu,\nu)$ such that

$$\int\|\mathbf{x}-\mathbf{y}\|^p\gamma^\star(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{y})=\mathcal{W}_p^p(\mu,\nu)\,.$$

Theorem 3.2: The Gluing Lemma

Let $\mu,\nu,\rho\in\mathcal{P}(\mathbb{R}^d)^3$, $\gamma_{\mu,\nu}\in\mathcal{C}(\mu,\nu)$ and $\gamma_{\nu,\rho}\in\mathcal{C}(\nu,\rho)$. Then there is a law $\gamma\in\mathcal{P}(\mathbb{R}^{3d})$ such that $\pi_{1:2d\#}\gamma=\gamma_{\mu,\nu}$ and $\pi_{2d+1:3d\#}\gamma=\gamma_{\nu,\rho}$.

Prove the following assertions:

1. $\mathcal{W}_p(\mu,\nu)$ is a metric.
2. $p\leq q\Rightarrow\mathcal{W}_p(\mu,\nu)\leq\mathcal{W}_q(\mu,\nu)$
3. $\mathcal{W}_2^2(\mathcal{N}\left(\mathbf{m}_0,\Sigma_0\right),\mathcal{N}\left(\mathbf{m}_1,\Sigma_1\right))=\|\mathbf{m}_0-\mathbf{m}_1\|^2+\mathcal{W}_2^2(\mathcal{N}\left(0,\Sigma_0\right),\mathcal{N}\left(0,\Sigma_1\right))$.

Some other important theorems are the following

Theorem 3.3: Wasserstein metrizes weak convergence

The Wasserstein metric is complete. Furthermore, for $p\geq1$, for a sequence of random variables $\mathbf{X}_n:(\Omega,\mathcal{F})\to(\mathbb{R},\mathcal{B}(\mathbb{R}))$, $n\in\mathbb{N}\cup\{\infty\}$, such that $\mathcal{L}\left(\mathbf{X}_n\right)\in\mathcal{P}_p(\mathbb{R}^d)$ for all n, then

$$\lim_{n\to\infty}\mathbf{X}_n\overset{w}{=}\mathbf{X}_\infty\iff\mathcal{W}_p(\mathcal{L}\left(\mathbf{X}_n\right),\mathcal{L}\left(\mathbf{X}_\infty\right))\,.$$

Theorem 3.4: Wasserstein 1 duality

For $(\mu, \nu) \in \mathcal{P}_1(\mathbb{R}^d)$,

$$\mathcal{W}_1(\mu, \nu) = \sup_{\|f\|_{\text{Lip}}=1} \mu f - \nu f \,. \tag{12}$$

Theorem 3.5: Curse of dimensionality

Let $\mu = \mathcal{N}(0, \Sigma)$ in $\mathbb{R}^d$. Then, there exist some constant c such that for every n

$$\inf_{\mathbf{x}_{1:n} \in (\mathbb{R}^d)^n} \mathcal{W}_1(\mu, \hat{\mu}[\mathbf{x}_{1:n}]) \geq \frac{c}{n^{1/d}} \,. \tag{13}$$

4 Projected / Sliced Wasserstein distances

Due to the curse of dimensionality, the Wasserstein is difficult to use in high dimensions. A solution is to project in smaller dimensions $1 \leq k < d$

$$\overline{\mathcal{W}}_{p,k}(\mu, \nu) = \sup_{\mathrm{P} \in \mathrm{Gr}_k(d)} \mathcal{W}_p(\mathrm{P}_{\#}\mu, \mathrm{P}_{\#}\nu) \,, \tag{14}$$

where $\mathrm{Gr}_k(d)$ is the set of linear projections P into a linear subspace of $\mathbb{R}^d$ of dimension k.

1. Prove the following assertions
   (a) $\overline{\mathcal{W}}_{p,1}(\mu, \nu)$ is a metric.
   (b) Let $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)^2$ and $\mathbf{x}_{1:n} \in \mathbb{R}^{dn}$, $\mathbf{y}_{1:n} \in \mathbb{R}^{dn}$. Show that

$$\overline{\mathcal{W}}_{p,k}(\mu, \nu) \geq \overline{\mathcal{W}}_{p,k}(\hat{\mu}[\mathbf{x}_{1:n}], \hat{\mu}[\mathbf{y}_{1:n}]) - \overline{\mathcal{W}}_{p,k}(\mu, \hat{\mu}[\mathbf{y}_{1:n}]) - \overline{\mathcal{W}}_{p,1}(\hat{\mu}[\mathbf{x}_{1:n}], \nu) \,. \tag{15}$$

2. Admit the following theorem, shown in [2]:

   Theorem 4.1: Expectation of empirical projected Wasserstein

   Suppose that $\mu \in \mathcal{P}_{2p}(\mathbb{R}^d)$ for $p \geq 1$ satisfies the projection Poincaré inequality [2, Definition 3.2]. Then for all $n \geq 1$ and $k \leq p$

$$\mathbb{E}\left[\overline{\mathcal{W}}_{p,k}(\mu, \hat{\mu}[\mathbf{X}_{1:n}])\right] \leq C_0 n^{-\frac{1}{2p}} + C_1 n^{-\frac{1}{\max(2,p)}} \sqrt{dk \log(n)} + C_2 n^{-\frac{1}{p}} dk \log(n) \,, \tag{16}$$

   where $\mathbf{X}_{1:n}$ are i.i.d. according to μ.

   Using this theorem, propose a way of estimating if two datasets follow the same distribution.

5 To go further

Total variation distance : The total variation distance is extremely useful in probability, namely in the theory of Markov chains and Markov processes. It is a natural *norm* over the space of *signed measures* which makes the latter a Banach space. A good introduction for such framework is given in [1, Appendix D]

Wasserstein distance : The proof of the existence of optimal transport is rather technique and rely on the Prokorov compactness criteria. It can be found in [1, Theorem 20.1.1]. For the Gluing Lemma, it is a rather standard result in probability and can be found on most textbooks. A good reference for the results used here is [5, Chapter 6].

Projected / Slice Wasserstein : We followed mostly [4, 2] but we refer the interested reader to the discussion on those papers concerning other available results. A good introduction to the subject is the introduction chapter of [3].

6 References

References

[1] R. Douc et al. *Markov Chains*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, 2018. xviii+757. ISBN: 978-3-319-97703-4 978-3-319-97704-1. DOI: 10.1007/978-3-319-97704-1. URL: https://doi.org/10.1007/978-3-319-97704-1.

[2] Tianyi Lin et al. "On Projection Robust Optimal Transport: Sample Complexity and Model Misspecification". In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. PMLR, Mar. 2021, pp. 262–270. (Visited on 08/19/2025).

[3] Kimia Nadjahi. "Sliced-Wasserstein Distance for Large-Scale Machine Learning : Theory, Methodology and Extensions". Theses. Institut Polytechnique de Paris, Nov. 2021. DOI: 10.70675/d6f72967z029bz40daz9e60z4b2e4e312
URL: https://theses.hal.science/tel-03533097 (visited on 09/12/2026).

[4] François-Pierre Paty and Marco Cuturi. "Subspace Robust Wasserstein Distances". In: *Proceedings of the 36th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, May 24, 2019, pp. 5072–5081. URL: https://proceedings.mlr.press/v97/paty19a.html (visited on 09/12/2026).

[5] Cedric Villani. *Optimal Transport: Old and New*. Ed. by M. Berger et al. Grundlehren Der Mathematischen Wissenschaften 338. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. ISBN: 978-3-540-71049-3 978-3-540-71050-9. DOI: 10.1007/978-3-540-71050-9.