Limit Distributions for Regularized Wasserstein Distances on Finite Spaces

Third annual RTG 2088 Workshop

Marcel Klatt

September 26, 2018

Institute for Mathematical Stochastics
University of Göttingen

Computational burden of Wasserstein distances

In general, the computational cost to calculate the Wasserstein distance

$$W_p(r, s) := \left\{ \min_{\pi \in \Pi(r,s)} \sum_{i,j=1}^{N} d^p(x_i, y_j)\pi_{ij} \right\}^{1/p}$$

is of order $\mathcal{O}(N^3 \log(N))$.

Computational burden of Wasserstein distances

In general, the computational cost to calculate the Wasserstein distance

$$W_p(r, s) := \left\{ \min_{\pi \in \Pi(r,s)} \sum_{i,j=1}^{N} d^p(x_i, y_j)\pi_{ij} \right\}^{1/p}$$

is of order $\mathcal{O}(N^3 \log(N))$. There exist some workarounds, e.g.

- Exploiting the underlying metric structure (Ling & Okada (2007))
- Graph sparsification (Pele & Werman (2009))
- Specialized algorithms (Gottschlich & Schuhmacher (2014))
- Subsampling methods (Sommerfeld, Schrieber & Munk (2018))

$\vdots$

Computational burden of Wasserstein distances

In general, the computational cost to calculate the Wasserstein distance

$$W_p(r, s) := \left\{ \min_{\pi \in \Pi(r,s)} \sum_{i,j=1}^{N} d^p(x_i, y_j)\pi_{ij} \right\}^{1/p}$$

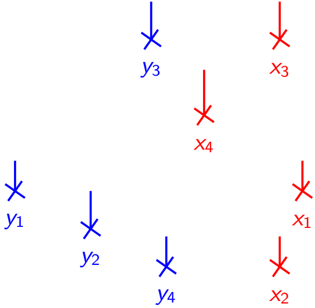
is of order $\mathcal{O}(N^3 \log(N))$. There exist some workarounds, e.g.

- Exploiting the underlying metric structure (Ling & Okada (2007))
- Graph sparsification (Pele & Werman (2009))
- Specialized algorithms (Gottschlich & Schuhmacher (2014))
- Subsampling methods (Sommerfeld, Schrieber & Munk (2018))

⋮

- **Regularization methods** (Cuturi (2013), Dessein et al. (2016))

Regularized Wasserstein distance

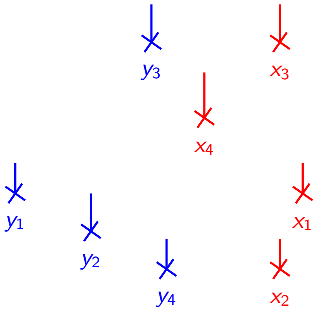


Basic idea:

- Let $E\colon \mathbb{R}^{N\times N} \to \mathbb{R}$ be the entropy

$$E(\pi) := \begin{cases} -\sum_{i,j=1}^{N} \pi_{ij} \log(\pi_{ij}), & \pi \in \mathbb{R}_{+}^{N\times N} \\ +\infty, & \text{else.} \end{cases}$$

Regularized Wasserstein distance



Basic idea:

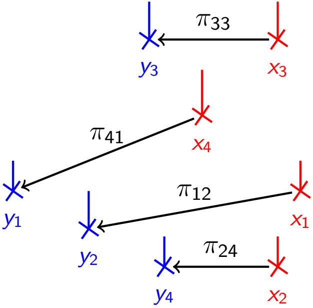
- Let $E\colon \mathbb{R}^{N\times N} \to \mathbb{R}$ be the entropy

$$E(\pi) := \begin{cases} -\sum_{i,j=1}^{N} \pi_{ij}\log(\pi_{ij}), & \pi \in \mathbb{R}_{+}^{N\times N} \\ +\infty, & \text{else.} \end{cases}$$

- For $\lambda > 0$, find the (*unique*) entropy regularized transport plan

$$\pi_\lambda(r,s) = \operatorname*{argmin}_{\pi\in\Pi(r,s)} \sum_{i,j=1}^{N} d^p(x_i,x_j)\pi_{ij} - \lambda E(\pi)\,.$$

Regularized Wasserstein distance



$\lambda = 0$

Basic idea:

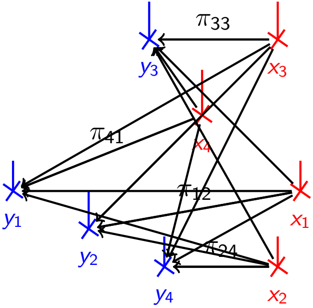
- Let $E \colon \mathbb{R}^{N \times N} \to \mathbb{R}$ be the entropy

$$E(\pi) := \begin{cases} -\sum_{i,j=1}^{N} \pi_{ij} \log(\pi_{ij}), & \pi \in \mathbb{R}_+^{N \times N} \\ +\infty, & \text{else.} \end{cases}$$

- For $\lambda > 0$, find the (*unique*) entropy regularized transport plan

$$\pi_\lambda(r, s) = \operatorname*{argmin}_{\pi \in \Pi(r,s)} \sum_{i,j=1}^{N} d^p(x_i, x_j)\pi_{ij} - \lambda E(\pi) .$$

Regularized Wasserstein distance



λ large

Basic idea:

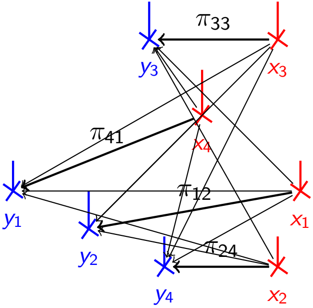
- Let $E\colon \mathbb{R}^{N\times N} \to \mathbb{R}$ be the entropy

$$E(\pi) := \begin{cases} -\sum_{i,j=1}^{N} \pi_{ij} \log(\pi_{ij}), & \pi \in \mathbb{R}_{+}^{N\times N} \\ +\infty, & \text{else.} \end{cases}$$

- For $\lambda > 0$, find the (*unique*) entropy regularized transport plan

$$\pi_\lambda(r,s) = \underset{\pi\in\Pi(r,s)}{\operatorname{argmin}} \sum_{i,j=1}^{N} d^p(x_i,x_j)\pi_{ij} - \lambda E(\pi)\,.$$

Regularized Wasserstein distance



λ intermediate

Basic idea:

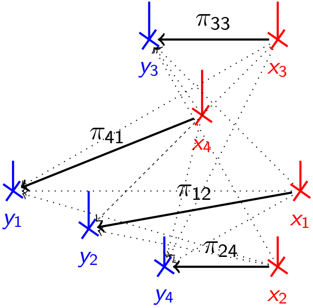
- Let $E\colon \mathbb{R}^{N\times N} \to \mathbb{R}$ be the entropy

$$E(\pi) := \begin{cases} -\sum_{i,j=1}^{N} \pi_{ij}\log(\pi_{ij}), & \pi \in \mathbb{R}_+^{N\times N} \\ +\infty, & \text{else.} \end{cases}$$

- For $\lambda > 0$, find the (*unique*) entropy regularized transport plan

$$\pi_\lambda(r,s) = \operatorname*{argmin}_{\pi\in\Pi(r,s)} \sum_{i,j=1}^{N} d^p(x_i,x_j)\pi_{ij} - \lambda E(\pi)\,.$$

Regularized Wasserstein distance



λ small

Basic idea:

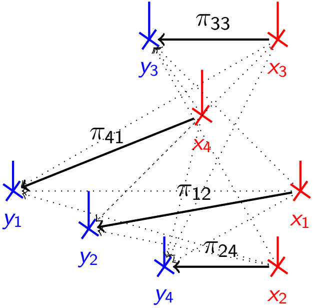
- Let $E\colon \mathbb{R}^{N\times N} \to \mathbb{R}$ be the entropy

$$E(\pi) := \begin{cases} -\sum_{i,j=1}^{N} \pi_{ij} \log(\pi_{ij}), & \pi \in \mathbb{R}_+^{N\times N} \\ +\infty, & \text{else.} \end{cases}$$

- For $\lambda > 0$, find the (*unique*) entropy regularized transport plan

$$\pi_\lambda(r, s) = \operatorname*{argmin}_{\pi\in\Pi(r,s)} \sum_{i,j=1}^{N} d^p(x_i, x_j)\pi_{ij} - \lambda E(\pi)\,.$$

Regularized Wasserstein distance



λ small

Basic idea:

- Let $E\colon \mathbb{R}^{N\times N} \to \mathbb{R}$ be the entropy

$$E(\pi) := \begin{cases} -\sum_{i,j=1}^{N} \pi_{ij}\log(\pi_{ij}), & \pi \in \mathbb{R}_+^{N\times N} \\ +\infty, & \text{else.} \end{cases}$$

- For $\lambda > 0$, find the (*unique*) entropy regularized transport plan

$$\pi_\lambda(r,s) = \operatorname*{argmin}_{\pi\in\Pi(r,s)} \sum_{i,j=1}^{N} d^p(x_i,x_j)\pi_{ij} - \lambda E(\pi)\,.$$

Regularized Wasserstein distance

For $\lambda > 0$, define the regularized Wasserstein distance as

$$W_{\lambda,p}(r,s) := \left\{ \sum_{i,j=1}^{N} d^p(x_i,x_j)\pi_\lambda(r,s)_{ij} \right\}^{1/p}\,.$$

Why entropic regularization?

$$\min_{\pi \in \Pi(r,s)} \sum_{i,j=1}^{N} d^p(x_i, x_j)\pi_{ij} - \lambda E(\pi) \tag{1}$$

Introducing two dual variables $f, g \in \mathbb{R}^N$ for each marginal constraint, the Lagrangian of (1) reads

$$\mathcal{L}(\pi, f, g) = \langle \pi,\, d^p \rangle - \lambda E(\pi) - \langle f,\, \pi \mathbb{1}_N - r \rangle - \langle g,\, \pi^T \mathbb{1}_N - s \rangle\,.$$

Why entropic regularization?

$$\min_{\pi \in \Pi(r,s)} \sum_{i,j=1}^{N} d^p(x_i, x_j)\pi_{ij} - \lambda E(\pi) \qquad (1)$$

Introducing two dual variables $f, g \in \mathbb{R}^N$ for each marginal constraint, the Lagrangian of (1) reads

$$\mathcal{L}(\pi, f, g) = \langle \pi,\, d^p \rangle - \lambda E(\pi) - \langle f,\, \pi \mathbb{1}_N - r \rangle - \langle g,\, \pi^T \mathbb{1}_N - s \rangle \,.$$

Considering first order conditions results in

$$\pi = \operatorname{diag}(u)\, K \operatorname{diag}(v)$$

with

$$u := \exp\left(\frac{f}{\lambda}\right),\ K := \exp\left(-\frac{d^p}{\lambda}\right),\ v := \exp\left(\frac{g}{\lambda}\right)\,.$$

Why entropic regularization?

The dual variables u, v must satisfy the following equations which correspond to the mass conservation constraints inherent to $\Pi(r, s)$,

$$\mathrm{diag}(u)\,K\,\mathrm{diag}(v)\mathbb{1}_N = r, \quad \mathrm{diag}(v)\,K^T\,\mathrm{diag}(u)\mathbb{1}_N = s\,.$$

Why entropic regularization?

The dual variables u, v must satisfy the following equations which correspond to the mass conservation constraints inherent to $\Pi(r, s)$,

$$\mathrm{diag}(u)\, K\, \mathrm{diag}(v) \mathbb{1}_N = r, \quad \mathrm{diag}(v)\, K^T\, \mathrm{diag}(u) \mathbb{1}_N = s\,.$$

That problem is known as the matrix scaling problem and is solved iteratively, starting with $v^{(0)} = \mathbb{1}_N$ and updates

$$u^{(l+1)} := \frac{r}{Kv^{(l)}}, \quad v^{(l+1)} := \frac{s}{K^T u^{(l+1)}}\,.$$

These updates define **Sinkhorn's algorithm**.

Statistical framework

Let $\mathcal{X} = \{x_1, \ldots, x_N\}$ be a finite space with metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$.
Assume, we only have access to the measure r through its corresponding empirical version

$$\hat{r}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

derived by a sample of $\mathcal{X}$-valued random variables $X_1, \ldots, X_n \overset{i.i.d.}{\sim} r$.

Statistical framework

Let $\mathcal{X} = \{x_1, \ldots, x_N\}$ be a finite space with metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. Assume, we only have access to the measure r through its corresponding empirical version

$$\hat{r}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

derived by a sample of $\mathcal{X}$-valued random variables $X_1, \ldots, X_n \overset{i.i.d.}{\sim} r$.

Central question:

- How do the random quantities $\pi_\lambda(\hat{r}_n, s)$ and $W_{\lambda,p}(\hat{r}_n, s)$ relate to $\pi_\lambda(r, s)$ and $W_{\lambda,p}(r, s)$, respectively?

Limit laws for empirical regularized transport plans

The empirical regularized transport plan is defined as

$$\pi_\lambda(\hat{r}_n, s) = \underset{\pi \in \Pi(\hat{r}_n, s)}{\arg\min} \sum_{i,j=1}^{N} d^p(x_i, x_j)\pi_{ij} - \lambda E(\pi)\,.$$

Limit laws for empirical regularized transport plans

The empirical regularized transport plan is defined as

$$\pi_\lambda(\hat{r}_n, s) = \underset{\pi \in \Pi(\hat{r}_n, s)}{\arg\min} \sum_{i,j=1}^{N} d^p(x_i, x_j)\pi_{ij} - \lambda E(\pi)\,.$$

Theorem (K., Tameling & Munk (2018+))

With the sample size n approaching infinity, it holds for $r = s$ and $r \neq s$ that

$$\sqrt{n}\,\{\pi_\lambda(\hat{r}_n, s) - \pi_\lambda(r, s)\} \xrightarrow{\mathfrak{D}} \mathcal{N}_{N^2}(0, \Sigma_\lambda(r|s))\,.$$

Limit laws for empirical regularized transport plans

The empirical regularized transport plan is defined as

$$\pi_\lambda(\hat{r}_n, s) = \underset{\pi \in \Pi(\hat{r}_n, s)}{\arg\min} \sum_{i,j=1}^{N} d^p(x_i, x_j)\pi_{ij} - \lambda E(\pi)\,.$$

Theorem (K., Tameling & Munk (2018+))

With the sample size n approaching infinity, it holds for $r = s$ and $r \neq s$ that

$$\sqrt{n}\,\{\pi_\lambda(\hat{r}_n, s) - \pi_\lambda(r, s)\} \xrightarrow{\mathfrak{D}} \mathcal{N}_{N^2}(0, \Sigma_\lambda(r|s))\,.$$

Remark

Limit distributions for the (non-regularized) transport plan ($\lambda = 0$) are not known.

Proof strategy:

- We think of $\pi_\lambda(r, s)$ as a vector and consider the **functional**

$$\phi_\lambda \colon (r, s) \mapsto \underset{\pi \in \mathbb{R}^{N^2}}{\arg\min} \; \langle d^p, \pi \rangle - \lambda E(\pi)$$

$$\text{s.t.} \quad A_\star \pi = \Big[r \, , \, s_\star \Big]^T .$$

Advantage to (non-regularized) OT: Uniqueness of $\pi_\lambda(r, s)$

Proof strategy:

- We think of $\pi_\lambda(r, s)$ as a vector and consider the **functional**

$$\phi_\lambda \colon (r, s) \mapsto \underset{\pi \in \mathbb{R}^{N^2}}{\arg\min} \; \langle d^p, \pi \rangle - \lambda E(\pi)$$
$$\text{s.t.} \quad A_\star \pi = \left[r \, , \, s_\star \right]^T .$$

  Advantage to (non-regularized) OT: $\boxed{\text{Uniqueness of } \pi_\lambda(r, s)}$

- **Sensitivity analysis** of the optimal solution

Proof strategy:

- We think of $\pi_\lambda(r, s)$ as a vector and consider the **functional**

$$\phi_\lambda \colon (r, s) \mapsto \underset{\pi \in \mathbb{R}^{N^2}}{\arg\min} \ \langle d^p, \pi \rangle - \lambda E(\pi)$$
$$\text{s.t.} \quad A_\star \pi = \left[r \, , \, s_\star \right]^T .$$

Advantage to (non-regularized) OT: Uniqueness of $\pi_\lambda(r, s)$

- **Sensitivity analysis** of the optimal solution
  - ▷ State optimality conditions for $\pi_\lambda(r, s)$ (a.k.a. KKT-conditions)
  - ▷ Apply the implicit function theorem
  - ⇒ The function ϕ_λ is **differentiable**

Advantage to (non-regularized) OT: Non-Sparsity of $\pi_\lambda(r, s)$

Proof strategy:

- We think of $\pi_\lambda(r, s)$ as a vector and consider the **functional**

$$\phi_\lambda \colon (r, s) \mapsto \underset{\pi \in \mathbb{R}^{N^2}}{\arg\min} \ \langle d^p, \pi \rangle - \lambda E(\pi)$$
$$\text{s.t.} \quad A_\star \pi = \left[r \, , \, s_\star \right]^T .$$

Advantage to (non-regularized) OT: Uniqueness of $\pi_\lambda(r, s)$

- **Sensitivity analysis** of the optimal solution
  - ▷ State optimality conditions for $\pi_\lambda(r, s)$ (a.k.a. KKT-conditions)
  - ▷ Apply the implicit function theorem
  - $\Rightarrow$ The function ϕ_λ is **differentiable**

Advantage to (non-regularized) OT: Non-Sparsity of $\pi_\lambda(r, s)$

- Apply (multivariate) **delta method**

The covariance matrix $\Sigma_\lambda(r|s)$

According to the **implicit function theorem** we obtain that

$$\nabla\phi_\lambda(r, s) = D A_\star^T [A_\star\, D\, A_\star^T]^{-1} \,.$$

- $A_\star$ is the coefficient matrix encoding the marginal constraints
- D is a diagonal matrix with diagonal $\pi_\lambda(r, s)$

The covariance matrix $\Sigma_\lambda(r|s)$

According to the **implicit function theorem** we obtain that

$$\nabla\phi_\lambda(r,s) = DA_\star^T[A_\star\, D\, A_\star^T]^{-1}\,.$$

- $A_\star$ is the coefficient matrix encoding the marginal constraints
- D is a diagonal matrix with diagonal $\pi_\lambda(r,s)$

Hence, the (multivariate) delta method tells us that

$$\Sigma_\lambda(r|s) = \nabla_r\phi_\lambda(r,s)\,\Sigma(r)\,\nabla_r\phi_\lambda(r,s)^T\,.$$

Limit laws for empirical regularized Wasserstein distances

The empirical regularized OT-distance is defined as

$$W_{\lambda,p}(\hat{r}_n, s) := \left\{ \sum_{i,j=1}^{N} d^p(x_i, x_j) \pi_\lambda(\hat{r}_n, s)_{ij} \right\}^{1/p} .$$

Limit laws for empirical regularized Wasserstein distances

The empirical regularized OT-distance is defined as

$$W_{\lambda,p}(\hat{r}_n, s) := \left\{ \sum_{i,j=1}^{N} d^p(x_i, x_j) \pi_\lambda(\hat{r}_n, s)_{ij} \right\}^{1/p} .$$

Theorem (K., Tameling & Munk (2018+))

With the sample size n approaching infinity, it holds for $r = s$ and $r \neq s$ that

$$\sqrt{n}\left\{W_{\lambda,p}(\hat{r}_n, s) - W_{\lambda,p}(r, s)\right\} \xrightarrow{\mathfrak{D}} \mathcal{N}_1(0, \sigma_\lambda^2(r|s)) .$$

Finite sample performance

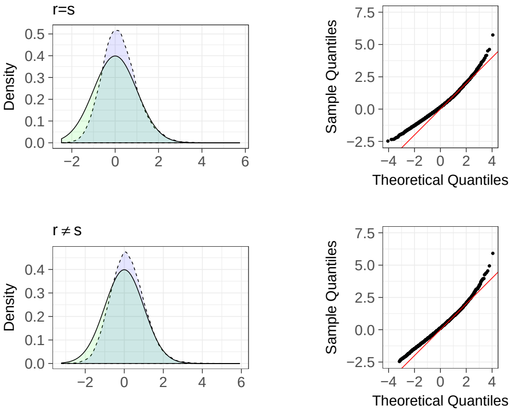


Figure 1: Density and Q-Q-plots in the one-sample case for $r = s$ and $r \neq s$. Comparison of the finite Sinkhorn divergence sample distribution on a regular grid of size 10×10 with regularization parameter $\lambda = 2q_{50}(d)$ and sample sizes $n = 25$ to the standard normal distribution.

Finite sample performance

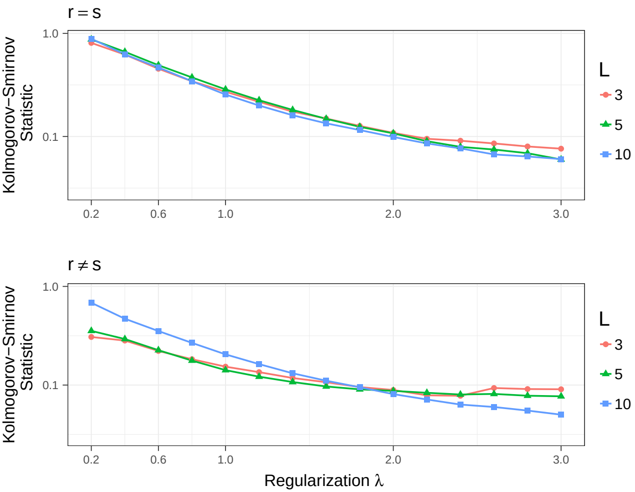


Figure 2: Kolmogorov-Smirnov distance on a logarithmic scale between the finite sample distribution ($n = 25$) and the theoretical normal distribution averaged over five measures.

Summary: Wasserstein vs. regularized Wasserstein

- **Different limit laws** under equality of measures (non-normal vs. normal)

Wasserstein	regularized Wasserstein
$n^{1/2p} W_p(\hat{r}_n, r)$	$\sqrt{n}\,\{W_{\lambda,p}(\hat{r}_n, r) - W_{\lambda,p}(r, r)\}$
$\downarrow \mathfrak{D}$	$\downarrow \mathfrak{D}$
$\left\{ \max_{f \in \Phi^*(r,r)} \langle \mathbf{G}, f \rangle \right\}^{1/p}$	$\mathcal{N}_1(0, \sigma^2_\lambda(r \mid r))$

Summary: Wasserstein vs. regularized Wasserstein

- **Different limit laws** under equality of measures (non-normal vs. normal)

Wasserstein	regularized Wasserstein
$n^{1/2p} W_p(\hat{r}_n, r)$	$\sqrt{n}\,\{W_{\lambda,p}(\hat{r}_n, r) - W_{\lambda,p}(r, r)\}$
$\downarrow \mathfrak{D}$	$\downarrow \mathfrak{D}$
$\left\{\max_{f \in \Phi^*(r,r)} \langle \mathbf{G}, f \rangle\right\}^{1/p}$	$\mathcal{N}_1(0, \sigma^2_\lambda(r\|r))$

- **Different scaling behavior**, i.e., for regularized Wasserstein the scaling behavior is independent of p

Summary: Wasserstein vs. regularized Wasserstein

- **Different limit laws** under equality of measures (non-normal vs. normal)

Wasserstein	regularized Wasserstein
$n^{1/2p} W_p(\hat{r}_n, r)$	$\sqrt{n}\,\{W_{\lambda,p}(\hat{r}_n, r) - W_{\lambda,p}(r, r)\}$
$\big\downarrow \mathfrak{D}$	$\big\downarrow \mathfrak{D}$
$\left\{\max_{f \in \Phi^*(r,r)} \langle \mathbf{G}, f\rangle\right\}^{1/p}$	$\mathcal{N}_1(0, \sigma_\lambda^2(r\|r))$

- **Different scaling behavior**, i.e., for regularized Wasserstein the scaling behavior is independent of p
- **Degeneracy**, i.e.

$$\lim_{\lambda \searrow 0} \sigma_\lambda^2(r|r) = 0\,.$$

Application and future work

- ? Statistical inference (e.g. How to apply the limit law for the regularized transport plan?)
- ? Similar approach for regularized Wasserstein barycenters?