Faster Wasserstein Distance Estimation with the Sinkhorn Divergence

Lénaïc Chizat[1], joint work with Pierre Roussillon[2], Flavien Léger[2], François-Xavier Vialard[3] and Gabriel Peyré[2]

July 8th, 2020 - Optimal Transport: Regularization and Applications

[1]CNRS and Université Paris-Sud [2]ENS Paris [3]Université Gustave Eiffel

Optimal Transport & Entropic Regularization

Statistical Optimal Transport

Estimation of the Squared Wasserstein Distance

Let μ and ν be probability densities on the unit ball in $\mathbb{R}^d$. Given

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \quad \text{and} \quad \hat{\nu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}$$

empirical distributions of n independent samples, estimate

$$W_2^2(\mu,\nu) := \min_{\gamma\in\Pi(\mu,\nu)} \int \|y-x\|_2^2 \,\mathrm{d}\gamma(x,y),$$

where $\Pi(\mu,\nu)$ is the set of transport plans*.

*Set of probability distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with respective marginals μ and ν.

How does entropic regularization help for this task?

[Refs for other approaches]:

Forrow et al. (2019). *Statistical optimal transport via factored couplings.*

Hütter, Rigollet (2019). *Minimax rates of estimation for smooth optimal transport maps.*

Niles-Weed, Berthet (2019). *Estimation of smooth densities in Wasserstein distance.*

Niles-Weed, Rigollet (2019). *Estimation of Wasserstein distances in the spiked transport model.*

Plug-in estimator

Theorem (CRLVP'20)

$$\mathbf{E}\left[\left|W_2^2(\hat{\mu}_n, \hat{\nu}_n) - W_2^2(\mu, \nu)\right|\right] \lesssim \begin{cases} n^{-2/d} & \textit{if } d > 4, \\ n^{-1/2}\log(n) & \textit{if } d = 4, \\ n^{-1/2} & \textit{if } d < 4. \end{cases}$$

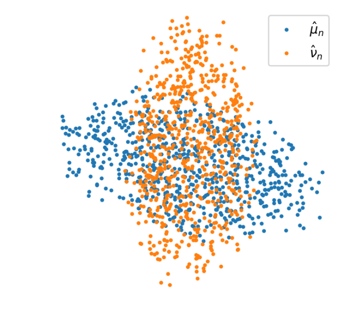
Proof idea. Bound $|\hat{W}_2^2 - W_2^2|$ by the supremum of an empirical process over convex 1-Lipschitz functions (Brenier). Then apply Dudley's chaining and Bronshtein's bound on the covering number.

Corollary

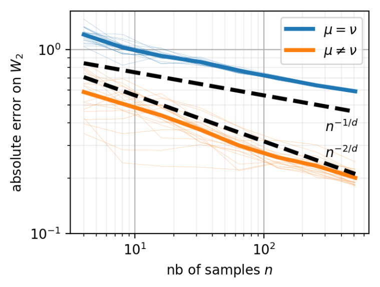
- If $W_2(\mu, \nu) \geq \alpha > 0$, same error bounds $\times \frac{1}{\alpha}$ for $W_2(\hat{\mu}_n, \hat{\nu}_n)$
- Faster than the rate $n^{-1/d}$ (which is when $\mu = \nu$)

Numerical illustration

Performance of the plug-in estimator $\hat{W}_{2,n} = W_2(\hat{\mu}_n, \hat{\nu}_n)$



Elliptically contoured distributions with compact support ($d = 2$)

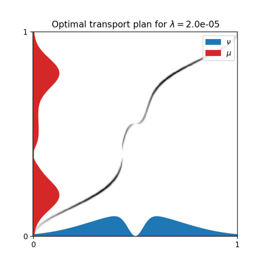


Estimation error on W_2 ($d = 8$)

Entropy Regularized Optimal Transport

Let $\lambda \geq 0$ and $H(\mu, \nu) = \int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\mu$ be the relative entropy.

$$T_\lambda(\mu, \nu) := \min_{\gamma \in \Pi(\mu,\nu)} \int \|y - x\|_2^2 \, \mathrm{d}\gamma(x, y) + 2\lambda H(\gamma, \mu \otimes \nu)$$



- a.k.a. the *Schrödinger bridge*
- favors diffuse solutions
- increases stability
- the higher λ, the easier to solve

Proposition (Dvurechensky et al., builds on Altschuler et al.,)

Sinkhorn's algo. computes $T_\lambda(\hat{\mu}_n, \hat{\nu}_n)$ to ϵ-accuracy in time $O(n^2 \lambda^{-1} \epsilon^{-1})$.

[Refs]:

Altschuler, Niles-Weed, Rigollet (2017). *Near-linear time approximation algorithms for optimal transport [...].*

Dvurechensky, Gasnikov, Kroshnin (2018). *Computational optimal transport [...]*

Discrete optimal transport via Sinkhorn

Shortcuts: $\hat{T}_{\lambda,n} = T_\lambda(\hat{\mu}_n, \hat{\nu}_n)$, $\hat{W}^2_{2,n} = W_2^2(\hat{\mu}_n, \hat{\nu}_n)$, $W_2^2 = W_2^2(\mu, \nu)$.

Error decomposition (I)

$$\mathbf{E}\big[|\hat{T}_{\lambda,n} - W_2^2|\big] \le \underbrace{\mathbf{E}\big[|\hat{T}_{\lambda,n} - \hat{W}^2_{2,n}|\big]}_{\substack{\text{Approximation error} \\ \lesssim \lambda \log(n)}} + \underbrace{\mathbf{E}\big[|\hat{W}^2_{2,n} - W_2^2|\big]}_{\substack{\text{Estimation error} \\ \lesssim n^{-2/d} \text{ (if } d > 4)}}$$

- With $\lambda \lesssim n^{-2/d}$, we get $\tilde{O}(n^{-2/d})$ accuracy (if $d > 4$)
- That's how regularization is analyzed in prior work

Can we use larger values of λ?

[Refs]:
Niles-Weed (2018). *An explicit analysis of the entropic penalty in linear programming.*

Naive unsuccessful attempt

Shortcuts: $\hat{T}_{\lambda,n} = T_\lambda(\hat{\mu}_n, \hat{\nu}_n)$, $T_\lambda = T_\lambda(\mu, \nu)$, $W_2^2 = W_2^2(\mu, \nu)$.

Error decomposition (II)

$$\mathbf{E}\left[|\hat{T}_{\lambda,n} - W_2^2|\right] \le \underbrace{\mathbf{E}\left[|\hat{T}_{\lambda,n} - T_\lambda|\right]}_{\substack{\text{Estimation error} \\ \lesssim (1+\lambda^{-d/2})n^{-1/2}}} + \underbrace{|T_\lambda - W_2^2|}_{\substack{\text{Approximation error} \\ \lesssim \lambda(1+\log(1/\lambda))}}$$

$\rightsquigarrow$ With $\lambda = n^{-1/(d+2)}$, we get $\mathbf{E}\left[|\hat{T}_\lambda - W_2^2|\right] \lesssim n^{-1/(d+2)} \log(n)$

Drawback of T_λ: poor approximation error

NB: estimation error bound potentially not tight

[Refs]:
Genevay et al. (2019). *Sample Complexity of Sinkhorn Divergences.*
Mena, Niles-Weed (2019). *Statistical bounds for entropic optimal transport [...]*

Improving the Approximation Error

Sinkhorn divergence

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

- It is positive definite: $S_\lambda(\mu,\nu) \geq 0$ with equality iff $\mu = \nu$
- Interpolation properties:

$$\begin{cases} \lim\limits_{\lambda\to 0} S_\lambda(\mu,\nu) = W_2^2(\mu,\nu) \\ \lim\limits_{\lambda\to\infty} S_\lambda(\mu,\nu) = \|\mathbf{E}_{X\sim\mu}[X] - \mathbf{E}_{Y\sim\nu}[Y]\|_2^2 \end{cases}$$

- As λ increases:
  - Increasing statistical and computational efficiency
  - Decreasing discriminative power

Can we quantify the trade-offs at play?

[Refs]:
Genevay, Peyré, Cuturi (2019). *Learning generative models with Sinkhorn divergences.*
Feydy, Séjourné, Vialard, Amari, Trouvé, Peyré (2019). *Interpolating between Optimal Transport and MMD.*

Dynamic entropy regularized optimal transport

Let $H(\mu) = \int \log(\mu(x))\mu(x)\,\mathrm{d}x$ and μ, ν with bounded densities.

Theorem (Yasue formulation of the Schrödinger problem)

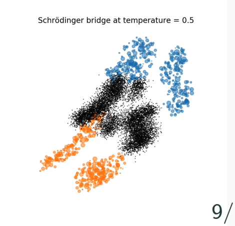
$$T_\lambda(\mu,\nu) + d\lambda\log(2\pi\lambda) + \lambda(H(\mu)+H(\nu)) =$$

$$\min_{\rho,v}\int_0^1\int_{\mathbb{R}^d}\Big(\underbrace{\|v(t,x)\|_2^2}_{\textit{Kinetic energy}} + \frac{\lambda^2}{4}\underbrace{\|\nabla_x \log(\rho(t,x))\|_2^2}_{\textit{Fisher information}}\Big)\rho(t,x)\,\mathrm{d}x\,\mathrm{d}t$$

where (ρ, v) *solves* $\partial_t\rho + \nabla\cdot(\rho v) = 0$, $\rho(0,\cdot) = \mu$ *and* $\rho(1,\cdot) = \nu$.

Definition (Fisher info. of the W_2-geodesic)

$$I(\mu,\nu) := \int_0^1\int_{\mathbb{R}^d}\|\nabla_x \log\rho(t,x)\|_2^2\rho(t,x)\,\mathrm{d}x\,\mathrm{d}t$$



[Refs]:
Chen, Georgiou, Pavon (2019). *On the relation between optimal transport [...].*
Conforti, Tamanini (2019). *A formula for the time derivative of the entropic cost.*

Tight approximation bounds

Recall assumptions: μ, ν have bounded densities and supports.

Theorem (CRLVP'20)

$$|S_\lambda(\mu,\nu) - W_2^2(\mu,\nu)| \leq \frac{\lambda^2}{4} \max\{I(\mu,\nu), (I(\mu)+I(\nu))/2\}.$$

If moreover the right-hand side is finite, it holds

$$S_\lambda(\mu,\nu) - W_2^2(\mu,\nu) = \frac{\lambda^2}{4}\Big(I(\mu,\nu) - (I(\mu)+I(\nu))/2\Big) + o(\lambda^2).$$

Proof idea. (1) Immediate from Yasue formula. (2) Variational analysis arguments to get the right derivative of $\lambda^2 \mapsto S_\lambda$ at 0.

- (in paper) bound $I(\mu,\nu)$ given regularity of Brenier potential
- from $\lambda \log(1/\lambda)$ to λ^2 for (almost) free!

Richardson extrapolation

We can cancel the term in λ^2 for (almost) free. Let

$$R_\lambda(\mu, \nu) := 2S_\lambda(\mu, \nu) - S_{\sqrt{2}\lambda}(\mu, \nu).$$

Proposition

If μ, ν have bounded densities and $I(\mu, \nu), I(\mu), I(\nu) < \infty$ then

$$|R_\lambda(\mu, \nu) - W_2^2(\mu, \nu)| = o(\lambda^2)$$

- Up to constants, T_λ, S_λ and R_λ have the same sample and computational complexities but better approximation errors
- *Open question*: when is the remainder in $O(\lambda^4)$?

[Ref]:
Bach (2020). *On the effectiveness of Richardson extrapolation in machine learning*.

Gaussian case

Let $\mu = \mathcal{N}(a, A)$, $\nu = \mathcal{N}(b, B)$ where $a, b \in \mathbb{R}^d$ and $A, B \in \mathcal{S}^d_{++}$.

If $a = b$, W_2 is the *Bures distance*:

$$W_2^2(\mu,\nu) = \mathrm{d}_B^2(A,B) := \operatorname{tr} A + \operatorname{tr} B - 2\operatorname{tr}(A^{1/2}BA^{1/2})^{1/2}.$$

Exploiting the closed-form expression for $T_\lambda(\mu,\nu)$, we prove:

Expansion Gaussian case

$$S_\lambda(\mu,\nu) - W_2^2(\mu,\nu) = -\frac{\lambda^2}{8}\,\mathrm{d}_B^2(A^{-1},B^{-1}) + \frac{\lambda^4}{384}\,\mathrm{d}_B^2(A^{-3},B^{-3}) + O(\lambda^5)$$

- Richardson extrapolation can boost approximation rates here
- Consistent with expansion in terms of $I(\mu,\nu)$, as it must.

[Refs]:
Chen, Georgiou, Pavon (2015). *Optimal steering of a linear stochastic system to a final probability distribution.*
Janati, Muzellec, Peyré, Cuturi (2020). *Entropic Optimal Transport between Gaussian Measures [...].*

Statistical & Computational Consequences

Sinkhorn Divergence Estimator

Shortcuts: $\hat{S}_{\lambda,n} = S_\lambda(\hat{\mu}_n, \hat{\nu}_n)$, $S_\lambda = S_\lambda(\mu,\nu)$, $W_2^2 = W_2^2(\mu,\nu)$.

Error decomposition (II)

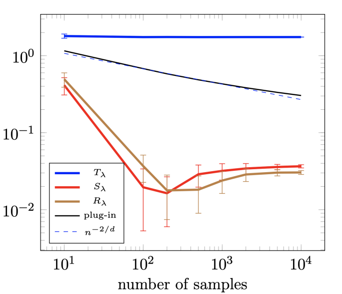
$$\mathbf{E}\left[|\hat{S}_{\lambda,n} - W_2^2|\right] \le \underbrace{\mathbf{E}\left[|\hat{S}_{\lambda,n} - S_\lambda|\right]}_{\substack{\text{Estimation error} \\ \lesssim (1+\lambda^{-d/2})n^{-1/2}}} + \underbrace{|S_\lambda - W_2^2|}_{\substack{\text{Approximation error} \\ \lesssim \lambda^2}}$$

$\rightsquigarrow$ With $\lambda = n^{-1/(d+4)}$, we get $\mathbf{E}\left[|\hat{S}_{\lambda,n} - W_2^2|\right] \lesssim n^{-2/(d+4)}$

- We "almost" recover the rate of the plug-in estimator
- But with a much larger λ ! ($n^{-1/(d+4)}$ instead of $n^{-2/d}$)
- Rate further improved w/ Richardson extrapolation $R_\lambda(\hat{\mu}_n, \hat{\nu}_n)$

Numerical experiments (I): estimate W_2^2

μ, ν elliptically contoured, smooth densities, compact supports.



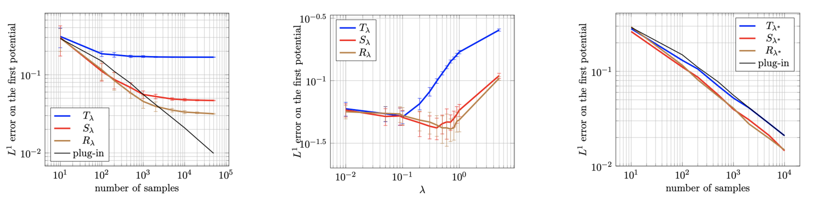
Absolute error on W_2^2 ($d = 10$, $\lambda = 1$).

- $\hat{S}_{\lambda,n}$ and $\hat{R}_{\lambda,n}$ quickly reach a good estimation
- then reach a plateau (the approximation error takes-over)
- difficult to interpret because W_2^2 is a scalar...

Numerical experiments (II): estimate dual potentials

Estimate φ, the Fréchet derivative of $\mu \mapsto W_2^2(\mu,\nu)$ ($d = 5$).

We plot the $L^1(\mu)$ estimation error.



(left) vs. n for $\lambda = 1$ (middle) vs. λ for $n = 10^4$ (right) vs. n for best λ.

Estimator	$\hat{T}_{\lambda,n}$	$\hat{S}_{\lambda,n}$	$\hat{R}_{\lambda,n}$
Time (s)	0.25	0.08	0.12

Table 1: Time to reach 0.03-accuracy via Sinkhorn's algorithm

Conclusion

- **Refined approximation error analysis**
- **Statistical & computational consequences**
- **Theory consistent with practical behavior**

[Paper :]

- Chizat, Roussillon, Léger, Vialard, Peyré (2020). Faster Wasserstein Distance Estimation with the Sinkhorn Divergence.