Entropic Regularization of Optimal Transport as a Statistical Regularization

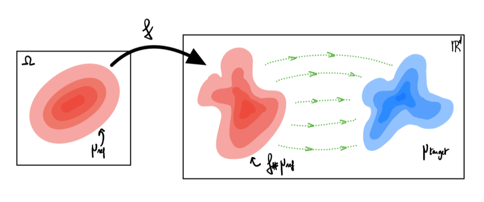
Lénaïc Chizat (EPFL)
based on joint work with Pierre Roussillon, Flavien Léger, François-Xavier Vialard and Gabriel Peyré

Dec. 13th, 2021 - Neurips OT Workshop

A motivating problem: density fitting

Find a map f such that $\mathrm{Loss}(f_{\#}\mu_{\text{ref}}, \mu_{\text{target}})$ is small.

- Choose an objective loss and a parametric family $\{f_\theta \,;\, \theta \in \Theta\}$
- Run a gradient-based algorithm to select θ



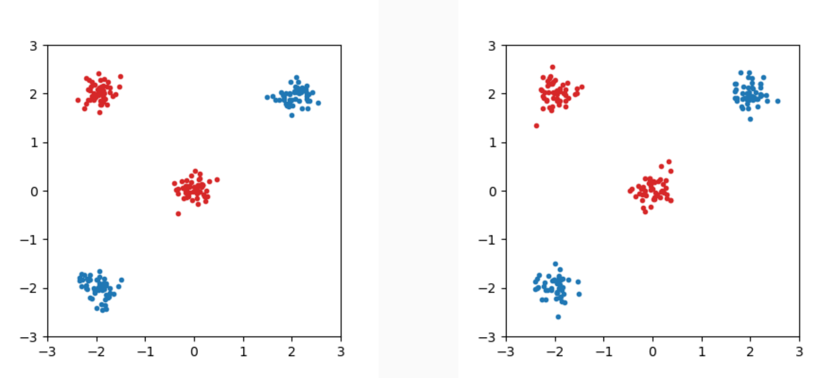
Examples: diffeomorphic matching, generative models

Important properties for the loss (Wasserstein, MMD, etc)

- favorable computational and statistical behavior
- informative gradient (strong when loss is high)

Illustration with the Sinkhorn divergence loss

Simplest case: $f_\theta = \theta$ (L^2-gradient descent); regularization $\lambda \geq 0$



$\lambda \ll 1$ (approx. Wasserstein) $\lambda \gg 1$ (approx. MMD)

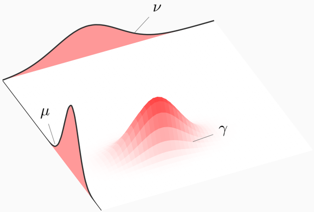
- in the end, we would like to understand the trade-offs at play in the choice of λ in such dynamics...
- ... but for now, we'll focus on the estimating the loss

Wasserstein loss

Definition (Set of transport plans between $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$)

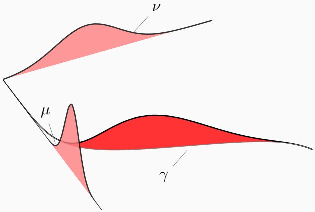
Nonnegative measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν:

$$\Pi(\mu,\nu) := \left\{\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) : \mathrm{proj}^1_{\#}\gamma = \mu,\ \mathrm{proj}^2_{\#}\gamma = \nu\right\}$$



Product coupling

$\gamma = \mu \otimes \nu$



Deterministic coupling

$\gamma = (Id \times T)_{\#}\mu$

Definition (Wasserstein loss)

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

Statistical & Computational Optimal Transport

Goal: estimate W_2^2 efficiently from samples

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.

How difficult is this task?

Can entropic regularization help?

[Related refs]:

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.*

Liang (2019). *On the Minimax Optimality of Estimating the Wasserstein Metric. ...*

Outline

The plug-in estimator

Entropic regularization

Improving the Approximation Error

Statistical & Computational Consequences

The plug-in estimator

Theorem (CRLVP'20)

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

- prev. known bound: $n^{-1/d}$ [e.g. Boissard & LeGouic (2014)]
- concentrates around its expectation (variance $\lesssim n^{-1/2}$)
- extended by [Manole & Niles-Weed, 2021], for $d > 4$,

$$\mathbf{E}\left[|W_p^p(\hat{\mu}_n, \hat{\nu}_n) - W_p^p(\mu, \nu)|\right] \lesssim \begin{cases} n^{-p/d} & \text{if } 1 \leq p \leq 2, \\ n^{-2/d} & \text{if } p \geq 2, \end{cases}$$

Also prove tightness of bounds & cover the non compact case

[Refs:]

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

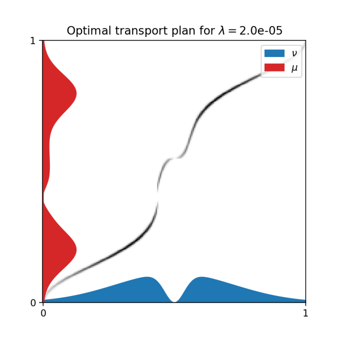
Manole, Niles-Weed (2021). *Sharp Convergence Rates for Empirical Optimal Transport with Smooth Costs.*

Entropic regularization

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
- assume $\lambda \leq 1/2$ in the following
- the higher λ, the easier to solve

[Refs]:

Léonard (2012). *From the Schrödinger problem to the Monge–Kantorovich problem.*

Kosowsky, Yuille (1994). *The invisible hand algorithm*

Cuturi (2013). *Sinkhorn Distances: Lightspeed Computation of Optimal Transport*

Computational bound to compute $T_\lambda(\hat{\mu}_n, \hat{\nu}_n)$ via Sinkhorn

Sinkhorn's iterations

Let $c_{i,j} = \frac{1}{2}\|x_i - y_j\|_2^2$, $v^{(0)} = 0 \in \mathbb{R}^n$ and compute for $k \geq 1$:

$$u_i^{(k)} = -\lambda \log \frac{1}{n} \sum_{j=1}^{n} e^{(v_j^{(k-1)} - c_{i,j})/\lambda} \quad \text{and} \quad v_j^{(k)} = -\lambda \log \frac{1}{n} \sum_{i=1}^{n} e^{(u_i^{(k)} - c_{i,j})/\lambda}.$$

The current estimate is $\hat{T}_{\lambda,n}^{(k)} = \frac{2}{n} \sum_i (u_i^{(k)} + v_i^{(k)})$.

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

After k iterations, it holds

$$|\hat{T}_{\lambda,n}^{(k)} - \hat{T}_{\lambda,n}| \lesssim \lambda^{-1} k^{-1}.$$

$\rightsquigarrow$ Reaches 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 [...]*

Approximation error

Proposition (Approximation error)

$$\mathbb{E}[|W_2^2(\mu,\nu) - T_\lambda(\mu,\nu)|] \lesssim \lambda \log(1/\lambda)$$

- Remember that $T_0 = W_2^2$ by definition.
- Simple proof: approximate the optimal transport plan with a plan of finite entropy that is piecewise proportional to $\mu \otimes \nu$
- bound tight for densities (see asymptotic expansion later)
- finer results for the discrete case [Niles-Weed (2018)]

[Refs]:
Pal (2019). *On the difference between entropic cost and the optimal transport cost.*
Genevay, Chizat, Bach, Cuturi, Peyré (2018). *Sample Complexity of Sinkhorn divergences.*
Niles-Weed (2018). *An explicit analysis of the entropic penalty in linear programming.*

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)

$$\mathbf{E}\big[|\hat{T}_{\lambda,n} - W_2^2|\big] \leq \underbrace{\mathbf{E}\big[|\hat{T}_{\lambda,n} - \hat{W}^2_{2,n}|\big]}_{\substack{\text{Approximation error} \\ \lesssim \lambda \log(1/\lambda)}} + \underbrace{\mathbf{E}\big[|\hat{W}^2_{2,n} - W_2^2|\big]}_{\substack{\text{Estimation error} \\ \lesssim n^{-2/d} \text{ (if } d > 4)}}$$

- With $\lambda \asymp n^{-2/d}$, we get $\tilde{O}(n^{-2/d})$ accuracy (if $d > 4$)

Can we see entropy as a statistical regularization instead ?
Can we use larger values of λ?

Standard error decomposition of a regularized estimator

k: nb of Sinkhorn iterations

n: nb of samples

λ: regularization strength

$$\underbrace{|\hat{T}^{(k)}_{\lambda,n} - T_0|}_{\text{Total error}} \leq \underbrace{|\hat{T}^{(k)}_{\lambda,n} - \hat{T}_{\lambda,n}|}_{\text{Optimization error}} + \underbrace{|\hat{T}_{\lambda,n} - T_\lambda|}_{\text{Estimation error}} + \underbrace{|T_\lambda - T_0|}_{\text{Approximation error}}$$

- in the following we ignore the optimization error
- we focus on expectation bounds as all quantities concentrate rapidly
- online algorithms would require a different error decomposition

[Refs]:
Bottout, Bousquet (2007). *The Tradeoffs of Large Scale Learning*.

Estimation error of T_λ

Theorem (Estimation)

$$\mathbb{E}[|T_\lambda(\hat{\mu}_n, \hat{\nu}_n) - T_\lambda(\mu, \nu)|] \lesssim \lambda^{-d/2} n^{-1/2}$$

T_λ is also stable under deterministic sampling, see [CRLVP,20]. For smooth densities and a regular grid:

$$|T_\lambda(\mu_n, \nu_n) - T_\lambda(\mu, \nu)| \lesssim \min\{\lambda^{-1} n^{-2/d}, n^{-1/d}\}$$

[Refs]:
Genevay, Chizat, Bach, Cuturi, Peyré (2018). *Sample Complexity of Sinkhorn divergences.*
Mena, Niles-Weed (2018). *Statistical bounds for entropic optimal transport.*

Naive unsuccessful attempt

Error decomposition (II)

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

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

Drawback of T_λ: poor approximation error

NB: estimation error bound potentially not tight

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:

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

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

How to quantify the trade-offs at play?

$\rightsquigarrow$ interpret it as an estimator for W_2^2

[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)

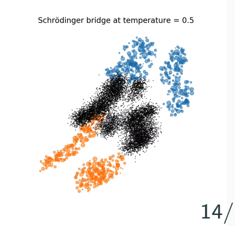
$$T_\lambda(\mu,\nu) + d\lambda\log(2\pi\lambda) + \lambda(H(\mu)+H(\nu)) =$$

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

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 (2020). *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, Conforti & Tamanini, 2020)

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

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!
- extended to a general setting in [Conforti & Tamanini, 2020]

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*:

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

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)

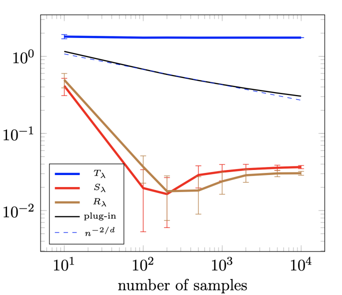
$$\mathbf{E}\left[|\hat{S}_{\lambda,n} - W_2^2|\right] \leq \underbrace{\mathbf{E}\left[|\hat{S}_{\lambda,n} - S_\lambda|\right]}_{\substack{\text{Estimation error} \\ \lesssim \lambda^{-d/2} n^{-1/2}}} + \underbrace{|S_\lambda - W_2^2|}_{\substack{\text{Approximation error} \\ \lesssim \lambda^2}}$$

$\rightsquigarrow$ 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.



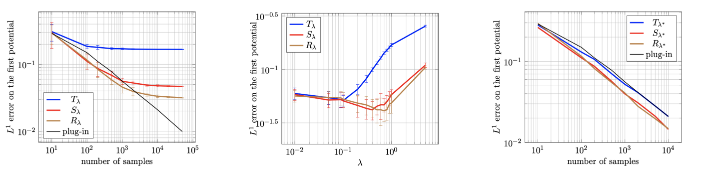
Absolute error on W_2^2 ($d = 10$, $\lambda = 1$).

- $\hat{S}_{\lambda,n}$ 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...

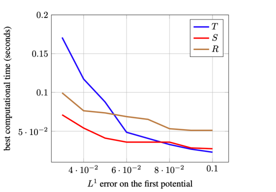
Numerical experiments (II): estimate dual potentials

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

We plot the $L^1(\mu)$ estimation error ($d = 5$).



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



Computational time to reach a target accuracy (optimizing over n and λ)

[Refs]:
Pooladian, Niles-Weed (2021) *Entropic estimation of optimal transport maps*

In a nutshell

To estimate $W_2^2(\mu,\nu)$: $S_\lambda(\hat{\mu}_n, \hat{\mu}_n)$ is "better" than $W_2^2(\hat{\mu}_n, \hat{\nu}_n)$!

Future directions

- **Which statistical bounds can be improved?**
- **Consequences for density fitting algorithms**

[Paper :]

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