Part 2: Computation and applications

Outline

Part 2: Computation and applications

- Exact and approximate computation
- Some statistical properties
- OT as a loss function
- Application: Generative models

Quick recap

Recall that we defined the primal problem (KP):

$$\min_{\boldsymbol{\pi} \in \Pi(\mu,\nu)} \int_{\mathcal{X}^2} c(x,y) \mathrm{d}\boldsymbol{\pi}(x,y),$$

and the corresponding dual problem (DP):

$$\max \left\{ \int_{\mathcal{X}} \varphi \, \mathrm{d}\mu + \int_{\mathcal{X}} \psi \, \mathrm{d}\nu : \varphi \oplus \psi \leq c, \text{ and } \varphi \in L_1(\mu), \psi \in L_1(\nu) \right\},$$

and that under general conditions, $\min\,(\mathrm{KP}) = \max\,(\mathrm{DP})$.

Discrete optimal transport

Important in applications, amenable to computation.

Suppose $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$, so that

$$\Pi(\mu, \nu) = \left\{ \boldsymbol{\pi} \in \mathbb{R}_+^{n \times m} : \boldsymbol{\pi} \mathbb{1}_m = a \text{ and } \boldsymbol{\pi}^\top \mathbb{1}_n = b \right\}.$$

Denoting $c_{i,j} = c(x_i, y_j)$, (KP) can be expressed

$$\min_{\boldsymbol{\pi} \in \Pi(\mu,\nu)} \sum_{i,j} c_{i,j} \boldsymbol{\pi}_{i,j} = \min_{\boldsymbol{\pi} \in \Pi(\mu,\nu)} \langle c, \boldsymbol{\pi} \rangle.$$

Similarly, (DP) can be expressed

$$\max \left\{ \langle \varphi, a \rangle + \langle \psi, b \rangle : (\varphi, \psi) \in \mathbb{R}^n \times \mathbb{R}^m \text{ and } \varphi_i + \psi_j \leq c_{i,j} \right\}.$$

Discrete optimal transport

Since (KP) is a linear program and $\Pi(\mu, \nu)$ is non-empty and bounded, min (KP) is attained at an *extremal point* of $\Pi(\mu, \nu)$.

- See Bertsimas and Tsitsiklis (1997, Theorem 2.7).

Extremal points of $\Pi(\mu, \nu)$ have a particular structure:

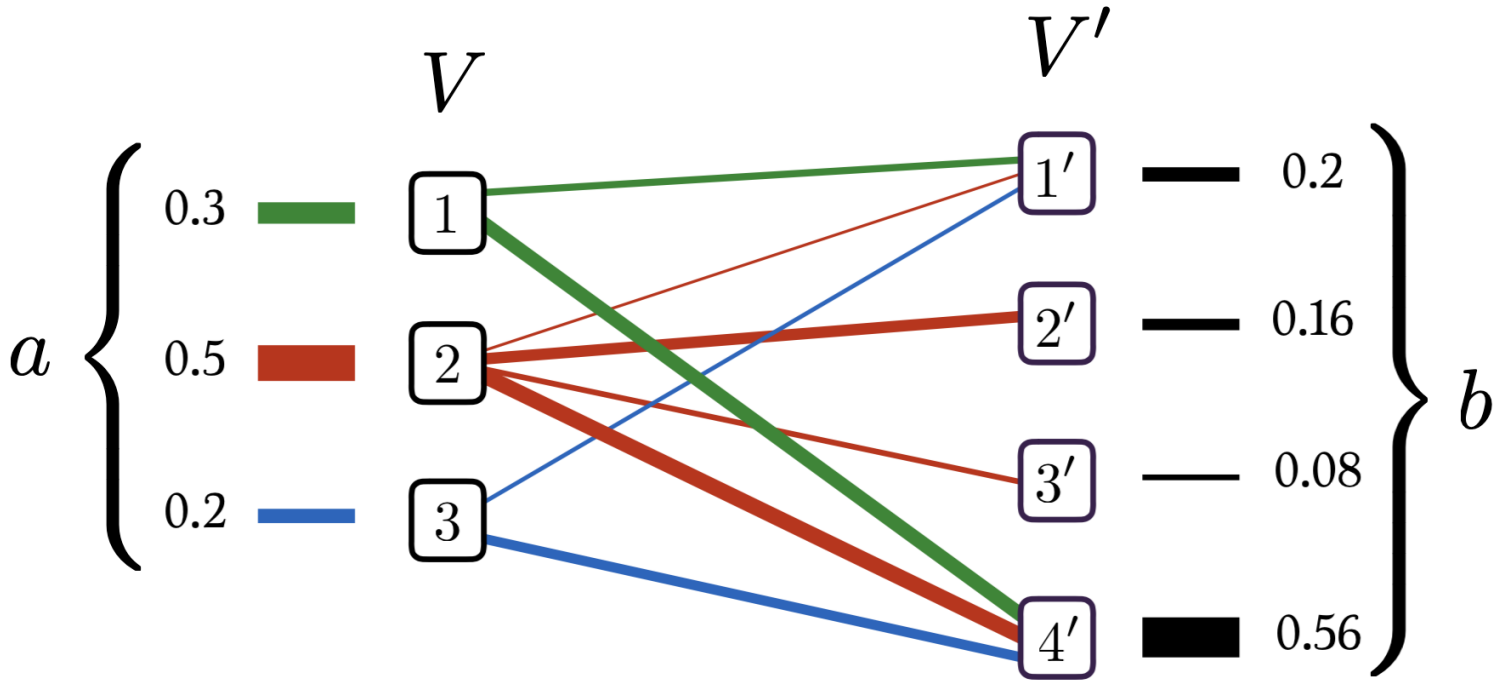
- Let $V = (1, \ldots, n), V' = (1', \ldots, m')$ be two sets of nodes.
- Let $E = \{(i, j') : 1 \leq i \leq n, 1 \leq j' \leq m\}$ be the set of edges.
- Let $\mathcal{G} = (V \cup V', E)$ be the bipartite graph between V and V'.

Any $\boldsymbol{\pi} \in \Pi(\mu, \nu)$ corresponds to a flow over $\mathcal{G}$ with

- a_i flowing out of node i,
- b_j flowing into node j'.

Fact: If $\boldsymbol{\pi}$ is an extremal point of $\Pi(\mu, \nu)$, and $F(\boldsymbol{\pi}) \subset E$ are the edges (i, j') such that $\boldsymbol{\pi}_{i,j} > 0$, then $\mathcal{G}(\boldsymbol{\pi}) = (V \cup V', F(\boldsymbol{\pi}))$ has no cycles.

Discrete optimal transport



- Feasible, but not optimal: Cycle $((1, 1'), (2, 1'), (2, 4'), (1, 4'))$.

Figure from Peyré and Cuturi (2019).

Exact computation

Can compute optimal $\boldsymbol{\pi}^\star$ with the network simplex algorithm.

- Acts in the primal, performs updates on feasible $\boldsymbol{\pi}$.

Can compute optimal $(\varphi^\star, \psi^\star)$ with dual ascent methods.

- Acts in the dual, performs updates on feasible (φ, ψ).

Special case: $n = m$ and $a = b = \frac{1}{n}\mathbb{1}_n$.

- Birkhoff's theorem: extremal points of $\Pi(\mu, \nu)$ are permutation matrices, so (KP) reduces to

$$\min_{\sigma \in \mathrm{Perm}(n)} \frac{1}{n} \sum_{i=1}^{n} c_{i,\sigma(i)}.$$

- Can be solved with the Hungarian algorithm, costs $\mathcal{O}(n^3)$.

Entropic regularization (Cuturi, 2013)

For $\varepsilon > 0$, the entropically regularized OT problem (ε-EOT) is defined

$$\min_{\boldsymbol{\pi} \in \Pi(\mu,\nu)} \langle c, \boldsymbol{\pi} \rangle + \varepsilon \mathrm{KL}(\boldsymbol{\pi} | \mu \otimes \nu).$$

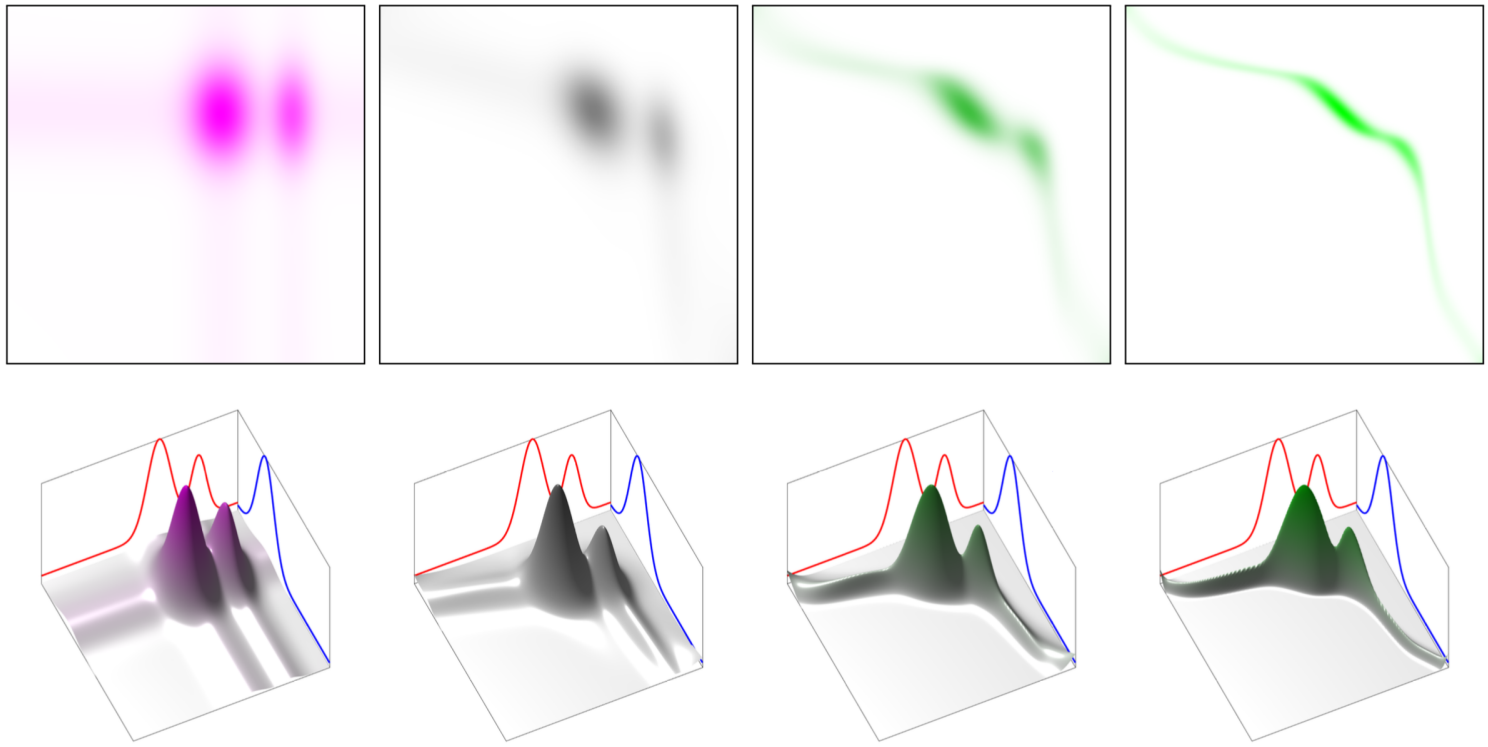
Due to 1-strong convexity of $\mathrm{KL}(\cdot|\mu \otimes \nu)$, (EOT) is ε-strongly convex.

- Unique minimizer $\boldsymbol{\pi}^\varepsilon$.
- $\min$ (ε-EOT) $\to \min$ (KP) as $\varepsilon \to 0$, and

  $$\boldsymbol{\pi}^\varepsilon \to \operatorname{argmin} \{\mathrm{KL}(\boldsymbol{\pi}|\mu \otimes \nu) : \boldsymbol{\pi} \in \Pi(\mu,\nu), \langle c, \boldsymbol{\pi} \rangle = \min \text{ (KP)}\}.$$

- $\boldsymbol{\pi}^\varepsilon \to \mu \otimes \nu$ as $\varepsilon \to +\infty$.

Entropic regularization



$\varepsilon = 10$ $\varepsilon = 1$ $\varepsilon = 10^{-1}$ $\varepsilon = 10^{-2}$

Figure from Peyré and Cuturi (2019).

Entropic regularization

Define the matrix $K^\varepsilon \in \mathbb{R}_+^{n \times m}$ by $K^\varepsilon_{i,j} = e^{-\frac{c_{i,j}}{\varepsilon}}$.

- Gibbs kernel associated with the cost matrix c.

Proposition

The minimizer $\boldsymbol{\pi}^\varepsilon$ *is given by*

$$\boldsymbol{\pi}^\varepsilon_{i,j} = u^\varepsilon_i K^\varepsilon_{i,j} v^\varepsilon_j,$$

for two vectors $(u^\varepsilon, v^\varepsilon) \in \mathbb{R}_+^n \times \mathbb{R}_+^m$.

Proof sketch: For dual variables $(\varphi, \psi) \in \mathbb{R}^n \times \mathbb{R}^m$, the Lagrangian of ($\varepsilon$-EOT) is given by

$$\mathcal{L}(\boldsymbol{\pi}, \varphi, \psi) = \langle c, \boldsymbol{\pi} \rangle + \varepsilon \mathrm{KL}(\boldsymbol{\pi} | \mu \otimes \nu) - \langle \varphi, \boldsymbol{\pi} \mathbb{1}_m - a \rangle - \langle \psi, \boldsymbol{\pi}^\top \mathbb{1}_n - b \rangle.$$

Differentiate with respect to $\boldsymbol{\pi}_{i,j}$ and set equal to zero.

Sinkhorn's algorithm

Set $v^{(0)} = \mathbb{1}_m$ and iterate until convergence:

$$u^{(\ell+1)} = \frac{a}{K v^{(\ell)}}, \qquad v^{(\ell+1)} = \frac{b}{K^\top u^{(\ell+1)}}.$$

- One iteration costs $\mathcal{O}(n^2)$, parallelizable on GPUs.
- Linear convergence of $u^{(\ell)}$ and $v^{(\ell)}$ in the Hilbert metric.
- δ-accurate solution in $\mathcal{O}(n^2 \log(1/\delta))$.
- Can be used to build δ-accurate solution of (KP) in $\tilde{\mathcal{O}}(n^2/\delta^2)$.

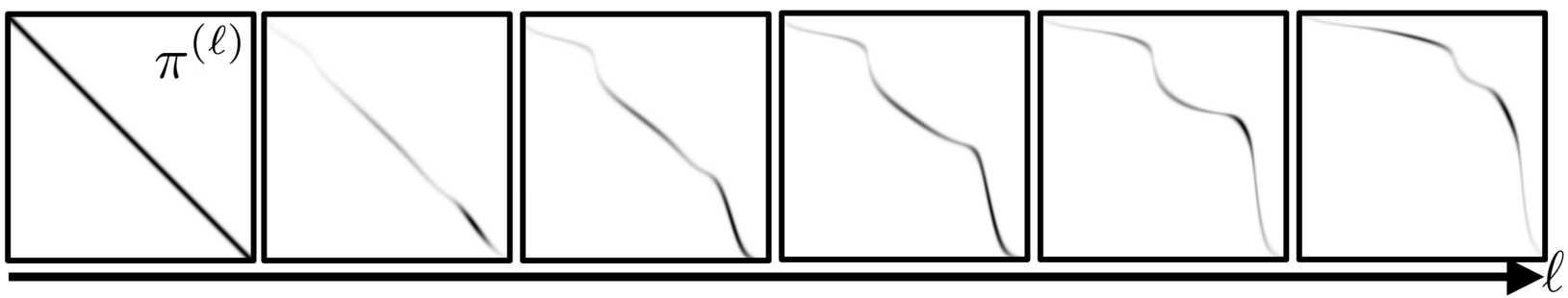


Figure from Peyré and Cuturi (2019).

Statistical properties

Let $\mathcal{X} = \mathbb{R}^d$, $\mu, \nu \ll \text{Leb}$, and $\{x_i\}_{i=1}^n \sim \mu$, $\{y_j\}_{j=1}^n \sim \nu$ i.i.d. Then,

$$\mathbb{E}|\mathcal{W}_p^p(\hat{\mu}_n, \hat{\nu}_n) - \mathcal{W}_p^p(\mu, \nu)| = \mathcal{O}(n^{-p/d}), \qquad p \in [1, d/2].$$

Rates can adapt to

- lower dimensional structures (Weed and Bach, 2019),
- smoothness in densities (Weed and Berthet, 2019).

However, $\mathcal{W}_p^p(\hat{\mu}_n, \hat{\nu}_n)$ concentrates rapidly around $\mathbb{E}\mathcal{W}_p^p(\hat{\mu}_n, \hat{\mu}_n)$:

$$\sqrt{n}\left[\mathcal{W}_p^p(\hat{\mu}_n, \hat{\mu}_n) - \mathbb{E}\mathcal{W}_p^p(\hat{\mu}_n, \hat{\mu}_n)\right] \to \mathcal{N}(0, \sigma^2(\mu, \nu)), \quad \text{in distribution.}$$

- See del Barrio and Loubes (2017).

On the other hand, if $\mathcal{S}_\varepsilon(\mu, \nu) = \min$ (ε-EOT), then

$$\mathbb{E}|\mathcal{S}_\varepsilon(\hat{\mu}_n, \hat{\mu}_n) - \mathcal{S}_\varepsilon(\mu, \nu)| = \mathcal{O}(n^{-1/2}).$$

- See Genevay et al. (2019); Mena and Weed (2019).

OT as a loss function

Suppose we have observed data $\{x_i\}_{i=1}^n \sim \mu$, where μ is unknown.

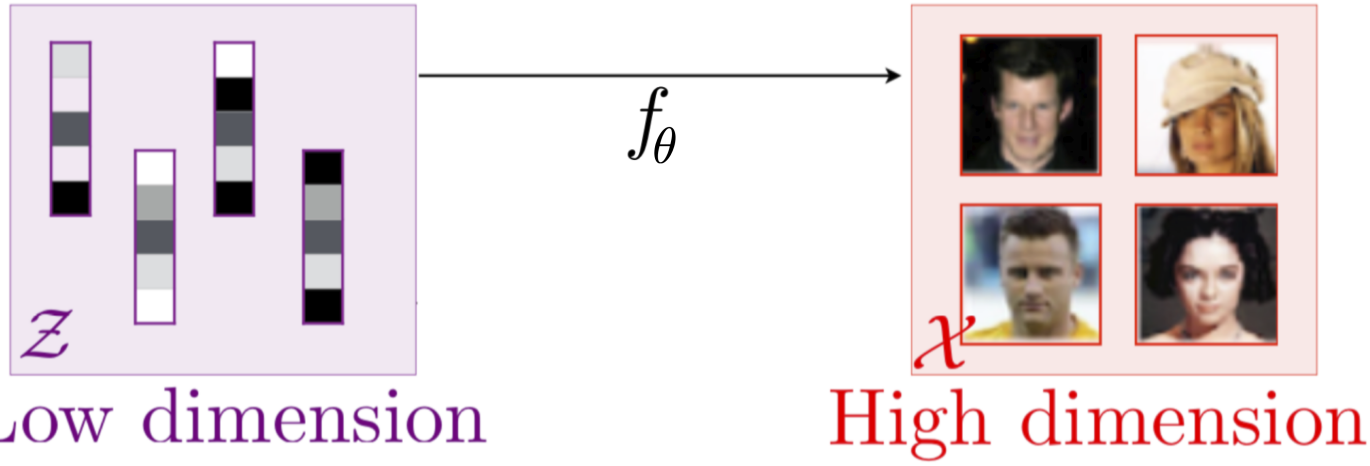
To learn about μ, we specify a model:

- $\mathcal{M} = \{\mu_\theta : \theta \in \mathcal{H}\}$, $\mathcal{H}$ parameter space.
- In practice, μ does typically not belong to $\mathcal{M}$.

If applicable, could e.g. maximize the likelihood to estimate θ.

- What if the likelihood function is not tractable or does not exist?
- What if μ and $\mathcal{M}$ are supported on different lower-dimensional structures?

Example: Image generation

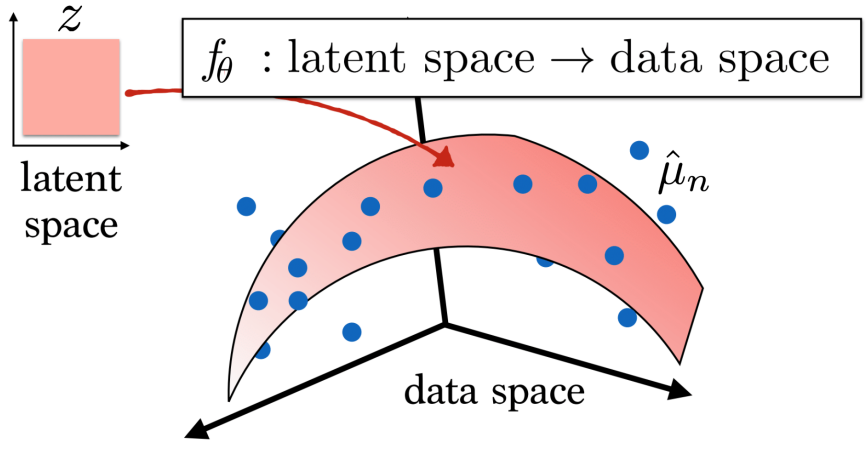


Misspecified: $\text{supp}(\mu) \neq \text{supp}(\mathcal{M})$.

However, the functions

- $\theta \mapsto \mathcal{W}_p(\mu, \mu_\theta)$,
- $\theta \mapsto \mathcal{W}_p(\hat{\mu}_n, \mu_\theta)$,
- $\theta \mapsto \mathbb{E}_\theta \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n})^*$,

still make sense, unlike information-based losses.



*$\mathbb{E}_\theta$ w.r.t. $\{y_i\}_{i=1}^n \sim \mu_\theta$, $\hat{\mu}_{\theta,n} = n^{-1} \sum \delta_{y_i}$. Figures from Peyré (2018); Cuturi (2018).

Generative models

Suppose that we can generate "synthetic" data from the model:

- For any $\theta \in \mathcal{H}$, can simulate $\{y_i\}_{i=1}^n \sim \mu_\theta$.
- For instance, $y_i = f_\theta(z_i)$, where $z_i \sim \mathcal{N}(0, \mathcal{I})$.

Simple example: g-and-k distribution on $\mathbb{R}$, defined by

$$F_\theta^{-1}(r) = a + b\left(1 + 0.8\frac{1-\exp(-gz(r)}{1+\exp(-gz(r)}\right)\left(1+z(r)^2\right)^k z(r),$$

where $z(r)$ is the r-th quantile of $\mathcal{N}(0,1)$.

- Parameter $\theta = (a, b, g, k) \in [0, 10]^4$.
- Intractable likelihood, but can simulate: $\left(F_\theta^{-1}\right)_{\#} \text{Unif}[0,1] = \mu_\theta$.

Minimum Wasserstein estimation

Minimum Wasserstein estimator (MWE):

$$\hat{\theta}_n = \operatorname*{argmin}_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n, \mu_\theta),$$

- First studied by Bassetti et al. (2006) in well-specified models.

Minimum expected Wasserstein estimator (MEWE):

$$\hat{\theta}_n = \operatorname*{argmin}_{\theta \in \mathcal{H}} \mathbb{E}_\theta \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}),$$

- Can be approximated in generative models, with for instance Monte Carlo EM.

MWE convergence

The MWE and MEWE exist, are measurable, and converge almost surely (Bernton et al., 2019):

$$\limsup_{n\to\infty} \operatorname*{argmin}_{\theta\in\mathcal{H}} \mathcal{W}_p(\hat{\mu}_n, \mu_\theta) \subset \operatorname*{argmin}_{\theta\in\mathcal{H}} \mathcal{W}_p(\mu, \mu_\theta).$$

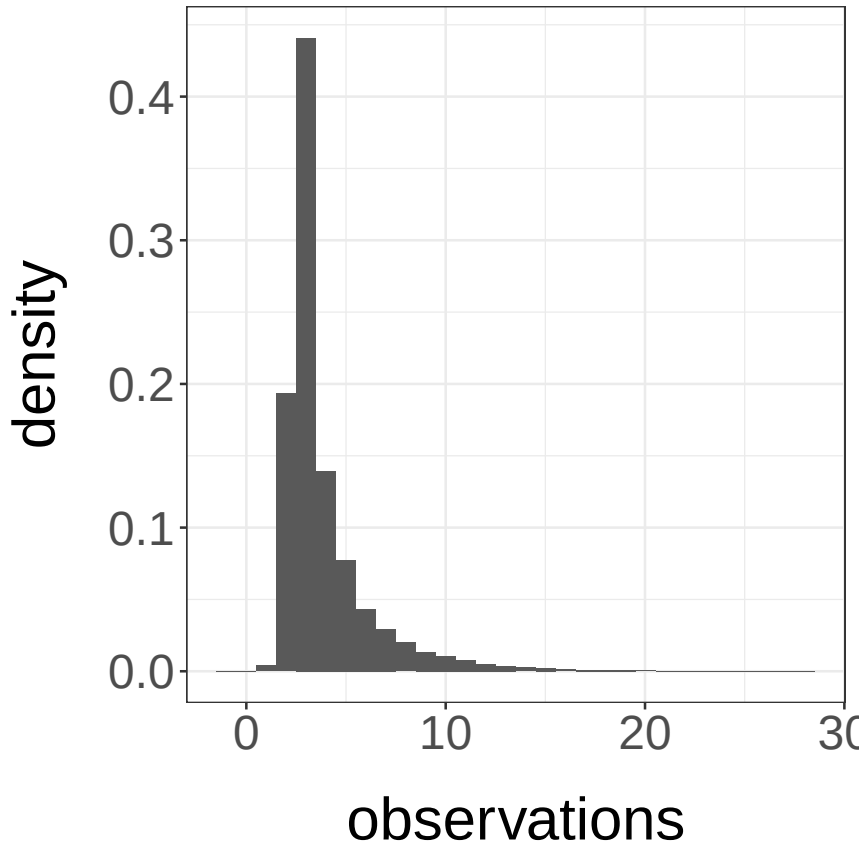
- If $\hat{\theta}_n$ and $\theta_\star = \operatorname{argmin}_{\theta\in\mathcal{H}} \mathcal{W}_p(\mu, \mu_\theta)$ are unique, $\hat{\theta}_n \to \theta_\star$.

For data in $\mathbb{R}$ and $p = 1$, can also derive the asymptotic distribution:

$$\sqrt{n}(\hat{\theta}_n - \theta_\star) \xrightarrow{w} \operatorname*{argmin}_{u\in\mathcal{H}} \int_{\mathbb{R}} |G(t) - \langle u, D_{\theta_\star}(t)\rangle| \mathrm{d}t,$$

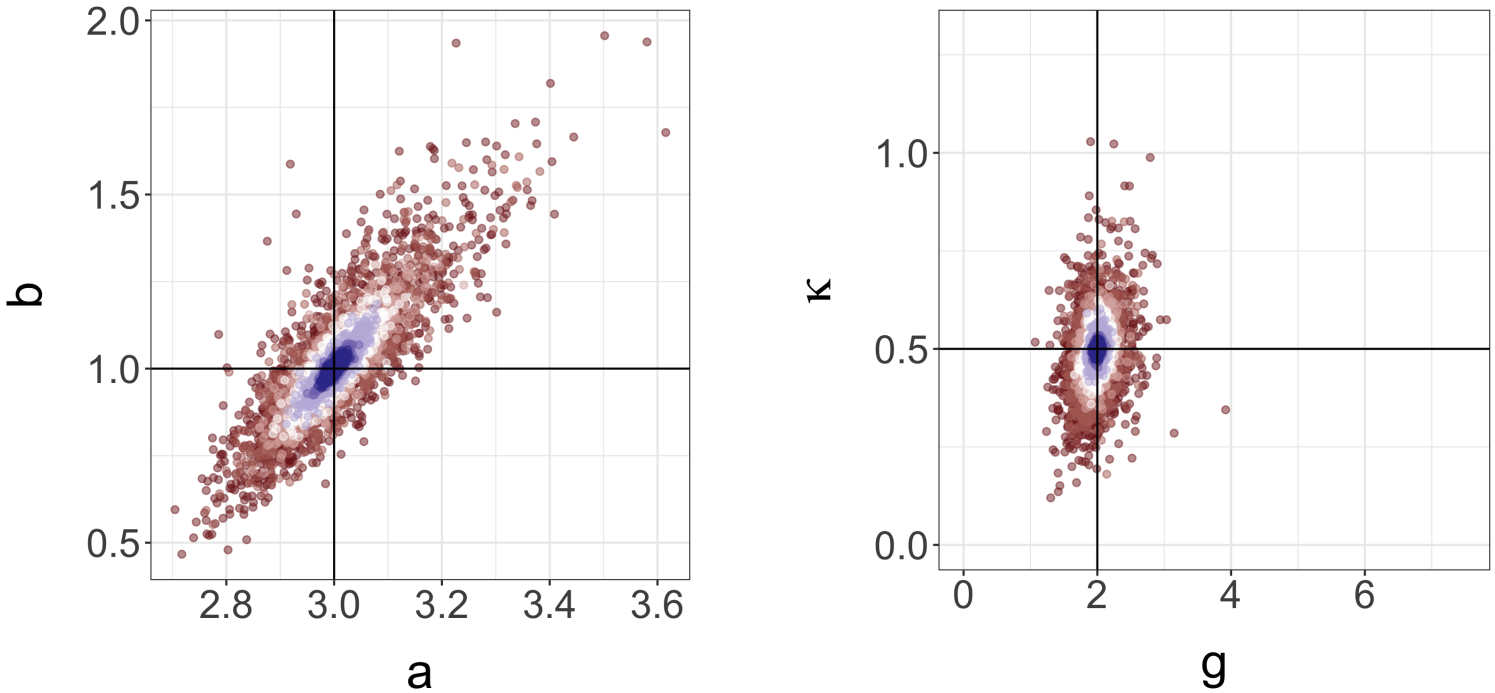
where G is a μ-Brownian bridge, $D_{\theta_\star}$ denotes a "derivative".

Example: g-and-k distribution



Histogram: $n = 1,000$ draws from g-and-k, $\theta_\star = (3, 1, 2, 0.5)$.

Example: g-and-k distribution



MEWE for different values of n, $\theta_\star$ in solid lines.

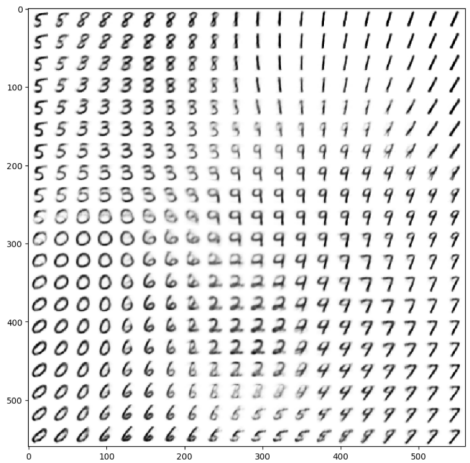
EOT as a loss function

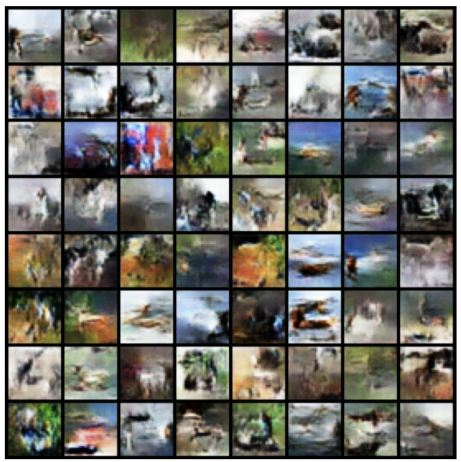
The Wasserstein loss is *not* differentiable in general.

The entropically regularized Wasserstein loss *is* differentiable.

- Approximate MWE with $\hat{\theta}_n = \operatorname{argmin}_{\theta \in \mathcal{H}} \mathbb{E}_\theta \mathcal{S}_\varepsilon(\hat{\mu}_n, \hat{\mu}_{\theta,n})$.
- Compute with SGD, auto-differentiation.

Applied by Genevay et al. (2018) in image generation:





Generators trained on MNIST and CIFAR 10.

OT as a loss function: Open questions

Many interesting questions to tackle:

- Is the slow rate of convergence of $\mathcal{W}_p(\hat{\mu}_n, \hat{\nu}_n)$ problematic for inference when $d > 1$?
- What are the statistical properties of the estimators based on (EOT)?
- What are the statistical properties of the approximations arising from computation?
- .....

Thanks!

References

Bassetti, F., Bodini, A., and Regazzini, E. (2006). On minimum Kantorovich distance estimators. *Statistics & probability letters*, 76(12):1298–1302.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019). On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676.

Bertsimas, D. and Tsitsiklis, J. N. (1997). *Introduction to linear optimization*. Athena Scientific, Belmont, MA.

Cuturi, M. (2013). Sinkhorn distances: lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2292–2300.

del Barrio, E. and Loubes, J.-M. (2017). Central limit theorems for empirical transportation cost in general dimension. *arXiv preprint arXiv:1705.01299*.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. (2019). Sample complexity of Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583.

Genevay, A., Peyre, G., and Cuturi, M. (2018). Learning generative models with Sinkhorn divergences. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1608–1617.

Gramfort, A., Peyré, G., and Cuturi, M. (2015). Fast optimal transport averaging of neuroimaging data. In *Proc. IPMI'15*, pages 261–272.

Kantorovich, L. V. (1942). On the translocation of masses. *Dokl. Akad. Nauk SSSR*, 37(7-8):227–229.

McCann, R. J. (1995). Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80:309–323.

Mena, G. and Weed, J. (2019). Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *arXiv preprint arXiv:1905.11882*.

Monge, G. (1781). *Mémoire sur la théorie des d'eblais et des remblais*. De l'Imprimerie Royale.

Papadakis, N., Peyré, G., and Oudet, E. (2014). Optimal transport with proximal splitting. *SIAM Journal on Imaging Sciences*, 7(1):212–238.

Peyré, G. and Cuturi, M. (2019). Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607.

Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Progress in Nonlinear Differential Equations and their applications*, 87.

Solomon, J., de Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. (2015). Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (Proc. SIGGRAPH 2015)*, 34(4):66:1–66:11.

Villani, C. (2008). *Optimal transport, old and new*. Springer-Verlag New York.

Weed, J. and Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648.

Weed, J. and Berthet, Q. (2019). Estimation of smooth densities in Wasserstein distance. *arXiv preprint arXiv:1902.01778*.