Quadratically Regularized Optimal Transport

Marcel Nutz
Columbia University

March 2026

marcelnutz.com

Joint Work with

Alberto González-Sanz

Eustasio del Barrio

Stephan Eckstein

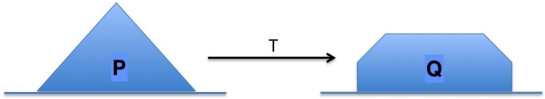
Andrés Riveros Valdevenito

Outline

Optimal Transport in Monge's Formulation

Given:

- Probability measures P, Q on compact subsets $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$
- Cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, continuous



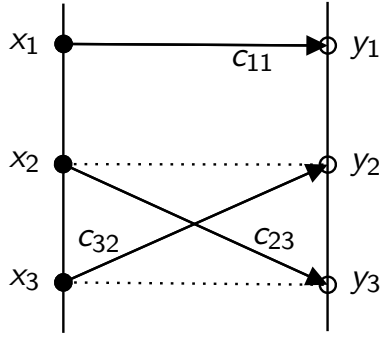
Objective:

- Find a map $T : \mathcal{X} \to \mathcal{Y}$ satisfying $Q = T_{\#}P := P \circ T^{-1}$ such as to minimize the total cost,

$$\min_T \int c(x, T(x))\, P(dx)$$

Empirical Measures: $n \times n$ Assignment Problem

- Given: points $(x_i)_{1 \le i \le n}$ and $(y_i)_{1 \le i \le n}$
- $P = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$ and $Q = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$
- A transport map T = a permutation $\sigma \in \Sigma(n)$
- $c(x_i, y_j) = c_{ij}$ cost matrix
- Cost of a permutation σ is $n^{-1} \sum_i c_{i,\sigma(i)}$



Monge–Kantorovich Optimal Transport

Relaxation:

- Allow to split mass
- Find a probability π on $\mathcal{X} \times \mathcal{Y}$ with marginals P, Q such as to minimize the cost:

$$\mathrm{OT}(P, Q) = \inf_{\pi \in \Pi(P,Q)} \int c(x, y)\, \pi(dx, dy)$$

with $\Pi(P, Q) = \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \pi(\cdot \times \mathcal{Y}) = P(\cdot),\ \pi(\mathcal{X} \times \cdot) = Q(\cdot)\}$

- $\pi \in \Pi(P, Q)$ is a Monge transport if $\pi(dx, dy) = P(dx) \otimes \delta_{T(x)}(dy)$
- In many practical examples, solution is Monge, hence has sparse support

Optimal Transport Distances

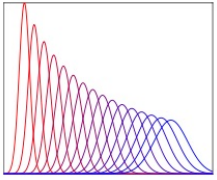
- For suitable cost c, the value $\mathrm{OT}(P, Q)$ quantifies similarity of P, Q
- For $\mathcal{X} = \mathcal{Y}$, it can be used to define a proper metric on $\mathcal{P}(\mathcal{X})$: Wasserstein or Earth's Movers distance

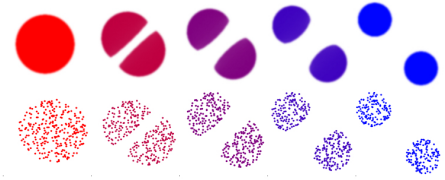
Examples:

- $\mathcal{W}_1(P, Q) = \mathrm{OT}(P, Q)$ for $c(x, y) = \|x - y\|$ "distance cost"
- $\mathcal{W}_2(P, Q) = \mathrm{OT}(P, Q)^{1/2}$ for $c(x, y) = \|x - y\|^2$ "quadratic cost"

- A natural way to lift ground cost/distance to space of distributions
- Can compare discrete and continuous distributions, etc.
- Defines geodesics, averages (barycenters), ... between distributions

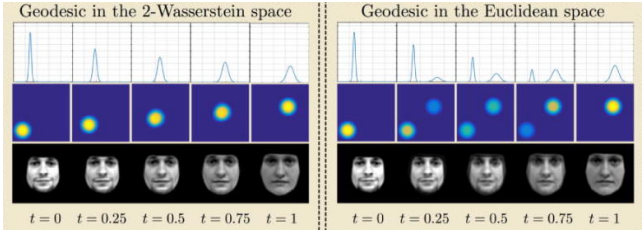
Optimal Transport Distances





$\mathcal{W}_2$ interpolation

(Figures from Peyré&Cuturi 2019)



(Figure from Kolouri et al. 2017)

Sample Complexity: Curse of Dimensionality

- Let $X_1, \dots, X_n$ be i.i.d. samples from P and $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ the associated empirical measure
- Similarly $Y_1, \dots, Y_n$ and Q_n (independent)
- Empirical transport (assignment) problem $\mathrm{OT}(P_n, Q_n)$, "plug-in estimator"
- Let $P \neq Q$. For, e.g., $c(x, y) = \|x - y\|^2$ defining the 2-Wasserstein distance,

$$\mathbb{E}[|\mathrm{OT}(P_n, Q_n) - \mathrm{OT}(P, Q)|] \lesssim n^{-2/d}$$

  sharp rate for dimension $d \geq 5$ (Manole, Niles-Weed '24)

- Exact computation of assignment problem has cost $O(n^3)$

Recent survey: Chewi, Niles-Weed, Rigollet '25, *Statistical Optimal Transport*

Outline

Entropic Optimal Transport

- Regularization parameter $\varepsilon > 0$
- Entropic optimal transport (EOT) problem:

$$\boxed{\mathrm{EOT}_\varepsilon := \inf_{\pi \in \Pi(P,Q)} \int_{\mathcal{X}\times\mathcal{Y}} c\, d\pi + \varepsilon D_{KL}(\pi | P \otimes Q)}$$

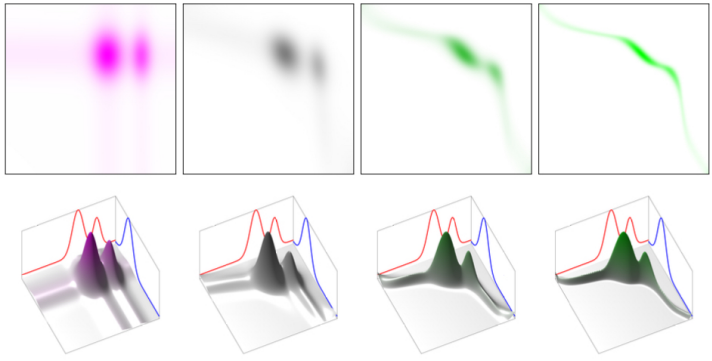
- $D_{KL}(\cdot | P \otimes Q)$ is Kullback–Leibler divergence (relative entropy) wrt. $P \otimes Q$,

$$D_{KL}(\pi | P \otimes Q) := \begin{cases} \int \log\left(\frac{d\pi}{d(P\otimes Q)}\right) d\pi, & \pi \ll P \otimes Q, \\ \infty, & \pi \not\ll P \otimes Q. \end{cases}$$

- Unique minimizer π_ε
- $\mathrm{EOT}_0 = \mathrm{OT}$

- EOT is tradeoff between transport cost and entropy
- "Interpolates" between $P \otimes Q$ and optimal transport

Entropic Regularization



(Figure from Peyré–Cuturi 2019)

Dual Problem and Potentials

- Subsequently, $\varepsilon = 1$ for notational simplicity

The dual EOT problem is

$$\sup_{f \in \mathcal{C}(\mathcal{X}), g \in \mathcal{C}(\mathcal{Y})} \int \left\{ f(x) + g(y) - e^{f(x)+g(y)-c(x,y)} \right\} P(dx) Q(dy) + 1.$$

- Unique (up to constant) solution $(f, g) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y})$
- Called (Schrödinger) potentials
- Potentials (f, g) give the density of the optimal coupling π:

$$\frac{d\pi}{d(P \otimes Q)}(x, y) = e^{f(x)+g(y)-c(x,y)}$$

First-Order Condition of Optimality (Schrödinger System)

- Potentials (f, g) satisfy a first-order condition:

$$1 = \int_{\mathcal{Y}} e^{f(x)+g(y)-c(x,y)}\, Q(dy),$$
$$1 = \int_{\mathcal{X}} e^{f(x)+g(y)-c(x,y)}\, P(dx)$$

- This can be written as

$$f(x) = -\log \int_{\mathcal{Y}} e^{g(y)-c(x,y)}\, Q(dy),$$
$$g(y) = -\log \int_{\mathcal{X}} e^{f(x)-c(x,y)}\, P(dx)$$

- Alternatingly solve the two equations (Gauss–Seidel)?

Sinkhorn's Algorithm

Algorithm (Sinkhorn–Knopp, IPFP)

Initialize at $f_0 := 0$. *Define for* $t = 0, 1, \ldots$ *the iterates*

$$g_t(y) := -\log \int_{\mathcal{X}} e^{f_t(x) - c(x,y)}\, P(dx),$$

$$f_{t+1}(x) := -\log \int_{\mathcal{Y}} e^{g_t(y) - c(x,y)}\, Q(dy).$$

- Can also be motivated as block-coordinate ascent in the dual problem

$$\sup_{f,g} G(f,g), \quad G(f,g) = \int f(x)\, P(dx) + \ldots$$

$$g_t = \arg\max G(f_t, \cdot), \qquad f_{t+1} = \arg\max G(\cdot, g_t)$$

- Strong concavity $\Rightarrow$ linear convergence , i.e., $\|f_t - f\|_\infty \le \lambda^t$ for some $\lambda < 1$.
  (Carlier '22)

Smoothness of EOT potentials

First-Order Condition as a Convolution:

$$e^{-f(x)} = \int_{\mathcal{Y}} e^{g(y)} e^{-c(x,y)}\, Q(dy)$$

- Similarly for $e^{-g(y)}$
- f, g are "as smooth as c"
- Uniform over marginals:

$$\|f\|_{\mathcal{C}^k(\mathcal{X})} + \|g\|_{\mathcal{C}^k(\mathcal{Y})} \leq C_k, \quad k \in \mathbb{N}$$

with C_k depending on $\|c\|_{\mathcal{C}^k(\mathcal{X}\times\mathcal{Y})}$ but not on P, Q

Sample Complexity of EOT

- Let $X_1, \ldots, X_n$ be i.i.d. samples from P and $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$
- Similarly $Q_n = \frac{1}{n} \sum_{i=1}^n \delta_{Y_i}$ (independent)
- EOT avoids curse of dimensionality:

$$\mathbb{E}[|\mathrm{EOT}(P, Q) - \mathrm{EOT}(P_n, Q_n)|] \lesssim n^{-\frac{1}{2}}$$

- Large literature, sample complexity results for optimal cost, potentials, couplings
- As well as central limit theorems
- Mostly for $c \in C^\infty$ and based on uniform smoothness of potentials
- Alternative approach for non-smooth cost using strong concavity of dual problem

Genevay et al. '19, Mena, Weed '19, del Barrio et al. '23, Goldfeld et al. '24, González-Sanz et al. '24, Goldfeld et al. '24b, Rigollet, Stromme '25, González-Sanz, Hundrieser '23

Overspreading and Numerical Issues

Overspreading:

- Recall that optimal coupling π has density of the form

$$\frac{d\pi_\varepsilon}{d(P \otimes Q)}(x, y) = \exp\left(\frac{f(x) + g(y) - c(x, y)}{\varepsilon}\right)$$

- In particular, $\pi_\varepsilon \sim P \otimes Q$, hence π_ε always has "full support"
- Even if the optimal transport that it approximates is given by a Monge map
- Related to infinite slope of $\varphi(t) = t \log t$ at $t = 0$

Numerical Issues for $\varepsilon \ll 1$:

$$g_t(y) := -\varepsilon \log \int_{\mathcal{X}} e^{\frac{f_t(x) - c(x,y)}{\varepsilon}} P(dx),$$

$$f_{t+1}(x) := -\varepsilon \log \int_{\mathcal{Y}} e^{\frac{g_t(y) - c(x,y)}{\varepsilon}} Q(dy).$$

Outline

Quadratically Regularized Optimal Transport

- Replace KL divergence by squared L^2-norm (or χ^2 divergence) :

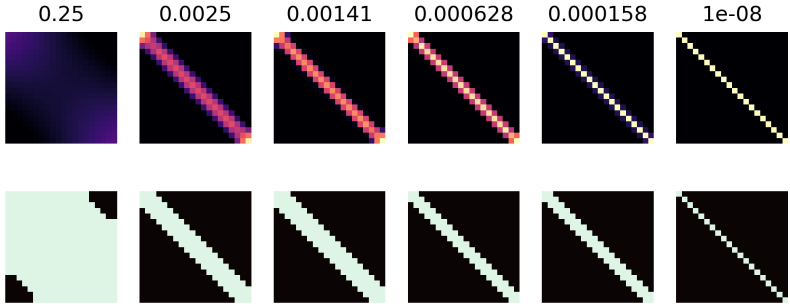
$$\mathrm{QOT}_\varepsilon(P,Q) := \inf_{\pi\in\Pi(P,Q)} \int_{\mathcal{X}\times\mathcal{Y}} c\, d\pi + \frac{\varepsilon}{2}\left\| \frac{d\pi}{d(P\otimes Q)} \right\|^2$$

- Solution is the $L^2(P\otimes Q)$-projection of $-c/\varepsilon$ onto set of couplings
- $\mathrm{QOT}_\varepsilon(P,Q) - \mathrm{OT}(P,Q) = \mathcal{O}(\varepsilon^{\frac{2}{d+2}})$

Muzellec et al. '17, Blondel et al. '18, Essid, Solomon '18, Dessein et al. '18, Seguy et al. '18, Genevay et al. '19, Lorenz et al. '21, Eckstein, Kupper '21, Lorenz, Mahler '22

Eckstein, Nutz '24, Garriz Molina, Gonzalez Sanz, Mordant '24

Sparsity

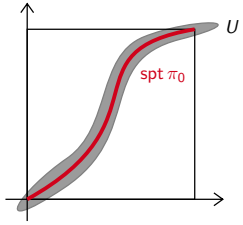


Optimal coupling π_ε (top) and its support (bottom) for different ε

$N = 20$ data points per marginal, $P = Q = N^{-1} \sum_{i=0}^{N-1} \delta_{i/N}$ and $c(x, y) = \|x - y\|^2$

Theoretical Results on Sparsity

- For quadratic cost $c(x, y) = \|x - y\|^2$ and marginals P, Q with densities
- $\mathrm{spt}(\pi_\varepsilon)$ is in a neighborhood of the Monge graph for small ε



- For $d = 1$, sections of $\mathrm{spt}(\pi_\varepsilon)$ shrink at rate $\varepsilon^{\frac{1}{3}}$
- For general d, diameter of sections conjectured to shrink at rate $\varepsilon^{\frac{1}{d+2}}$
- Known only in very special cases ($P = Q$ or radially symmetric marginals)
- Non-sharp upper bounds are known in general

Nutz '24, Wiesel, Xu '24, González-Sanz, Nutz '24

Dual Problem for QOT

The dual problem is

$$\sup_{f\in\mathcal{C}(\mathcal{X}),g\in\mathcal{C}(\mathcal{Y})} \int \left\{ f(x)+g(y)-\frac{1}{2\varepsilon}\big((f(x)+g(y)-c(x,y)\big)_+^2 \right\} P(dx)Q(dy)$$

- Optimizers f, g again called potentials
- Dual problem not strictly/strongly concave
- Potentials are non-unique in general
- Uniqueness (up to constant) holds if one marginal has connected support
- Potentials again give the density of the optimal coupling:

$$\frac{d\pi}{d(P\otimes Q)}(x,y)=\frac{1}{\varepsilon}\big(f(x)+g(y)-c(x,y)\big)_+$$

- $(\cdot)_+$ creates sparse support

Non-Smoothness of QOT Potentials

- Dual problem is not strongly concave
- Potentials are non-unique in general
- Potentials (f, g) satisfy a first-order condition:

$$\varepsilon = \int_{\mathcal{Y}} \big(f(x) + g(y) - c(x,y)\big)_+ \, Q(dy),$$
$$\varepsilon = \int_{\mathcal{X}} \big(f(x) + g(y) - c(x,y)\big)_+ \, P(dx)$$

- Limited smoothness: formally,

$$\nabla f(x) = \frac{\int_{\{f(x)+g(\cdot)-c(x,\cdot)>0\}} \nabla_x c(x,\cdot)\, dQ}{Q\{f(x) + g(\cdot) - c(x,\cdot) > 0\}} = \text{mean}(\nabla_x c(x,\cdot))$$

  under the measure $Q|_{\{f(x)+g(\cdot)-c(x,\cdot)>0\}}$
- Even if c is smooth, ∇f may be discontinuous
- Better regularity if marginals are smooth, but not for empirical marginals
- Expect worse sample complexity than EOT

Prior Result on Sample Complexity

Reminder

- $X_1, \ldots, X_n$ be i.i.d. samples from P and $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$
- Similarly $Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ (independent)
- Empirical problem $\mathrm{QOT}(P_n, Q_n)$ with optimal coupling π_n and potentials (f_n, g_n)

Prior Result (Bayraktar, Eckstein, Zhang '25)

- General φ-divergence with convex conjugate $\psi \in \mathcal{C}^k$
- Bound on expected absolute cost difference:

$$\mathbb{E}[|\mathrm{ROT}(P, Q) - \mathrm{ROT}(P_n, Q_n)|] \lesssim n^{-k/d}, \quad \text{for } d > 2k, \quad \text{if } c \in \mathcal{C}^k, \quad \psi \in \mathcal{C}^k$$

- For $\psi \in \mathcal{C}^\infty$ (e.g., EOT), dimension dependence can be eliminated
- For QOT: $\psi \in \mathcal{C}^{1,1}$. Pretend this is $\mathcal{C}^2$ ... giving $\lesssim n^{-2/d}$
- Suggests same curse of dimensionality as unregularized OT

Sample Complexity of QOT

But actually:

- Rate does not depend on the dimension d
- All key objects converge at parametric rate $\lesssim n^{-1/2}$
- and even satisfy CLTs

Assumptions:

- **Marginals:** Population marginals P, Q have compact and convex supports Ω, Ω', and bounded densities.
- **Quadratic transport cost:** $c(x, y) = \frac{1}{2}\|x - y\|^2$

González-Sanz, del Barrio, Nutz '25, González-Sanz, Eckstein, Nutz '25

CLT for Costs: We have

$$\sqrt{n}(\mathrm{QOT}(P_n, Q_n) - \mathrm{QOT}(P, Q)) \xrightarrow{w} N(0, \sigma^2)$$

where σ^2 admits a formula in terms of (f, g)

CLT for Potentials: Let (f_n, g_n) be empirical potentials and (f, g) the population potentials (unique modulo constant). Then

$$\sqrt{n}\begin{pmatrix} f_n - f \\ g_n - g \end{pmatrix} \xrightarrow{w} (\dots) \quad \text{holds in } \mathcal{C}(\Omega) \times \mathcal{C}(\Omega')/\sim$$

where $(\dots)$ has an expression in terms of a Gaussian process and (f, g).

CLT for Couplings: For any bounded and measurable function $\eta : \Omega \times \Omega' \to \mathbb{R}$,

$$\sqrt{n}\left(\int \eta d(\pi_n - \pi)\right) \xrightarrow{w} N(0, \sigma^2(\eta))$$

where $\sigma^2(\eta)$ admits a formula in terms of (f, g).

González-Sanz, del Barrio, Nutz '25

Error Bound (EB) and Polyak–Łojasiewicz (PL) Inequality

Assumptions:

- P has convex support, density bounded above and below
- Transport cost c is L-Lipschitz

$\rightarrow$ Unique potentials (f_*, g_*)

Consider the dual objective function $\Gamma : L^2(P) \times L^2(Q) \to \mathbb{R}$,

$$\Gamma(f,g) := \int \left(f(x) + g(y) - \frac{1}{2\varepsilon} \left(f(x) + g(y) - c(x,y) \right)_+^2 \right) d(P \otimes Q)(x,y).$$

The gradient at $(f,g) \in L^2(P) \times L^2(Q)$ is

$$\mathrm{D}\Gamma(f,g) = \begin{pmatrix} \mathrm{D}_1\Gamma \\ \mathrm{D}_2\Gamma \end{pmatrix}(f,g) = \begin{pmatrix} 1 - \frac{1}{\varepsilon} \int \left(f(\cdot) + g(y) - c(\cdot,y) \right)_+ dQ(y) \\ 1 - \frac{1}{\varepsilon} \int \left(f(x) + g(\cdot) - c(x,\cdot) \right)_+ dP(x) \end{pmatrix}$$

Error Bound (EB) and Polyak–Łojasiewicz (PL) Inequality

Theorem: For all $(f, g) \in L^\infty(\Omega) \times L^\infty(\Omega')$

Error bound:

$$\|f \oplus g - f_* \oplus g_*\|_{L^2(P\otimes Q)} \leq \gamma \max(\|f \oplus g - f_* \oplus g_*\|_\infty, \varepsilon)\, \|\mathrm{D}\Gamma(f, g)\|_{L^2(P)\times L^2(Q)}$$

PL inequality:

$$\|\mathrm{D}\Gamma(f, g)\|^2_{L^2(P)\times L^2(Q)} \geq \frac{1}{\gamma \max\left(\|f \oplus g - f_* \oplus g_*\|_\infty, \varepsilon\right)} \left(\mathrm{QOT}_\varepsilon(P, Q) - \Gamma(f, g)\right)$$

where

$$\gamma = 16 \left(\delta_P^{-1} \max\left(\frac{8L}{\varepsilon}, 1\right)^d\right) \frac{\Lambda_P^2}{\lambda_P^2} \frac{\left(\lceil 8L \operatorname{diam}(\Omega)/\varepsilon \rceil\right)^{d+2}}{\inf_{y\in\Omega'} Q\left(B_{\frac{\varepsilon}{8L}}(y)\right)}$$

with constants s.t. $P(B_r(x)) \geq \delta_P \min(r^d, 1)$ and $\lambda_P \leq dP/dx \leq \Lambda_P$

González-Sanz, Nutz, Riveros Valdevenito '26

Linear Convergence of Algorithms (GD, CGD)

Gradient ascent with step size $\eta > 0$: initialize at $(f_0, g_0) \in L^\infty(\Omega) \times L^\infty(\Omega')$, iterate

$$\begin{pmatrix} f_{n+1} \\ g_{n+1} \end{pmatrix} = \begin{pmatrix} f_n \\ g_n \end{pmatrix} + \eta \, \mathrm{D}\Gamma \begin{pmatrix} f_n \\ g_n \end{pmatrix}, \quad n \geq 0.$$

For every $\eta < \varepsilon$, the suboptimality gap $\Delta_n := \mathrm{QOT}_\varepsilon(P, Q) - \Gamma(f_n, g_n)$ satisfies

$$\Delta_n \leq (1-q)^n \Delta_0 \qquad \text{for} \qquad q := \frac{\eta\left(1 - \frac{\eta}{\varepsilon}\right)}{\gamma \max(2\|f_0 \oplus g_0 - f_* \oplus g_*\|_\infty, \varepsilon)}.$$

Coordinate gradient ascent:

$$f_{n+1} := f_n + \eta \, \mathrm{D}_1\Gamma \begin{pmatrix} f_n \\ g_n \end{pmatrix}, \qquad g_{n+1} := g_n + \eta \, \mathrm{D}_2\Gamma \begin{pmatrix} f_{n+1} \\ g_n \end{pmatrix}, \qquad n \geq 0.$$

For every $\eta \in (0, \varepsilon/\sqrt{2})$,

$$\Delta_n \leq (1-q)^n \Delta_0 \qquad \text{for} \qquad q := \frac{\eta\left(1 - \frac{\eta}{2\varepsilon}\right)}{2\gamma \max(2\|f_0 \oplus g_0 - f_* \oplus g_*\|_\infty, \varepsilon)}.$$

Linear Convergence of Algorithms (CD)

Coordinate ascent: initialize at $g_0 \in L^\infty(\Omega')$, iterate

$$f_n := \arg\max_f \Gamma(f, g_n), \qquad g_{n+1} := \arg\max_g \Gamma(f_n, g), \qquad n \geq 0.$$

This is equivalent to solving the first-order conditions:

$$\text{define } f_n(x) \text{ via} \quad \varepsilon = \int \big(f_n(x) + g_n(y) - c(x,y)\big)_+ \, dQ(y),$$
$$\text{then } g_{n+1}(y) \text{ via} \quad \varepsilon = \int \big(f_n(x) + g_{n+1}(y) - c(x,y)\big)_+ \, dP(x).$$

Corresponds to Sinkhorn's algorithm of EOT. Suboptimality gap satisfies

$$\Delta_n \leq (1-q)^n \Delta_0 \qquad \text{for} \qquad q := \frac{\varepsilon}{2\gamma \max(2\|g_0 - g_*\|_\infty, \varepsilon)}.$$

González-Sanz, Nutz, Riveros Valdevenito '25 & '26, Lorenz, Mahler '19

Conclusions

- L^2-regularization yields sparse approximations of optimal transport and is nicer than you may think
- Sample complexity better than expected
- Computation, too
- Amenable to theoretical analysis
- Limited theoretical results, many open questions