Interpolating between Optimal Transport & KL regularized Optimal Transport with Rényi Divergences

joint work with

Jonas Bresch, TU Berlin

University of South Carolina, Columbia, 12.09.2024.

Graduate Colloquium (Alec Helm, Jonah Klein).

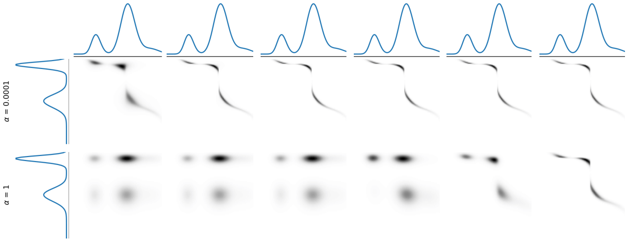
Optimal transport (OT) **distance on probability measures** via **transport plan**.

Problem: $O(N^3)$ for N samples.

Solution: Entropic OT (Cuturi, NeurIPS'13): add ε times KL-regularizer to OT problem for $\varepsilon > 0$.

Sinkhorn algorithm $\rightsquigarrow O(N^{1+\frac{1}{d}} \ln(N))$.

Problem in practice: need ε very small to get accurate plan, but $\rightsquigarrow$ numerical instabilities.

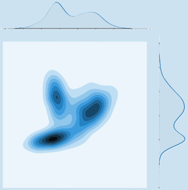


Our solution: Add instead ε times different (=α-Rényi) regularizer and let $\alpha \searrow 0$ instead of $\varepsilon \searrow 0$.

1. Tsallis divergence and α-Rényi divergence

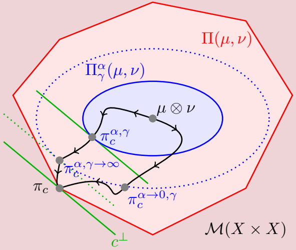
https://repository.aip.org/islandora/object/nbla:310720
https://en.wikipedia.org/wiki/Constantino_Tsallis#/media/File:Constantino_Tsallis_February_2010.jpg.

2. Optimal transport and its regularization

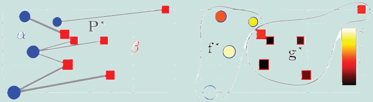


dsweber2.github.io/Optimal-Transport-Information-Geometry/

3. Rényi-regularized OT



4. Dual formulation



Péyre, Cuturi, 2019

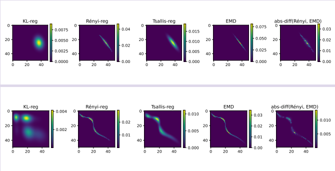
$$\min_{\pi\in\mathcal{P}(X^2)} \langle c,\pi\rangle + \varepsilon R_\alpha(\pi)$$

$$\max_{h\in\mathcal{C}(X^2)} \langle h,\pi\rangle - \varepsilon \ln(\gamma_h^\alpha)$$

5. Interpolation properties

$$\begin{tikzcd}
\mathrm{OT} & & \langle c,\cdot\rangle \\
& \mathrm{OT}_{\varepsilon,\alpha} \arrow[ul, "\varepsilon\searrow 0"] \arrow[ur, "\varepsilon\to\infty"] \arrow[dl, "\alpha\searrow 0"] \arrow[dr, "\alpha\nearrow 1"] & \\
\mathrm{OT} & & \mathrm{OT}_\varepsilon
\end{tikzcd}$$

6. Numerical results



Definition (α-Rényi divergence)

The α-Rényi divergence of *order* $\alpha \in (0,1)$ is

$$R_\alpha \colon \mathcal{P}(X) \times \mathcal{P}(X) \to [0, \infty], \qquad (\mu \mid \nu) \mapsto \frac{1}{\alpha - 1} \ln \left(\int_X \left(\frac{\rho_\mu(x)}{\rho_\nu(x)} \right)^\alpha \mathrm{d}\nu(x) \right).$$

where for $\sigma \in \mathcal{P}(X)$, ρ_σ is the density w.r.t. $\frac{1}{2}(\mu + \nu)$, and $\ln(0) \coloneqq -\infty$.

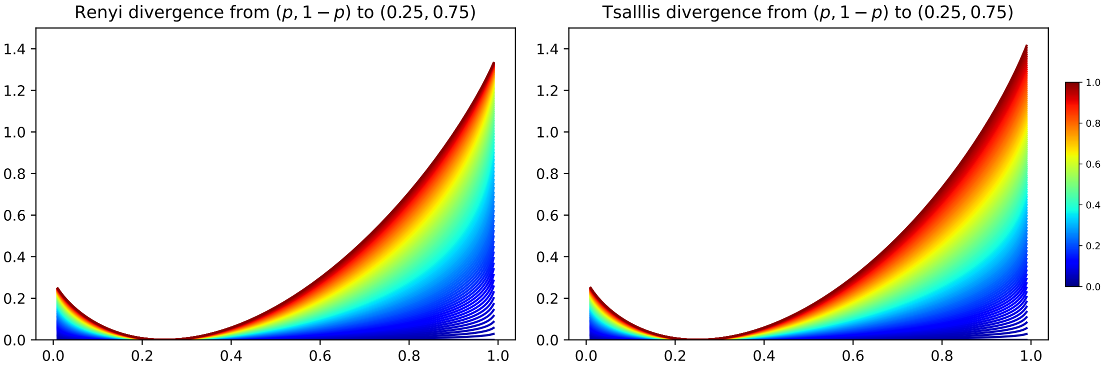
Muzellec et. al (AAAI 2017) examine Tsallis-regularized OT.

Definition (q-Tsallis divergence)

The q-Tsallis divergence of *order* $q > 0$, $q \neq 1$, is

$$T_q = \frac{1}{q-1} \left[\exp\left((q-1) R_q \right) - 1 \right] \colon \mathcal{P}(X) \times \mathcal{P}(X) \to [0, \infty], \qquad (\mu \mid \nu) \mapsto \frac{1}{q-1} \left[\int_X \left(\frac{\rho_\mu(x)}{\rho_\nu(x)} \right)^q \mathrm{d}\nu(x) - 1 \right]$$

Tsallis = 1st order approximation of Rényi since $\ln(y) \approx y - 1$ (1st order Taylor).



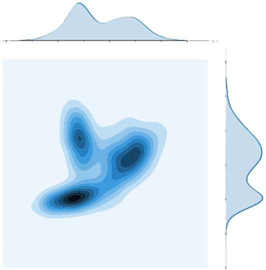
Theorem (Properties of the Rényi divergence)

- ***Divergence property:*** *$R_\alpha(\mu \mid \nu) \geq 0$ and $R_\alpha(\mu \mid \nu) = 0$ if and only if $\mu = \nu$.*
- *R_α is **nondecreasing** and continuous in $\alpha \in [0, 1]$ with $\lim_{\alpha \nearrow 1} R_\alpha = \mathrm{KL}$ pointwise.*
- *R_α jointly **convex**, jointly **weakly lower semicontinuous** for $\alpha \in (0, 1]$.*

Let (X, d) metric space, with d lower semicontinuous.

Let $p \in [1, \infty)$, $\mathcal{P}(X)$ the set of probability measures.

$$\mathcal{P}_p(X) := \left\{ \mu \in \mathcal{P}(X) : \int_X d(x, x_0)^p \, \mathrm{d}\mu(x) < \infty \right\}, \quad x_0 \in X.$$



On $\mathcal{P}_p(X)$, the **Wasserstein-p metric** is

$$\boxed{\mathrm{OT}(\mu, \nu)^p = \min_{\pi \in \Pi(\mu, \nu)} \int_{X \times X} d(x, y)^p \, \mathrm{d}\pi(x, y), \quad \mu, \nu \in \mathcal{P}_p(X),}$$

where the **transport polytope** is

$$\Pi(\mu, \nu) := \{\pi \in \mathcal{P}(X \times X) : \pi(A \times X) = \mu(A), \pi(X \times A) = \nu(A) \ \forall A\}$$

The product measure $\mu \otimes \nu \in \Pi(\mu, \nu)$.

Notation: $\langle f, \mu \rangle := \int_X f(x) \, \mathrm{d}\mu(x)$, so we can write $\mathrm{OT}(\mu, \nu)^p = \min\{\langle d^p, \pi \rangle : \pi \in \Pi(\mu, \nu)\}$.

Regularizer: Kullback-Leibler divergence

$$\mathrm{KL}(\cdot \mid \mu \otimes \nu) \colon \Pi(\mu, \nu) \to [0, \infty),$$

$$\pi \mapsto \int_{X \times X} \ln \left(\frac{\mathrm{d}\pi}{\mathrm{d}\mu \otimes \nu}(x, y) \right) \mathrm{d}\mu(x)\, \mathrm{d}\nu(y)$$

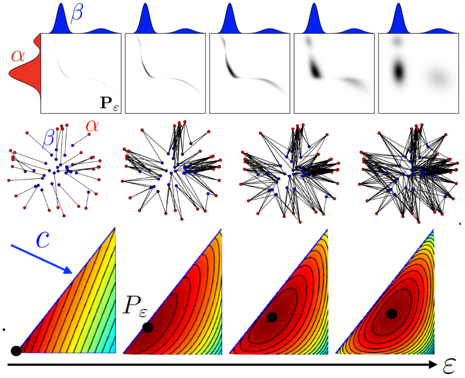
KL-regularized OT:

$$\mathrm{OT}_\varepsilon(\mu, \nu) := \min_{\pi \in \Pi(\mu, \nu)} \langle d^p, \pi \rangle + \varepsilon\, \mathrm{KL}(\pi \mid \mu \otimes \nu)$$

$$= \max_{f, g \in \mathcal{C}(X)} \left\langle f \oplus g - \varepsilon \exp\left(-\frac{1}{\varepsilon}(f \oplus g - d^p) \right), \mu \otimes \nu \right\rangle.$$

$(f \oplus g)(x, y) := f(x) + g(y)$ for $f, g \in \mathcal{C}(X)$.

Primal-dual relation: $\hat{\pi}^\varepsilon = \exp\left(\frac{\hat{f} \oplus \hat{g} - d^p}{\varepsilon} \right) \cdot \mu \otimes \nu$



Here, $c = d^p$.

$$\operatorname{argmin} \{\mathrm{KL}(\pi \mid \mu \otimes \nu) : \langle d^p, \pi \rangle = \mathrm{OT}(\mu, \nu)\} \xleftarrow{\varepsilon \searrow 0} \hat{\pi}_\varepsilon \xrightarrow{\varepsilon \to \infty} \mu \otimes \nu$$

$$\mathrm{OT}(\mu, \nu) \xleftarrow{\varepsilon \searrow 0} \mathrm{OT}_\varepsilon(\mu, \nu) \xrightarrow{\varepsilon \to \infty} \langle d^p, \mu \otimes \nu \rangle$$

Discretize $X \approx (x_i)_{i=1}^N$

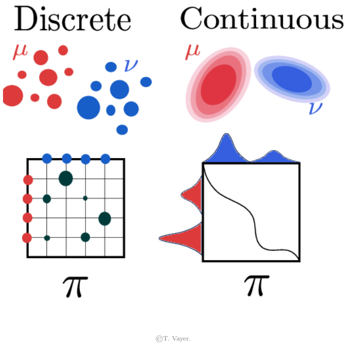
$\mu, \nu \in \mathcal{P}(X)$ become vectors $\boldsymbol{r} := (\mu(x_i))_{i=1}^N, \boldsymbol{c} := (\nu(x_i))_{i=1}^N \in \Sigma_N$, where

$$\Sigma_N := \{x \in [0,1]^N : \sum_{i=1}^N x_i = 1\}.$$

cost matrix: $\boldsymbol{M} := (d(x_i, x_j)^p)_{i,j=1}^N$.

transport polytope

$$\Pi(\boldsymbol{r}, \boldsymbol{c}) := \{\boldsymbol{P} \in \Sigma_{N \times N} : \boldsymbol{P} \mathbb{1}_N = \boldsymbol{r}, \boldsymbol{P}^{\mathrm{T}} \mathbb{1}_N = \boldsymbol{c}\}$$



Optimal transport plan as KL-projection of Gibbs kernel

$$\hat{\boldsymbol{P}}^\varepsilon = \operatorname*{argmin}_{\boldsymbol{P} \in \Pi(\boldsymbol{r}, \boldsymbol{c})} \mathrm{KL}\left(\boldsymbol{P} \,\middle|\, \exp\left(\frac{-\boldsymbol{M}}{\varepsilon}\right)\right)$$

Sinkhorn algorithm finds this projection via matrix scaling.

First: A different way of regularizing

For regularization parameter $\gamma \in [0, \infty]$ and $\alpha \in (0, 1)$, the **restricted transport polytope**,

$$\Pi_\gamma^\alpha(\mu, \nu) \coloneqq \{\pi \in \Pi(\mu, \nu) : R_\alpha(\pi \mid \mu \otimes \nu) \le \gamma\},$$

is weakly compact, since $R_\alpha(\cdot \mid \mu \otimes \nu)$ is weakly lsc. and $\Pi(\mu, \nu)$ is weakly compact.

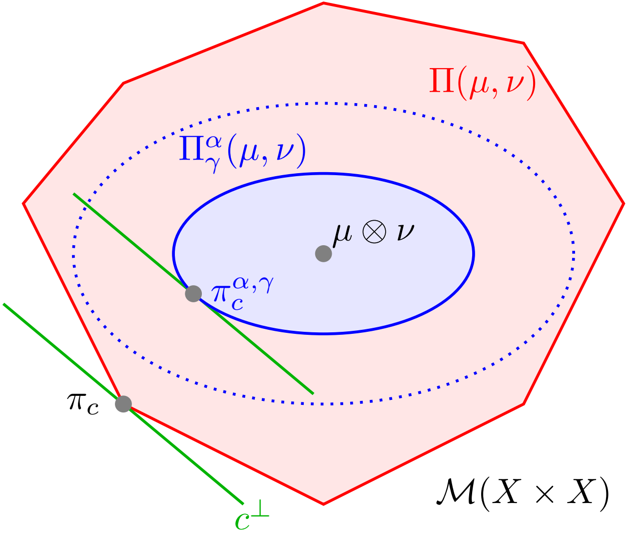
Definition (Rényi-Sinkhorn distance)

The **Rényi-Sinkhorn distance** between $\mu, \nu \in \mathcal{P}_p(X)$ is

$$\mathrm{d}_{\gamma,\alpha} \colon \mathcal{P}_p(X) \times \mathcal{P}_p(X) \to \mathbb{R}, \qquad (\mu, \nu) \mapsto \min\left\{\langle d^p, \pi\rangle^{\frac{1}{p}} : \pi \in \Pi_\gamma^\alpha(\mu, \nu)\right\}. \tag{1}$$

Theorem (Bresch, S. '24)

- *For $(\mu, \nu) \in \mathcal{P}_p(X)$, the optimization problem (1) is* ***convex*** *and has a* ***unique*** *minimizer.*
- $\mathcal{P}_p(X)^2 \ni (\mu, \nu) \mapsto \mathbb{1}_{[\mu \neq \nu]}(\mu, \nu)\mathrm{d}_{\gamma,\alpha}(\mu, \nu)$ *is a* ***metric*** *for $\alpha \in (0, 1)$, $\gamma \in [0, \infty]$.*



Transport polytope $\Pi(\mu, \nu)$, restricted transport polytope $\Pi_\gamma^\alpha(\mu, \nu)$ for $c = d^p$.

(Plot inspired by (Cuturi, 2013).)

Instead of restricting the problems domain, penalize the Rényi divergence constraint in (1).

DEFINITION (DUAL RÉNYI-DIVERGENCE-SINKHORN DISTANCE)

The **dual Rényi-Divergence-Sinkhorn distance** for $\alpha \in (0,1)$, $\varepsilon \in [0,\infty)$ is

$$\mathrm{d}^{\alpha,\varepsilon} \colon \mathcal{P}_p(X) \times \mathcal{P}_p(X) \to \mathbb{R}, \qquad (\mu,\nu) \mapsto \langle d^p, \pi^{\alpha,\varepsilon}(\mu\nu)\rangle^{\frac{1}{p}},$$

$$\text{where } \pi^{\alpha,\varepsilon}(\mu,\nu) \in \operatorname{argmin}\left\{\langle d^p, \pi\rangle + \varepsilon R_\alpha(\pi \mid \mu \otimes \nu) : \pi \in \Pi(\mu,\nu)\right\}. \tag{2}$$

THEOREM (LAGRANGIAN POINT OF VIEW AND PRE-METRIC [BRESCH, S. '24])

Let $(\mu,\nu) \in \mathcal{P}_p(X)$.

- *The optimization problem* (2) *is **convex** and has a **unique** minimizer.*
- *Rényi-Sinkhorn* $\mathrm{d}_{\gamma,\alpha}(\mu,\nu)$ *and dual Rényi-Sinkhorn* $\mathrm{d}^{\alpha,\lambda}(\mu,\nu)$ *are **equivalent**:*

$$\text{for } \gamma > 0, \text{ there exists } \varepsilon \in [0,\infty], \text{ such that } \langle d^p, \pi^{\alpha,\varepsilon}(\mu,\nu)\rangle = \mathrm{d}_{\gamma,\alpha}(\mu,\nu)^p.$$

Definition (Rényi-regularized OT [Bresch, S. '24])

The **Rényi-regularized OT** problem is

$$\mathrm{OT}_{\varepsilon,\alpha}\colon \mathcal{P}_p(X) \times \mathcal{P}_p(X) \to [0,\infty), \ (\mu,\nu) \mapsto \min_{\pi \in \Pi(\mu,\nu)} \langle c, \pi \rangle + \varepsilon R_\alpha(\pi \mid \mu \otimes \nu).$$

Theorem ($\mathrm{OT}_{\varepsilon,\alpha}$ is a pre-metric [Bresch, S. '24])

$\mathcal{P}_p(X)^2 \ni (\mu,\nu) \mapsto \mathbb{1}_{[\mu \neq \nu]} \mathrm{OT}_{\varepsilon,\alpha}(\mu,\nu)$ *is a metric for* $\alpha \in (0,1), \varepsilon \in [0,\infty)$.

Lemma (Monotonicity of Rényi regularized OT [Bresch, S. '24])

Let $\mu,\nu \in \mathcal{P}_p(X)$, $\alpha, \alpha' \in (0,1)$ *and* $\varepsilon, \varepsilon' \geq 0$ *with* $\alpha > \alpha'$ *and* $\varepsilon < \varepsilon'$. *Then, we have*

$$\mathrm{OT}_{\varepsilon',\alpha}(\mu,\nu) \geq \mathrm{OT}_{\varepsilon,\alpha}(\mu,\nu) \geq \mathrm{OT}_{\varepsilon,\alpha'}(\mu,\nu).$$

Dual formulation, Representation of $\pi^{\alpha,\lambda}$

From now on: X **compact**. The dual space of all finite signed Borel measures on X, $\mathcal{M}(X)$, is $\mathcal{C}(X)$, the space of real-valued continuous functions on X.

Recall $(f \oplus g)(x, y) \coloneqq f(x) + g(y)$.

Theorem (Dual problem, dual representation [Bresch, S. '24])

We have the strong duality

$$\mathrm{OT}_{\varepsilon,\alpha}(\mu, \nu) = \max_{\substack{f,g \in \mathcal{C}(X) \\ f \oplus g \le d}} \langle f \oplus g, \mu \otimes \nu \rangle - \varepsilon \ln\left(\left\langle (d - f \oplus g)^{\frac{\alpha}{\alpha-1}}, \mu \otimes \nu \right\rangle \right) + C_{\alpha,\lambda}. \tag{3}$$

The optimal dual potentials $\hat{f}, \hat{g} \in \mathcal{C}(X)$ from (3) are unique $\mathrm{supp}(\mu \otimes \nu)$-a.e. up to additive constants and the unique optimal plan is

$$\pi^{\alpha,\varepsilon} \propto (d - \hat{f} \oplus \hat{g})^{\frac{1}{\alpha-1}} \cdot (\mu \otimes \nu).$$

Proof idea. Use Fenchel-Rockafellar theorem, extend objective to $\mathcal{M}(X) \times \mathcal{M}(X)$ by ∞.

Theorem (Interpolation properties).

$$\mathrm{OT}(\mu,\nu) \xleftarrow[\varepsilon \searrow 0]{\alpha \searrow 0 \text{ or}} \mathrm{OT}_{\varepsilon,\alpha}(\mu,\nu)$$

$$\mathrm{OT}_{\varepsilon,\alpha}(\mu,\nu) \xrightarrow{\varepsilon \to \infty} \langle d, \mu \otimes \nu \rangle$$

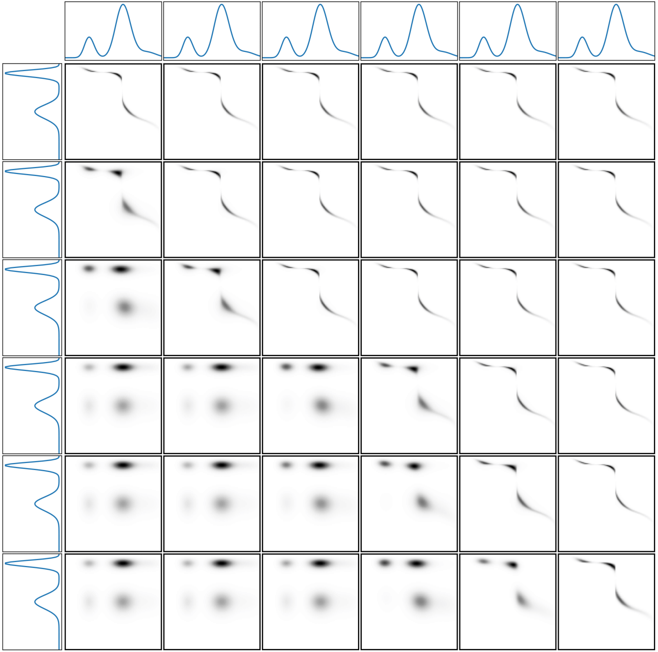
$$\mathrm{OT}_{\varepsilon,\alpha}(\mu,\nu) \xrightarrow{\alpha \nearrow 1} \mathrm{OT}_{\varepsilon}(\mu,\nu)$$

$$\text{some } \pi \in \Pi(\mu,\nu) \text{ s.t. } \langle d, \pi \rangle = \mathrm{OT}(\mu,\nu) \xleftarrow{\alpha \searrow 0} \pi_{\varepsilon,\alpha}(\mu,\nu)$$

$$\pi_{\varepsilon,\alpha}(\mu,\nu) \xrightarrow{\varepsilon \to \infty} \mu \otimes \nu$$

$$\pi_{\varepsilon,\alpha}(\mu,\nu) \xrightarrow{\varepsilon \searrow 0} \operatorname*{argmin}_{\substack{\pi \in \Pi(\mu,\nu),\\ \langle d,\pi \rangle = \mathrm{OT}(\mu,\nu)}} R_\alpha(\pi \mid \mu \otimes \nu)$$

$$\pi_{\varepsilon,\alpha}(\mu,\nu) \xrightarrow{\alpha \nearrow 1} \pi_\varepsilon$$



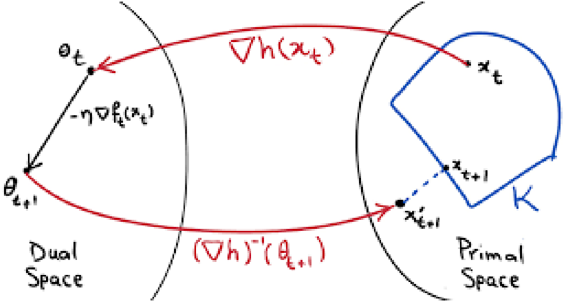
Solve

$$\min_{x \in K} f(x), \qquad \text{where } K \subset \mathbb{R}^n \text{ compact.}$$

via the updates

$$x^{(k+1)} = \operatorname*{argmin}_{y \in K} D_h\left(y \,\middle|\, (\nabla h)^{-1}\left(\nabla h(x^{(k)}) - \eta_k \nabla f(x^{(k)})\right)\right), \qquad x^{(0)} \in K,\ \eta_k > 0, \tag{4}$$

for a convex function $h\colon \mathbb{R}^n \to \mathbb{R}$ with special properties.



https://www.cs.cmu.edu/~15850/notes/lec20.pdf

Choose $K = \Sigma_N$ (probability simplex), $-h =$ Shannon entropy $\implies D_h = \mathrm{KL}$.

Rényi-regularized OT objective

$$\Pi(\boldsymbol{r}, \boldsymbol{c}) \to [0, \infty), \qquad \boldsymbol{P} \mapsto \langle \boldsymbol{M}, \boldsymbol{P} \rangle + \varepsilon R_\alpha(\boldsymbol{P} \mid \boldsymbol{r}\boldsymbol{c}^{\mathrm{T}}).$$

is not Lipschitz continuous, but locally Lipschitz on

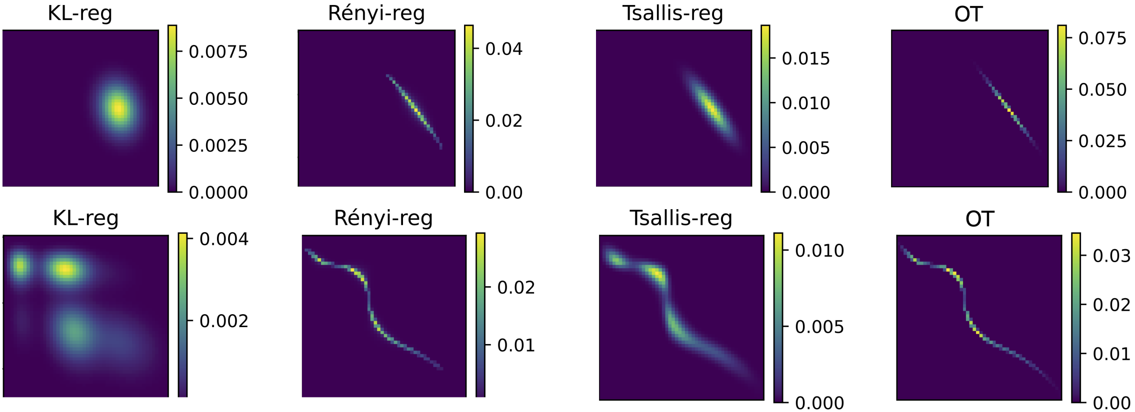
$$\{\boldsymbol{P} \in \Pi(\boldsymbol{r}, \boldsymbol{c}) : \boldsymbol{P}|_{\mathrm{supp}(\boldsymbol{r} \otimes \boldsymbol{c})} > 0\} = \Pi(\boldsymbol{c}, \boldsymbol{r}) \cap \mathbb{R}^N_{>0},$$

which suffices for convergence of a mirror descent with **special step size** $(\eta_k)_{k \in \mathbb{N}}$ (You, Li, 2022).

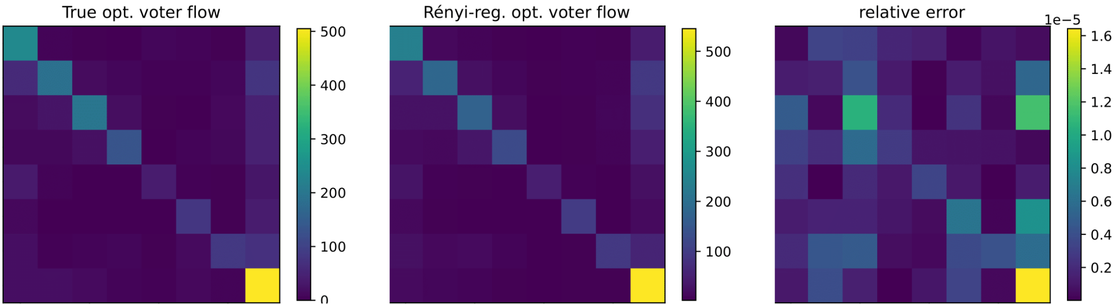
In each iteration one KL projection onto Σ_N (using Sinkhorn algorithm) is performed:

$$\boldsymbol{P}^{(k)} \leftarrow \mathrm{Sinkhorn}\left(\boldsymbol{P}^{(k-1)} \odot \exp\left(-\eta_k \boldsymbol{M} - \frac{\eta_k}{\lambda} \frac{\alpha}{\alpha - 1} \frac{(\boldsymbol{r}\boldsymbol{c}^{\mathrm{T}} \oslash \boldsymbol{P})^{1-\alpha}}{\langle \boldsymbol{P}^\alpha, (\boldsymbol{r}\boldsymbol{c}^{\mathrm{T}})^{1-\alpha} \rangle}\right); \boldsymbol{r}, \boldsymbol{c}\right), \qquad k \in \mathbb{N}$$

Rényi regularization yields more accurate plans



Regularized OT plans for Gaussian *(top)* and Poisson *(bottom)* marginals with regularization parameter $\lambda = 10$, Rényi order $\alpha = 0.01$, Tsallis order: $q = 2$.



regularizer, $\varepsilon = 1$	abs error $\pm$ std	KL error	mean squared error
KL	$2.4221 \times 10^1 \pm 2.848 \times 10^1$	8.422×10^2	9.008×10^4
Tsallis	$9.409 \pm 1.529 \times 10^1$	3.173×10^2	2.063×10^4
OT	$1.845 \times 10^1 \pm 2.358 \times 10^1$	7.655×10^2	5.738×10^4
$\frac{3}{10}$-Renyi	$\mathbf{6.611 \pm 7.868}$	2.128×10^2	6.759×10^3

- **Contribution.** Regularize optimal transport problem using the α-Rényi-divergences R_α for $\alpha \in (0, 1)$. Prove dual formulation and interpolation properties.
- **Prior work.** Regularization with $\mathrm{KL} = \lim_{\alpha \nearrow 1} R_\alpha$ and with q-Tsallis divergence
- **Method.** Solve primal problem with mirror descent and dual problem with subgradient descent.
- **Result.** Rényi-regularized OT plans outperform KL / Tsallis regularized OT plans on real and synthetic data.
- **Novelty.** $R_\alpha \notin \{f\text{-divergence}, \text{Bregman divergence}\}$ and R_α not "separable" due to the logarithm.

Thank you for your attention!

I am happy to take any questions.

Paper link: `https://arxiv.org/abs/2404.18834`

My website: `https://viktorajstein.github.io`

[BT03] Amir Beck and Marc Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Oper. Res. Lett. **31** (2003), no. 3, 167–175.

[Cut13] Marco Cuturi, *Sinkhorn distances: lightspeed computation of optimal transport*, Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (Red Hook, NY, USA), NIPS'13, Curran Associates Inc., 2013, p. 2292–2300.

[MNPN17] Boris Muzellec, Richard Nock, Giorgio Patrini, and Frank Nielsen, *Tsallis regularized optimal transport and ecological inference*, Proceedings of the AAAI conference on Artificial Intelligence (Hilton San Francisco, San Francisco, California, USA), vol. 31, 2017.

[NS21] Sebastian Neumayer and Gabriele Steidl, *From optimal transport to discrepancy*, Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging: Mathematical Imaging and Vision (2021), 1–36.

[NY83] Arkadij Semenovič Nemirovskij and David Borisovich Yudin, *Problem complexity and method efficiency in optimization*, Wiley, New York, 1983.

[PC19] Gabriel Peyré and Marco Cuturi, *Computational optimal transport*, Found. Trends Mach. Learn. **11** (2019), no. 5-6, 355–607.

[Rén61] Alfréd Rényi, *On measures of entropy and information*, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics (Statistical Laboratory of the University of California, Berkeley, California, USA), vol. 4, University of California Press, 1961, pp. 547–562.

[Tsa88] Constantino Tsallis, *Possible generalization of boltzmann-gibbs statistics*, Journal of statistical physics **52** (1988), 479–487.

[vEH14] Tim van Erven and Peter Harremos, *Rényi divergence and Kullback-Leibler divergence*, IEEE Trans. Inf. Theory **60** (2014), no. 7, 3797–3820.

$\mathrm{OT}_{\varepsilon,\alpha}(\mu, \mu) \neq 0$

To obtain valid, differentiable distance:

$$D_{\varepsilon,\alpha}(\mu, \nu) \coloneqq \mathrm{OT}_{\varepsilon,\alpha}(\mu, \nu) - \frac{1}{2}\,\mathrm{OT}_{\varepsilon,\alpha}(\mu, \mu) - \frac{1}{2}\,\mathrm{OT}_{\varepsilon,\alpha}(\nu, \nu).$$

Can be used for gradient flows.