

Interpolating between Optimal Transport and KL regularized Optimal Transport using Rényi Divergences

Jonas Bresch Viktor Stein

- **Contribution.** Regularize OT between probability distributions $\mu, \nu \in \mathcal{P}(\mathbb{X}), \mathbb{X} \subseteq \mathbb{R}^d$

$$\operatorname{argmin}_{\pi \in \Pi(\mu, \nu)} \langle \mathbf{c}, \pi \rangle := \int_{\mathbb{X} \times \mathbb{X}} \mathbf{c}(x, y) \, d\pi(x, y), \quad \text{with cost } \mathbf{c}: \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}_+$$

using the α -Rényi-divergences R_α for $\alpha \in (0, 1)$.

- **Prior work.** Regularization with KL = $\lim_{\alpha \nearrow 1} R_\alpha$ [Cut13] and with q -Tsallis divergence $T_q = (q-1)^{-1} [\exp((q-1)R_q) - 1]$ for $q > 0$ [Muz+17].
- **Method.** Solve primal problem with mirror descent [BT03] and dual problem with subgradient descent.
- **Result.** Rényi-regularized OT plans outperform KL / Tsallis regularized OT plans on real and synthetic data.
- **Novelty.** $R_\alpha \notin \{f\text{-divergence, Bregman divergence}\}$ and R_α not “separable” due to $\ln$.

Preliminaries: OT and Rényi divergence

The transport polytope is

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

and $\mathcal{D}_p := \{\mathbf{c} = d^p \mid d: \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}_+ \text{ is lower semicontinuous metric}\}$ the set of cost functions, where $p \in [1, \infty)$ and $\mathcal{P}_p(\mathbb{X})$ is the set of probability measures with finite p -th moment. The α -Rényi-divergence of order $\alpha \in (0, 1)$ is

$$R_\alpha: \mathcal{P}(\mathbb{X}) \times \mathcal{P}(\mathbb{X}) \rightarrow \mathbb{R}_+, \quad (\mu \mid \nu) \mapsto \frac{1}{\alpha-1} \ln \left(\int_{\mathbb{X}} \left(\frac{d\mu(x)}{d\tau(x)} \right)^\alpha \left(\frac{d\nu(x)}{d\tau(x)} \right)^{1-\alpha} d\tau(x) \right), \quad (1)$$

where a finite reference measure τ , e.g. $\tau = \mu + \nu \in \mathcal{M}(\mathbb{X})_+$, is chosen such that $\mu, \nu \ll \tau$, and $\ln(0) := -\infty$.

Remark. $R_\alpha(\cdot \mid \mu \otimes \nu)$ is well defined on $\Pi(\mu, \nu)$, i.e. $\pi \in \Pi(\mu, \nu)$ implies that $\pi \ll \mu \otimes \nu$.

Theorem. (Pre-convex conjugate of $R_\alpha(\cdot \mid \theta)$)

For $\theta \in \mathcal{M}(\mathbb{X})$ we have

$$[R_\alpha(\cdot \mid \theta)]_*(g) = \begin{cases} \ln(\langle |g|^{\frac{\alpha}{\alpha-1}}, \theta \rangle) + \text{const}_\alpha, & \text{if } g \in \mathcal{C}_0(\mathbb{X}), g < 0, \\ +\infty & \text{else.} \end{cases}$$

Rényi-Regularized OT Problem

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

$$\Pi_\gamma^\alpha(\mu, \nu) := \{\pi \in \Pi(\mu, \nu) : R_\alpha(\pi \mid \mu \otimes \nu) \leq \gamma\}, \quad (2)$$

is weakly compact, since $R_\alpha(\cdot \mid \mu \otimes \nu)$ is weakly lsc and $\Pi(\mu, \nu)$ is weakly compact. The **Rényi-Sinkhorn distance** between $\mu, \nu \in \mathcal{P}_p(\mathbb{X})$ with cost $\mathbf{c} \in \mathcal{D}_p$ is

$$d_{\mathbf{c}, \gamma, \alpha}: \mathcal{P}_p(\mathbb{X}) \times \mathcal{P}_p(\mathbb{X}) \rightarrow \mathbb{R}, \quad (\mu, \nu) \mapsto \min \left\{ \langle \mathbf{c}, \pi \rangle^{\frac{1}{p}} : \pi \in \Pi_\gamma^\alpha(\mu, \nu) \right\}. \quad (3)$$

Remark. (3) is a convex problem, with unique minimizer.

Theorem. $\mathcal{P}_p(\mathbb{X})^2 \ni (\mu, \nu) \mapsto \mathbf{1}_{[\mu \neq \nu]} d_{\mathbf{c}, \gamma, \alpha}(\mu, \nu)$ is a metric for $\alpha \in (0, 1), \gamma \in [0, \infty], \mathbf{c} \in \mathcal{D}_p$.

The Dual Point of View - Penalizing the Constraint

Instead of restricted the domain of the optimization problem, one can penalize the Rényi divergence constraint in (3). The **dual Rényi-Divergence-Sinkhorn distance** for $\alpha \in (0, 1)$, $\lambda \in (0, \infty]$ is

$$d_{\mathbf{c}}^{\alpha, \lambda}: \mathcal{P}_p(\mathbb{X}) \times \mathcal{P}_p(\mathbb{X}) \rightarrow \mathbb{R}, \quad (\mu, \nu) \mapsto \langle \mathbf{c}, \pi_{\mathbf{c}}^{\alpha, \lambda}(\mu, \nu) \rangle^{\frac{1}{p}}, \quad (4)$$

$$\text{where } \pi_{\mathbf{c}}^{\alpha, \lambda}(\mu, \nu) \in \operatorname{argmin} \left\{ \langle \mathbf{c}, \pi \rangle + \frac{1}{\lambda} R_\alpha(\pi \mid \mu \otimes \nu) : \pi \in \Pi(\mu, \nu) \right\}. \quad (5)$$

Theorem. Problem (5) has a unique solution for every $\alpha \in (0, 1), \lambda > 0$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{X})$.

Lagrangian Reformulation for $\Pi_\gamma^\alpha(\mu, \nu)$

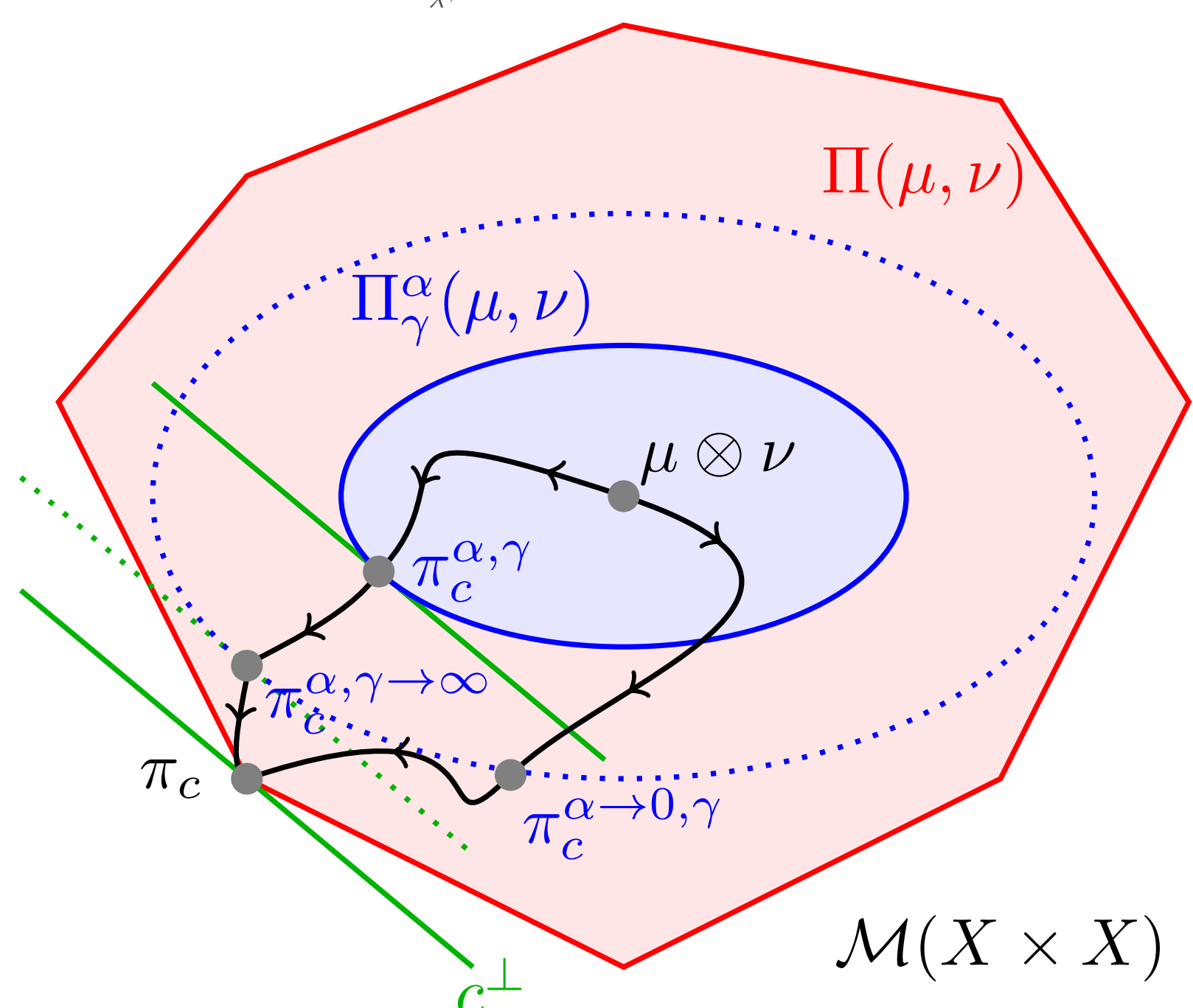
Rényi-Sinkhorn distance $d_{\mathbf{c}, \gamma, \alpha}(\mu, \nu)$ and **dual Rényi-Sinkhorn distance** are equivalent: for $\mu, \nu \in \mathcal{P}_p(\mathbb{X})$ and $\gamma > 0$, there exists $\lambda \in [0, \infty]$, such that $\langle \mathbf{c}, \pi_{\mathbf{c}}^{\alpha, \lambda}(\mu, \nu) \rangle = d_{\mathbf{c}, \gamma, \alpha}(\mu, \nu)^p$.

Rényi-regularized OT pre-metric

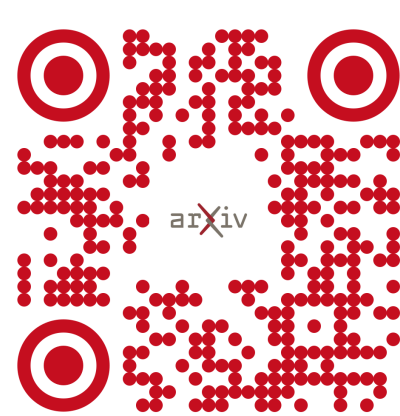
The problem (5) yields the regularized OT problem

$$\operatorname{OT}_{\frac{1}{\lambda}, \alpha}: \mathcal{P}_p(\mathbb{X}) \times \mathcal{P}_p(\mathbb{X}) \rightarrow [0, \infty), \quad (\mu, \nu) \mapsto \min \left\{ \langle \mathbf{c}, \pi \rangle + \frac{1}{\lambda} R_\alpha(\pi \mid \mu \otimes \nu) : \pi \in \Pi(\mu, \nu) \right\}. \quad (6)$$

Theorem. $\mathcal{P}(\mathbb{X})^2 \ni (\mu, \nu) \mapsto \mathbf{1}_{[\mu \neq \nu]} \operatorname{OT}_{\frac{1}{\lambda}, \alpha}(\mu, \nu)$ is a metric for $\alpha \in (0, 1), \lambda \in (0, \infty], \mathbf{c} \in \mathcal{D}_p$.



Transport polytope $\Pi(\mu, \nu)$ with the restricted transport polytope $\Pi_\gamma^\alpha(\mu, \nu)$. Convergence of $\pi_{\mathbf{c}}^{\alpha, \gamma} \rightarrow \pi_{\mathbf{c}}$ to the unregularized OT plan with $\alpha \rightarrow 0$ and $\gamma \rightarrow \infty$. (Plot inspired by [Cut13].)



Dual formulation of the Rényi-Sinkhorn distance

If $\mathbb{X} \subset \mathbb{R}^d$ **compact**, then the Fenchel-Rockafellar theorem yields the strong duality

$$(6) = \max \left\{ \langle \mathbf{f} \oplus \mathbf{g}, \mu \otimes \nu \rangle - \frac{1}{\lambda} \ln \left(\langle (\mathbf{c} - \mathbf{f} \oplus \mathbf{g})^{\frac{\alpha}{\alpha-1}}, \mu \otimes \nu \rangle \right) : \begin{matrix} \mathbf{f}, \mathbf{g} \in \mathcal{C}(\mathbb{X}) \\ \mathbf{f} \oplus \mathbf{g} < \mathbf{c} \end{matrix} \right\} + \text{const}_{\alpha, \lambda}, \quad (7)$$

where $(\mathbf{f} \oplus \mathbf{g})(x, y) := \mathbf{f}(x) + \mathbf{g}(y)$.

Theorem. (Representation of $\pi_{\mathbf{c}}^{\alpha, \lambda}$)

For unique solution $\pi_{\mathbf{c}}^{\alpha, \lambda}(\mu, \nu)$ of (5), and optimal dual potentials $\hat{\mathbf{f}}, \hat{\mathbf{g}} \in \mathcal{C}(\mathbb{X})$ from (7):

$$\pi_{\mathbf{c}}^{\alpha, \lambda} = \frac{(\mathbf{c} - \hat{\mathbf{f}} \oplus \hat{\mathbf{g}})^{\frac{1}{\alpha-1}}}{\langle (\mathbf{c} - \hat{\mathbf{f}} \oplus \hat{\mathbf{g}})^{\frac{1}{\alpha-1}}, \mu \otimes \nu \rangle} \mu \otimes \nu.$$

Corollary. We have $\operatorname{supp}(\pi_{\mathbf{c}}^{\alpha, \lambda}(\mu, \nu)) = \operatorname{supp}(\mu \otimes \nu)$.

Theorem. (Interpolating Properties in $(\lambda, \alpha) \in \mathbb{R}_{>0} \times (0, 1]$)

For fixed $\alpha \in (0, 1)$, we have

$$\operatorname{OT}_{\frac{1}{\lambda}, \alpha}(\mu, \nu) \rightarrow \operatorname{OT}(\mu, \nu) \quad \text{for } \lambda \rightarrow \infty \quad \text{and} \quad \operatorname{OT}_{\frac{1}{\lambda}, \alpha}(\mu, \nu) \rightarrow \langle \mathbf{c}, \mu \otimes \nu \rangle \quad \text{for } \lambda \rightarrow 0.$$

Moreover, for fixed $\lambda > 0$, we have

$$\operatorname{OT}_{\frac{1}{\lambda}, \alpha}(\mu, \nu) \rightarrow \operatorname{OT}(\mu, \nu) \quad \text{for } \alpha \searrow 0 \quad \text{and} \quad \operatorname{OT}_{\frac{1}{\lambda}, \alpha}(\mu, \nu) \rightarrow \operatorname{OT}_{\frac{1}{\lambda}, \alpha'}(\mu, \nu) \quad \text{for } \alpha \nearrow \alpha' \in (0, 1].$$

Remark. Additionally, we can show weak convergence in each case of the solution $\pi_{\mathbf{c}}^{\alpha, \lambda}$.

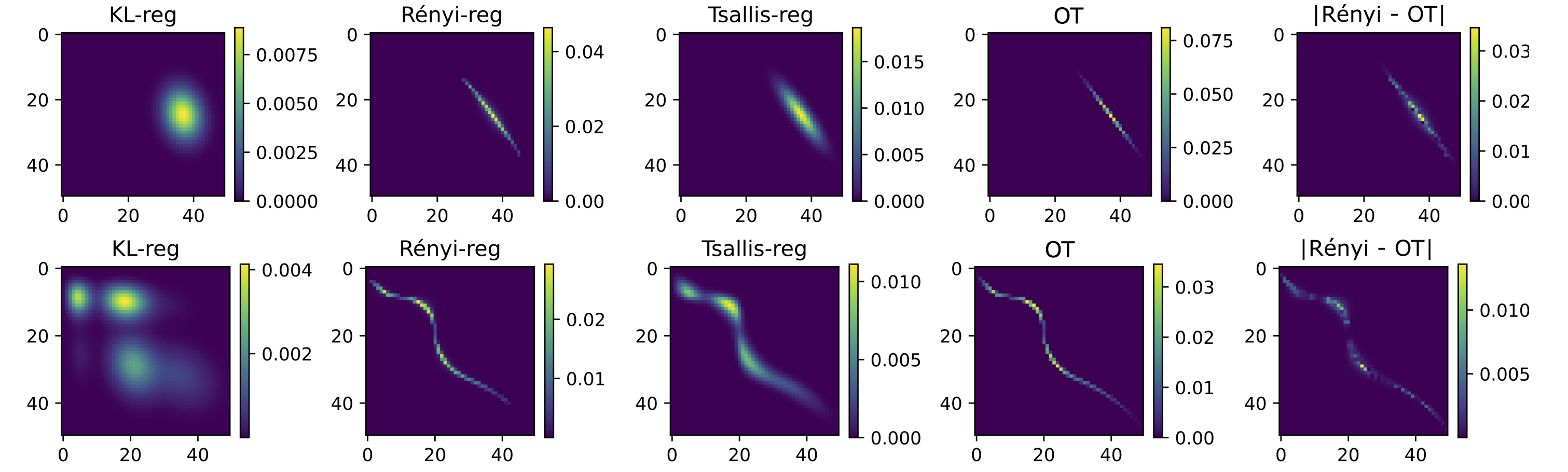
Proposition. The optimal dual solution $(\hat{\mathbf{f}}, \hat{\mathbf{g}}) \in \mathcal{C}(\mathbb{X})^2$ is unique up to additive constants: for two optimal pairs $(\hat{\mathbf{f}}, \hat{\mathbf{g}}), (\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) \in \mathcal{C}(\mathbb{X})^2$ there exists $\gamma \in \mathbb{R}$ such that $\hat{\mathbf{f}} - \tilde{\mathbf{f}} \equiv \gamma \equiv \tilde{\mathbf{g}} - \hat{\mathbf{g}}$.

Numerical Experiments - Better transport plans and voter migration

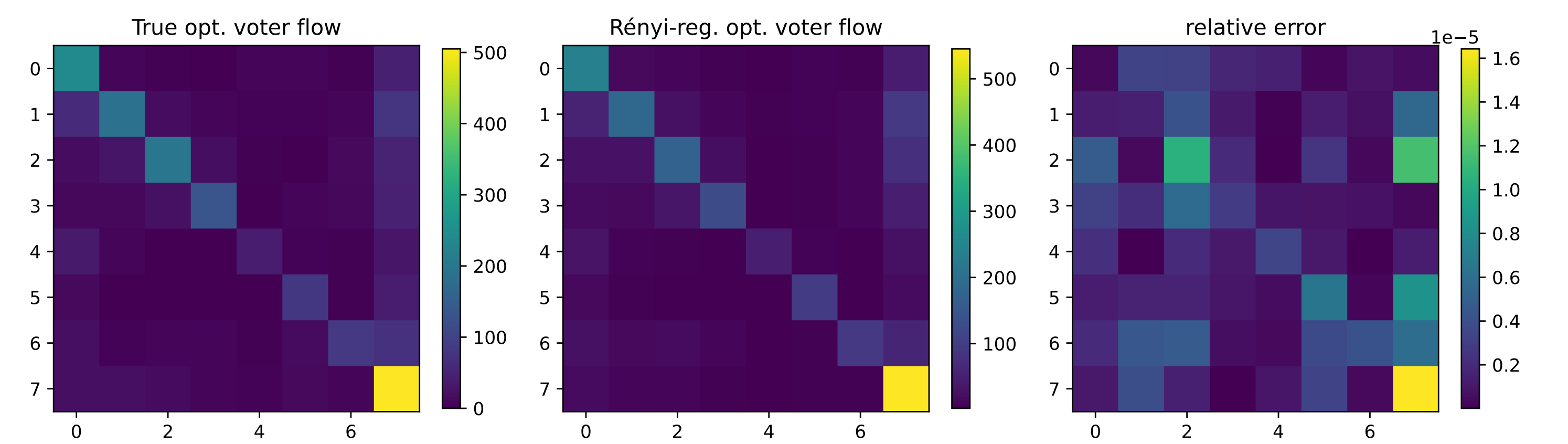
$\Phi_{\mathbf{c}}^{\alpha, \lambda}$ is not Lipschitz continuous in zero, but locally Lipschitz in

$$\{\pi \in \Pi(\mu, \nu) : \pi|_{\operatorname{supp}(\mu \otimes \nu)} > 0\} = \Pi(\mu, \nu) \cap \operatorname{int}(\operatorname{dom} h),$$

where h generates KL divergence $\implies$ modified mirror descent algorithm with Polyak step size converges [YL22]. In each iteration one Sinkhorn-Knopp projection is performed.



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



The true voter migration (left) evaluated by “Infratest dimap”. The Rényi-regularized OT plan $\pi_{\mathbf{c}}^{\alpha, \lambda}(\mu, \nu)$ with squared cost $\mathbf{c}(x, y) = |x - y|^2$, $\lambda = 1$, $\alpha = 0.3$ (middle). μ, ν are the Berlin elections results in 2021 and 2023 (source: “Federal Statistical Office”).

regularizer	absolute error $\pm$ std	KL error	MSE
KL, $\lambda = 1$	$2.4221 \times 10^1 \pm 2.848 \times 10^1$	8.422×10^2	9.008×10^4
Tsallis, $\lambda = 1, q = 1.4$	$9.112 \pm 1.368 \times 10^1$	3.173×10^2	1.724×10^4
None	$1.845 \times 10^1 \pm 2.358 \times 10^1$	7.655×10^2	5.738×10^4
Rényi, $\lambda = 1, \alpha = 0.3$	6.611 ± 7.868	2.128×10^2	6.759×10^3

{ KL, Tsallis, Rényi } regularizers for the optimal $\alpha \in \{k \cdot 10^{-1} : 1 \leq k \leq 9\}$ or $q \in \{k \cdot 10^{-1} : k \in \mathbb{N}\}$ each.

Conclusion

- Generalization of KL-regularized OT.
- Hard constraint from primal formulation stays hard constraint in dual problem, unlike in KL-regularized OT.
- $\alpha \ll 1$ small yields thin regularized OT plans, where $\alpha \searrow 0$ yields the unregularized OT plan (new convergence property).
- Fast computation via mirror descent.

Further work. Rényi-regularized OT barycenters, debiased Rényi-regularized transport (like Sinkhorn divergences [Fey+19]), convergence rates for $\lambda \rightarrow \{0, \infty\}$ and $\alpha \rightarrow \{0, 1\}$.

References

- [BT03] Amir Beck and Marc Teboulle. “Mirror descent and nonlinear projected subgradient methods for convex optimization”. In: *Operations Research Letters* 31.3 (2003), pp. 167–175. doi: 10.1016/S0167-6377(02)00231-6.
- [Cut13] Marco Cuturi. “Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances”. In: *Advances in Neural Information Processing Systems* 26 (2013).
- [EH14] Tim van Erven and Peter Harremoës. “Rényi Divergence and Kullback-Leibler Divergence”. In: *IEEE Transactions on Information Theory* 60.7 (2014), pp. 3797–3820. doi: 10.1109/TIT.2014.2320500.
- [Fey+19] Jean Feydy et al. “Interpolating between optimal transport and mmd using sinkhorn divergences”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 2681–2690.
- [Muz+17] Boris Muzellec et al. “Tsallis regularized optimal transport and ecological inference”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 2017.
- [YL22] Jun-Kai You and Yen-Huan Li. *Two Polyak-Type Step Sizes for Mirror Descent*. 2022. arXiv: 2210.01532 [math.OA].