Sinkhorn Divergences for Unbalanced Optimal Transport

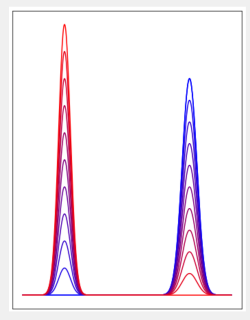
Providing a loss between arbitrary positive measures, fastly computable, with no bias.

Thibault Séjourné[1] Jean Feydy[1,2] François-Xavier Vialard[3] Alain Trouvé [2] Gabriel Peyré[1]

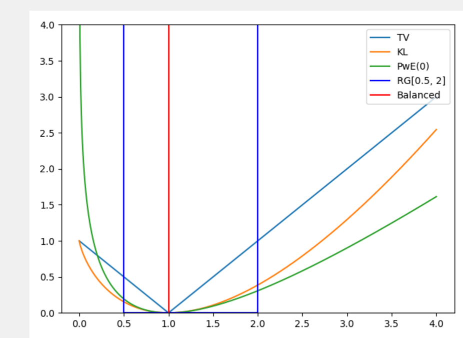
[1]DMA, École Normale Supérieure [2]CMLA, ENS Paris-Saclay [3]LIGM, UPEM

1 Csiszar divergences (Csi67) $\approx$ Vertical Geometry

- **Entropy** ϕ: positive, l.s.c., convex on $\mathbb{R}_+$ s.t. $\phi(1) = 0$
- **Recession constant:** $\phi'^{\infty} = \lim_{x \to \infty} \phi(x)/x$
- **Lebesgue decomposition:** $\forall(\alpha, \beta)$, $\alpha = \frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\beta + \alpha^\top$
- ϕ**-divergence:** $\mathrm{D}_\phi(\alpha, \beta) \overset{\text{def.}}{=} \int_{\mathcal{X}} \phi(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta})\mathrm{d}\beta + \phi'^{\infty} \int_{\mathcal{X}} \mathrm{d}\alpha^\top$



- **Balanced:** $\phi(x) = \iota_{\{1\}}(x)$ with $\mathrm{D}_\phi(\pi_1, \alpha) = \iota_{(=)}(\pi_1, \alpha)$.
- **KL:** $\phi(x) = x \log x - x + 1$
- **TV:** $\phi(x) = |x - 1|$
- **Range:** $\phi(x) = \iota_{[a,b]}(x)$ $(a \leq 1 \leq b)$, i.e $a\alpha \leq \pi_1 \leq b\alpha$.
- **Power entropy:** $\phi(x) = \frac{1}{p(p-1)}(x^p - p(x-1) - 1)$, $p \in \mathbb{R}$.

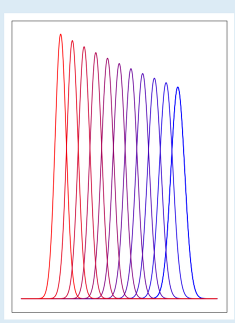


2 OT (Kan42) $\approx$ Vertical geometry

Consider a **Compact domain** $\mathcal{X}$. Take a cost C : $(x, y) \mapsto \mathrm{C}(x, y)$ continuous, symmetric, Lipschitz (e.g. $\frac{1}{p}\|x - y\|^p$). One defines

$$\mathrm{OT}_0(\alpha, \beta) \overset{\text{def.}}{=} \min_{\pi \geq 0} \left\{ \langle \pi, \mathrm{C} \rangle \ : \ \pi 1 = \alpha, \ \pi^\top 1 = \beta \right\},$$

where $\langle \pi, \mathrm{C} \rangle \overset{\text{def.}}{=} \int_{\mathcal{X}^2} \mathrm{C}(x, y)\mathrm{d}\pi(x, y)$. OT compares measures by accounting for the geometry. It metrizes the **convergence in law.**

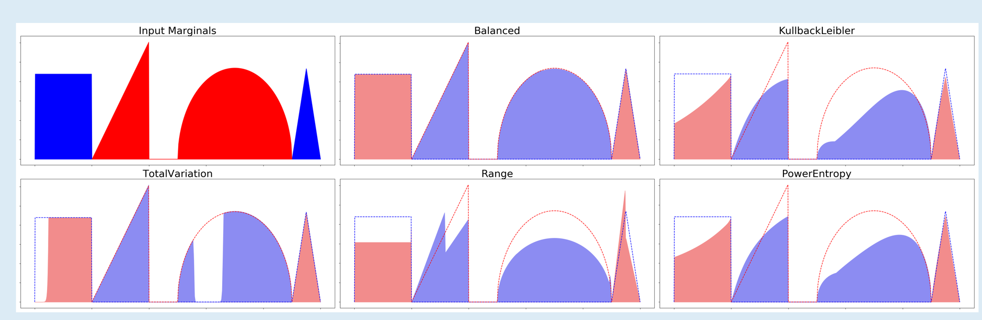


3 Unbalanced Optimal Transport (LMS15)

OT only compares normalized measures. Its generalization reads

$$\mathrm{OT}_0(\alpha, \beta) \overset{\text{def.}}{=} \inf_{\pi \geq 0} \langle \pi, \mathrm{C} \rangle + \rho \mathrm{D}_\phi(\pi_1, \alpha) + \rho \mathrm{D}_\phi(\pi_2, \beta).$$

It hybridizes vertical and horizontal geometries.



4 Entropic UOT

(U)OT has a complexity $O(n^3 \log n)$ and is non differentiable. Adding entropy improves both aspects (Sch31; KY94; Cut13; CPSV18).

$$\mathrm{OT}_\epsilon(\alpha, \beta) \overset{\text{def.}}{=} \inf_{\pi \geq 0} \langle \pi, \mathrm{C} \rangle + \rho \mathrm{D}_\phi(\pi_1, \alpha) + \rho \mathrm{D}_\phi(\pi_2, \beta) + \epsilon \mathbf{KL}(\pi, \alpha \otimes \beta)$$

Writing $\phi^*(x) = \sup_{y \geq 0} xy - \phi(y)$, the dual reads

$$\mathrm{OT}_\epsilon(\alpha, \beta) = \sup_{f, g \in \mathcal{C}(\mathcal{X})} \langle \alpha, -(\rho\phi)^*(-f) \rangle + \langle \beta, -(\rho\phi)^*(-g) \rangle - \epsilon \langle \alpha \otimes \beta, e^{\frac{f(x)+g(y)-\mathrm{C}(x,y)}{\epsilon}} - 1 \rangle.$$

5 Sinkhorn algorithm

Define the following operators

- **Softmin / LogSumExp**
  $\mathrm{Smin}_\alpha^\epsilon(f) \overset{\text{def.}}{=} -\epsilon \log \langle \alpha, e^{-f/\epsilon} \rangle$
- **Anisotropic Prox** (CR13)
  $\mathrm{aprox}(p) = \arg\min_{q \in \mathbb{R}} \epsilon e^{(p-q)/\epsilon} + \phi^*(q)$

The dual optimality condition defines the Sinkhorn algorithm

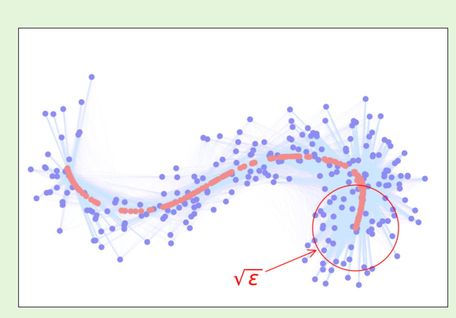
$$g_{t+1}(y) = -\,\mathrm{aprox}(\,-\mathrm{Smin}_\alpha^\epsilon\,(\mathrm{C}(.,y) - f_t)\,)$$
$$f_{t+1}(x) = -\,\mathrm{aprox}(\,-\mathrm{Smin}_\beta^\epsilon\,(\mathrm{C}(x,.) - g_{t+1})\,).$$

Ex: (Balanced) $\mathrm{aprox}(p) = p$, (ρKL) $\mathrm{aprox}(p) = \frac{\rho}{\rho+\epsilon}p$

6 Removing the entropic bias

When $\varepsilon > 0$, **fuzzy transport plans** induce shrinking artifacts (CR03):

Minimize $\mathrm{OT}_\varepsilon(\alpha, \beta)$ with respect to α $\longrightarrow$



$\Longrightarrow$ Use the **unbiased** Sinkhorn divergence (RTC17; GPC18; SZRM18):

$$\mathrm{S}_\varepsilon(\alpha, \beta) \ = \ \mathrm{OT}_\varepsilon(\alpha, \beta) - \tfrac{1}{2}\mathrm{OT}_\varepsilon(\alpha, \alpha) - \tfrac{1}{2}\mathrm{OT}_\varepsilon(\beta, \beta) + \frac{\varepsilon}{2}(m(\alpha) - m(\beta))^2,$$

$$\underbrace{\mathrm{OT}(\alpha, \beta)}_{\text{Wasserstein}} \xleftarrow{\varepsilon \to 0} \underbrace{\mathrm{S}_\varepsilon(\alpha, \beta)}_{\text{Easy to compute}} \xrightarrow{\varepsilon \to +\infty} \underbrace{\mathrm{MMD}_{-\mathrm{C}}(\alpha, \beta)}_{\text{Kernel MMD}}$$

7 Contributions 1 & 2

Theorem: If $e^{-\mathrm{C}(x,y)/\varepsilon}$ is a positive definite kernel, then for any strictly convex ϕ^*

$$\mathrm{S}_\varepsilon(\beta, \beta) \ = \ 0 \ \leqslant \ \mathrm{S}_\varepsilon(\alpha, \beta)$$
$$\mathrm{S}_\varepsilon(\alpha, \beta) \ = \ 0 \iff \alpha = \beta$$
$$\mathrm{S}_\varepsilon(\alpha_n, \beta) \ \to \ 0 \iff \alpha_n \rightharpoonup \beta$$
$$\mathrm{Loss}_\beta : \alpha \mapsto \mathrm{S}_\varepsilon(\alpha, \beta) \ \text{ is } \textbf{convex}.$$

Theorem: For all entropies mentioned above, the Sinkhorn algorithm converges linearly towards the optimal dual potentials (f, g) with a rate independent of the number of sample points.

8 Sample Complexity

Set (α_n, β_n) the sampled versions of (α, β) with n points. In $\mathbb{R}^d$ one has for **Balanced OT**:

Unregularized OT (Dud69; WB17):
$$\mathbb{E}_{\alpha \otimes \beta}\left[|\mathrm{OT}(\alpha, \beta) - \mathrm{OT}(\alpha_n, \beta_n)|\right] = O(n^{-1/d})$$

Compact supports (GCB+18):
$$\mathbb{E}_{\alpha \otimes \beta}\left[|\mathrm{OT}_\epsilon(\alpha, \beta) - \mathrm{OT}_\epsilon(\alpha_n, \beta_n)|\right] = O_{\varepsilon \to 0}(\varepsilon^{-\lfloor d/2 \rfloor} n^{-1/2})$$

Subgaussian measures (MW19):
$$\mathbb{E}_{\alpha \otimes \beta}\left[|\mathrm{OT}_\epsilon(\alpha, \beta) - \mathrm{OT}_\epsilon(\alpha_n, \beta_n)|\right] = O(\varepsilon^{1 - \lceil 5d/4 + 3 \rceil} n^{-1/2})$$

9 Contribution 3

Take positive measures (α, β), set $\bar{\alpha} = \alpha/m(\alpha)$ and $\bar{\beta} = \beta/m(\beta)$. Write

$$\alpha_n = \frac{m(\alpha)}{n}\sum_{i=1}^{n} \delta_{X_i} \qquad \beta_n = \frac{m(\beta)}{n}\sum_{i=1}^{n} \delta_{Y_i},$$

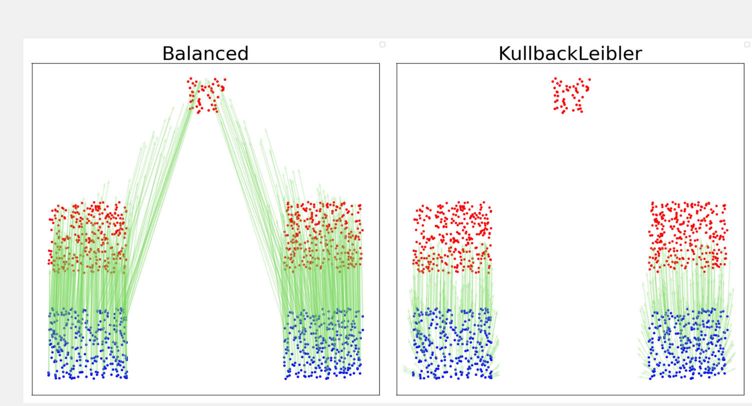
Theorem: Assume smooth C and ϕ^*, assume ϕ^* is strictly convex. Then

$$\mathbb{E}_{\bar{\alpha} \otimes \bar{\beta}}\left[|\mathrm{OT}_\epsilon(\alpha, \beta) - \mathrm{OT}_\epsilon(\alpha_n, \beta_n)|\right] = O_{\varepsilon \to 0}(\varepsilon^{-\lfloor d/2 \rfloor} n^{-1/2}) = O_{\varepsilon \to \infty}(n^{-1/2}).$$

Furthermore the rate is linear in $m(\alpha) + m(\beta)$.

10 Numerical Highlight

Unbalanced OT allows to **discard geometric outliers** !



Notation

Banach duality: $f \in \mathcal{C}(\mathcal{X})$, $\alpha \in \mathcal{M}_+(\mathcal{X})$

Dual Bracket: $\langle \alpha, f \rangle = \int_{\mathcal{X}} f \mathrm{d}\alpha = \mathbb{E}_\alpha[f]$

Discrete Encoding: $(\alpha_i)_i \in \mathbb{R}^N$, $(x_i)_i \in \mathbb{R}^{N \times D}$, $(f_i)_i \in \mathbb{R}^N$

Discrete Setting: $\alpha = \sum_i \alpha_i \delta_{x_i} \Rightarrow \langle \alpha, f \rangle = \sum_i \alpha_i f_i$

Check the repos at:

www.github.com/thibsej/unbalanced-ot-functionals
www.kernel-operations.io/geomloss

References

[CPSV18] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Scaling algorithms for unbalanced transport problems. to appear in Mathematics of Computation, 2018.

[CR03] H. Chui and A. Rangarajan. A new point matching algorithm for non-rigid registration. Computer Vision and Image Understanding, 89(2-3):114–141, 2003.

[CR13] P. L. Combettes and N. N. Reyes. Moreau's decomposition in banach spaces. Mathematical Programming, 139(1-2):103–114, 2013.

[Csi67] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. studia scientiarum Mathematicarum Hungarica, 2:229–318, 1967.

[Cut13] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Adv. in Neural Information Processing Systems, pages 2292–2300, 2013.

[Dud69] R. M. Dudley. The speed of mean Glivenko-Cantelli convergence. The Annals of Mathematical Statistics, 40(1):40–50, 1969.

[FSV+18] J. Feydy, T. Séjourné, F.-X. Vialard, S.-I. Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and mmd using Sinkhorn divergences. arXiv preprint arXiv:1810.08278, 2018.

[GCB+18] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of Sinkhorn divergences. arXiv preprint arXiv:1810.02733, 2018.

[GPC18] A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with Sinkhorn divergences. In International Conference on Artificial Intelligence and Statistics, pages 1608–1617, 2018.

[Kan42] L. Kantorovich. On the transfer of masses (in Russian). Doklady Akademii Nauk, 37(2):227–229, 1942.

[KY94] J. J. Kosowsky and A. L. Yuille. The invisible hand algorithm: Solving the assignment problem with statistical physics. Neural networks, 7(3):477–490, 1994.

[LMS15] M. Liero, A. Mielke, and G. Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. Inventiones mathematicae, pages 1–149, 2015.

[MW19] G. Mena and J. Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. arXiv preprint arXiv:1905.11882, 2019.

[RTC17] A. Ramdas, N. G. Trillos, and M. Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. Entropy, 19(2), 2017.

[Sch31] E. Schrödinger. Über die Umkehrung der Naturgesetze. Sitzungsberichte Preuss. Akad. Wiss. Berlin. Phys. Math., 144:144–153, 1931.

[SZRM18] T. Salimans, H. Zhang, A. Radford, and D. Metaxas. Improving gans using optimal transport. arXiv preprint arXiv:1803.05573, 2018.

[WB17] J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. arXiv preprint arXiv:1707.00087, 2017.

[Wil69] A. G. Wilson. The use of entropy maximizing models, in the theory of trip distribution, mode split and route split. Journal of Transport Economics and Policy, pages 108–126, 1969.