Energy-Guided Continuous Entropic Barycenter Estimation for General Costs

A. Kolesov[1], P. Mokrov[1], I. Udovichenko[1], M. Gazdieva[1], G. Pammer[2], A. Kratsios[3], E. Burnaev[1,4], A. Korotin[1,4]

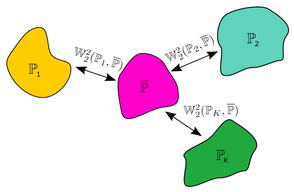
[1]Skolkovo Institute of Science and Technology (Moscow, Russia)
[2]Department of Mathematics, ETH Zürich (Zürich, Switzerland)
[3]Department of Mathematics, Vector Institute, McMaster University (Ontario, Canada)
[4]Artificial Intelligence Research Institute (Moscow, Russia)

Wasserstein-2 barycenter problem

OT barycenter $\overline{\mathbb{P}}$ is the average of distributions $\{\mathbb{P}_k\}_{k=1}^K$ w.r.t. a given transport cost function c.

Particular case:
The Wasserstein-2 barycenter with **Euclidean** quadratic cost $c(x,y) = \ell_2^2(x,y) \equiv \frac{1}{2}\|x-y\|_2^2$ is

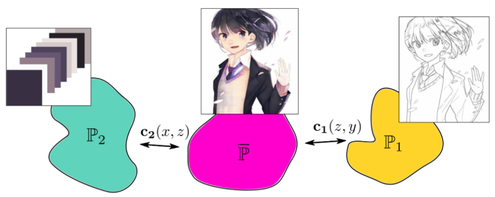
$$\overline{\mathbb{P}} = \arg\min_{\mathbb{Q}} \sum_{k=1}^K \lambda_k \mathbb{W}_2^2(\mathbb{P}_k, \mathbb{Q}) \text{ s.t. } \sum_{k=1}^K \lambda_k = 1, \lambda > 0.$$



The $\mathbb{W}_2^2$ barycenters are just intersections in data space that have no practical meaningfulness.

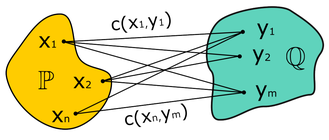
Common **Wasserstein-2 ($\mathbb{W}_2^2$) barycenter** solvers are **restricted** to using the quadratic Euclidean cost ℓ_2^2 only. They can not work with **general** costs.

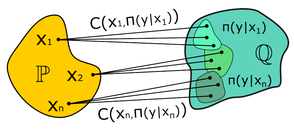
Question : How to build meaningful barycenters?



Background on (entropic) OT

	Classical OT	Weak OT
Transport cost:	$c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$	$C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$
Example:	$c(x,y) = \frac{1}{2}\|x-y\|_2^2$	$C(x, \pi(\cdot\|x)) = \mathbb{E}_{y\sim\pi(\cdot\|x)} c(x,y) - \epsilon H(\pi(\cdot\|x))$
Primal:	$\text{OT}_c(\mathbb{P},\mathbb{Q}) = \inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{(x,y)\sim\pi} c(x,y)$	$\text{EOT}_C(\mathbb{P},\mathbb{Q}) = \inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{x\sim\mathbb{P}} C(x,\pi(\cdot\|x))$
Dual:	$\text{OT}_c(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{C}(\mathcal{Y})} \mathbb{E}_{x\sim\mathbb{P}} f^c(x) + \mathbb{E}_{y\sim\mathbb{Q}} f(y)$	$\text{EOT}_C(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{C}(\mathcal{Y})} \mathbb{E}_{x\sim\mathbb{P}} f^C(x) + \mathbb{E}_{y\sim\mathbb{Q}} f(y)$
Conjugate :	$f^c(x) \stackrel{\text{def}}{=} \inf_{y\in\mathcal{Y}} \{c(x,y) - f(y)\}$	$f^C(x) \stackrel{\text{def}}{=} \inf_{\mu\in\mathcal{P}(\mathcal{Y})} \{C(x,\mu) - \mathbb{E}_{y\sim\mu} f(y)\}$





Weak EOT dual

For the entropic case, weak C-transform reads as:

$$f^C(x) \stackrel{\text{def}}{=} \min_{\mu\in\mathcal{P}(\mathcal{Y})} \left\{ \mathbb{E}_{y\sim\mu} c(x,y) - \epsilon H(\mu) - \mathbb{E}_{y\sim\mu} f(y) \right\}.$$

The inner minimization problem can be solved explicitly up to a normalization constant:

$$\frac{\mathrm{d}\mu_x^f(y)}{\mathrm{d}y} \stackrel{\text{def}}{=} \frac{1}{Z_c(f,x)} \exp\left(\frac{f(y) - c(x,y)}{\epsilon}\right).$$

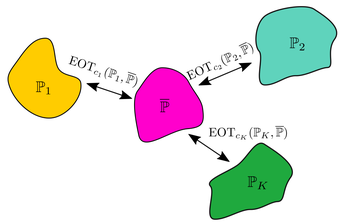
One can employ techniques from **energy-based modeling** to work with distributions of this type.

Weak EOT barycenter problem

Let $\mathbb{P}_k \in \mathcal{P}(\mathcal{X}_k)$ be given distributions **accessible by samples**; let $C_k : \mathcal{X}_k \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$ be appropriate cost functions, $k \in \{1,..,K\}$.

For positive weights λ_k s.t. $\sum_{k=1}^N \lambda_k = 1$ the **EOT barycenter problem** consists in finding a distribution $\overline{\mathbb{P}} = \mathbb{Q}^*$ that minimizes:

$$\mathcal{L}^* = \inf_{\mathbb{Q}\in\mathcal{P}(\mathcal{Y})} \sum_{k=1}^K \lambda_k \text{EOT}_{c_k}(\mathbb{P}_k, \mathbb{Q}) \qquad (1)$$



Dual problem for EOT barycenter

Our main theorem states that the primal EOT objective for barycenter can be attained via optimizing over the potentials.

Theorem 1 Problem (1) permits the following dual formulation:

$$\mathcal{L}^* = \sup_{\substack{f_1,\dots,f_K \in \mathcal{C}(\mathcal{Y}):\\ \sum_{k=1}^K \lambda_k f_k \equiv 0 \text{ (congruence)}}} \mathcal{L}(f_1,\dots,f_K), \qquad (2)$$

where

$$\mathcal{L}(f_1,\dots,f_K) \stackrel{\text{def}}{=} \sum_{k=1}^K \lambda_k \left\{ \mathbb{E}_{x_k\sim\mathbb{P}_k} f_k^{C_k}(x_k) \right\} \qquad \left[= -\epsilon \sum_{k=1}^K \lambda_k \left\{ \mathbb{E}_{x_k\sim\mathbb{P}_k} \log Z_{c_k}(f_k, x_k) \right\} \right].$$

Practical optimization procedure

We parametrize the potentials f_k with neural networks $f_{\theta,k}$, and the objective (2) becomes:

$$L(\theta) \stackrel{\text{def}}{=} -\epsilon \sum_{k=1}^K \lambda_k \left\{ \mathbb{E}_{x_k\sim\mathbb{P}_k} \log Z_{c_k}(f_{\theta,k}, x_k) \right\}.$$

The direct computation of the normalizing constant Z_{c_k} may be infeasible. Still, the gradient of L with respect to θ can be derived:

Theorem 2 The gradient of L satisfies

$$\frac{\partial}{\partial\theta} L(\theta) = -\sum_{k=1}^K \lambda_k \mathbb{E}_{x_k\sim\mathbb{P}_k} \left\{ \mathbb{E}_{y\sim\pi^{f_{\theta,k}}(\cdot|x_k)} \left[\frac{\partial}{\partial\theta} f_{\theta,k}(y) \right] \right\}.$$

Sampling from $\pi^{f_{\theta,k}}(\cdot \mid x_k)$ can be done via MCMC (e.g., ULA). The theorem leads to the following practical algorithm.

Algorithm 1: EOT barycenters via Energy-Based Modelling

Input: Distributions $\mathbb{P}_k$, $k \in \overline{K}$ accessible by samples; cost functions $c_k(x_k, y) : \mathcal{X}_k \times \mathcal{Y} \to \mathbb{R}$; the regularization coeff. $\epsilon > 0$; barycenter averaging coeff. $\lambda_k > 0 : \sum_{k=1}^K \lambda_k = 1$; MCMC procedure `MCMC_proc`; batch size $S > 0$; NNs $f_{\theta,k} : \mathcal{Y} \to \mathbb{R}$, s.t. $\sum_{k=1}^K \lambda_k f_{\theta,k} \equiv 0$.
Output: Trained NNs $f_{\theta^*,k}$ recovering the conditional OT plans between $\mathbb{P}_k$ and barycenter $\mathbb{Q}^*$.

for $iter = 1, 2, \dots$ **do**
  for $k = 1, 2, \dots, K$ **do**
    Sample batch $\{x_k^s\}_{s=1}^S \sim \mathbb{P}_k$;
    Draw $Y_k = \{y_k^s\}_{s=1}^S$ with MCMC: $y_k^s = \texttt{MCMC\_proc}\left(\frac{f_{\theta,k}(\cdot) - c(x_k^s,\cdot)}{\epsilon}\right)$;
    $\widehat{L}_k \leftarrow -\lambda_k \frac{1}{S}\left[\sum_{s=1}^S f_{\theta,k}(y_k^s)\right]$;
  $\widehat{L} \leftarrow \sum_{k=1}^K \widehat{L}_k$; Update θ by using $\frac{\partial \widehat{L}}{\partial\theta}$;

Proved universal approximation

Theorem 3 (Barycenter quality bounds decomposition)

Let $\{f_k\}_{k=1}^K, f_k \in \mathcal{C}(\mathcal{Y})$ be congruent potentials. The following holds:

$$\epsilon \sum_{k=1}^K \lambda_k \text{KL}\left(\mathbb{Q}^* \| \mathbb{Q}^{f_k}\right) \leq \overbrace{2\sum_{k=1}^K \lambda_k \mathbb{E}\text{Rep}_{X_k}(\mathcal{F}_k^{C_k}, \mathbb{P}_k)}^{\text{Estimation error (upper bound)}} + \overbrace{\left[\mathcal{L}^* - \max_{(f_1,\dots,f_K)\in\overline{\mathcal{F}}} \mathcal{L}(f_1,\dots,f_K)\right]}^{\text{Approximation error}},$$

where $\mathcal{F}_k^{C_k} \stackrel{\text{def}}{=} \{f_k^{C_k} \mid (f_1,\dots,f_K) \in \overline{\mathcal{F}}\}$, and the expectations are taken w.r.t. the random realization of the datasets $X_1{\sim}\mathbb{P}_1, \dots, X_K{\sim}\mathbb{P}_K$. Here $\text{Rep}_{X_k}(\mathcal{F}_k^{C_k}, \mathbb{P}_k)$ is the standard notion of the **representativeness** of the sample X_k w.r.t. functional class $\mathcal{F}_k^{C_k}$ and distribution $\mathbb{P}_k$.

Theorem 4 (Bound on $\mathbb{E}$Rep w.r.t. C_k-transform classes)

(a) Let $\mathcal{F}_k \subset \mathcal{C}(\mathcal{Y})$. Assume that $c_k(x,y)$ is Lipschitz in x with the same Lipschitz constant for all $y \in \mathcal{Y}$. Then

$$\mathbb{E}\text{Rep}_{X_k}(\mathcal{F}_k^{C_k}, \mathbb{P}_k) \leq O\left(N_k^{-1/(D_k+1)}\right).$$

(b) Let $c_k(x_k,y) = \frac{1}{2}\|u_k(x_k) - v(y)\|^2$, $\mathcal{F}_k$ be a **bounded** (w.r.t. the supremum norm) subset of $\mathcal{C}(\mathcal{Y})$, $u_k : \mathcal{X}_k \to \mathbb{R}^{D''}$ and $v : \mathcal{Y} \to \mathbb{R}^{D''}$ be continuous functions. Then

$$\mathbb{E}\text{Rep}_{X_k}(\mathcal{F}_k^{C_k}, \mathbb{P}_k) \leq O\left(N_k^{-1/2}\right).$$

Theorem 5 (Vanishing approximation error)

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an activation function. Assume that it is non-affine and there is an $\widetilde{x} \in \mathbb{R}$ at which σ is differentiable and $\sigma'(\widetilde{x}) \neq 0$. Then for every $\delta > 0$ there exist K multi-layer perceptrons $g_k : \mathbb{R}^D \to \mathbb{R}$ with activations σ for which the congruent functions $f_k = g_k - \sum_{k=1}^K \lambda_k g_k$ satisfy

$$\sum_{k=1}^K \lambda_k \text{KL}\left(\pi_k^* \| \pi^{f_k}\right) = (\mathcal{L}^* - \mathcal{L}(f_1,\dots,f_K))/\epsilon < \delta/\epsilon.$$

Furthermore, each g_k has width at-most $D+4$.

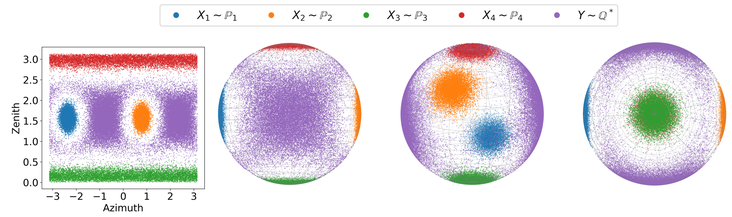
Sphere experiment

Our method can work 1) with arbitrary OT costs, 2) in a manifold-constrained setup.

Input distributions are 4 von Mises distributions.

Manifold is the unit 2D-sphere S^2 in $\mathbb{R}^3$.

Transport costs: $\forall k \in \overline{1,4} : c_k(x_k,y) = \frac{1}{2}\left(\arccos\langle x_k, y\rangle\right)^2$



(a) The unfolded sphere. (b) The sphere viewed from different viewpoints.

Ave, CelebA experiment

Input distributions are transformed CelebA faces $\mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_3$.

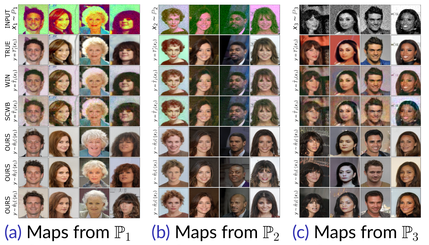
True unregularized barycenter is the CelebA dataset.

Manifold is represented by Style-GAN G that is trained on CelebA dataset.

Transport costs: $\forall k \in \overline{1,3} : c_k(x,z) = \frac{1}{2}\|x_k - G(z)\|_2^2$.

Solver	*FID*↓ of plans to barycenter		
	$k=1$	$k=2$	$k=3$
SCWB	56.7	53.2	58.8
WIN	49.3	46.9	61.5
Ours	**8.4**	**8.7**	**10.2**

FID scores of images mapped from inputs $\mathbb{P}_k$ to the barycenter.



(a) Maps from $\mathbb{P}_1$ to $\mathbb{Q}$. (b) Maps from $\mathbb{P}_2$ to $\mathbb{Q}$. (c) Maps from $\mathbb{P}_3$ to $\mathbb{Q}$.

Links

arXiv