Estimating Barycenters of Distributions with Neural Optimal Transport

Alexander Kolesov[1], Petr Mokrov[1], Igor Udovichenko[1], Milena Gazdieva[1], Gudmund Pammer[2], Evgeny Burnaev[1,3], Alexander Korotin[1,3]

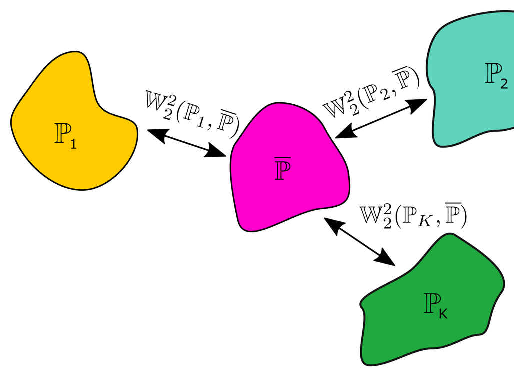
[1]Skolkovo Institute of Science and Technology (Moscow, Russia)
[2]Department of Mathematics, ETH Zürich (Zürich, Switzerland)
[3]Artificial Intelligence Research Institute (Moscow, Russia)

Wasserstein-2 barycenter problem

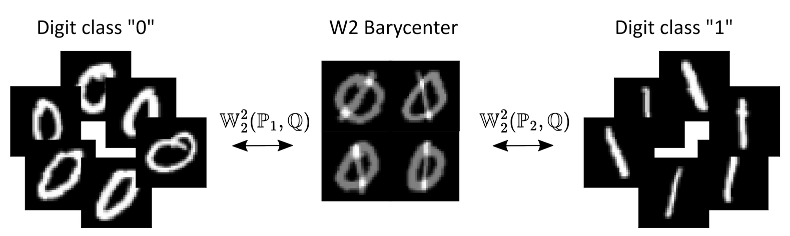
OT barycenter $\overline{\mathbb{P}}$ is the average of distributions $\{\mathbb{P}_k\}_{k=1}^{K}$ w.r.t. a given transport cost function c.

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

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

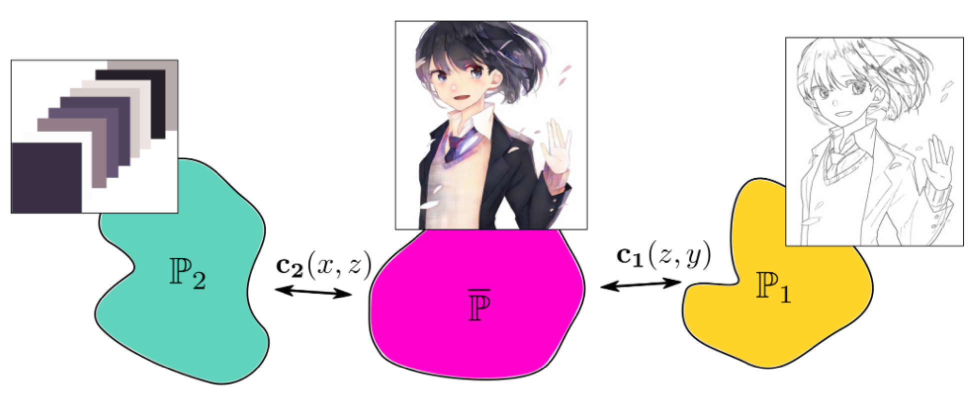


The $\mathbb{W}_2^2$ barycenters are just intersections in data space that have no practical meaningfulness.



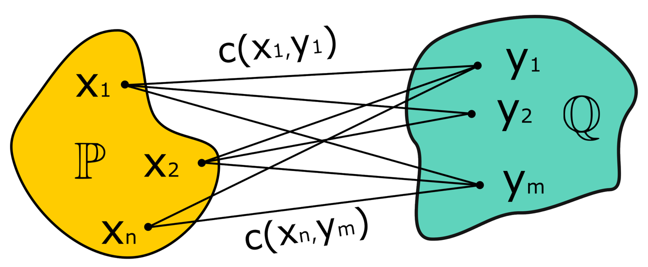
Common **Wasserstein-2 ($\mathbb{W}_2^2$) barycenter** solvers are **restricted** to using the quadratic Euclidean cost ℓ_2^2 only. They can not work with **general** costs.

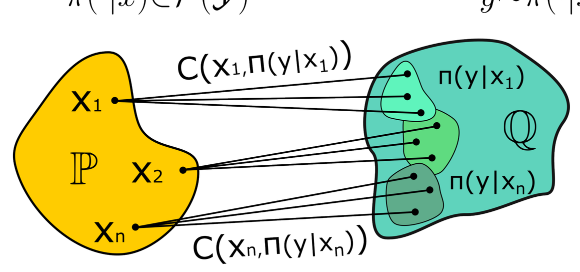
Question : How to build meaningful barycenters?



Background on 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:	$c(x,y) = \frac{1}{2}\|x-y\|_2^2$	$C(x,\pi(\cdot\|x)) = \mathbb{E}_{y\sim\pi(\cdot\|x)}\{c(x,y) + \gamma\mathcal{R}(\pi(\cdot\|x))\}$
Primal:	$\text{OT}_c(\mathbb{P},\mathbb{Q}) = \inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{(x,y)\sim\pi} c(x,y)$	$\text{OT}_C(\mathbb{P},\mathbb{Q}) = \inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{x\sim\mathbb{P}} C(x,\pi(\cdot\|x))$
Dual:	$\text{OT}_c(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{C}(\mathcal{Y})} \mathbb{E}_{x\sim\mathbb{P}} f^c(x) + \mathbb{E}_{y\sim\mathbb{Q}} f(y)$	$\text{OT}_C(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{C}(\mathcal{Y})} \mathbb{E}_{x\sim\mathbb{P}} f^C(x) + \mathbb{E}_{y\sim\mathbb{Q}} f(y)$
Conjugate :	$f^c(x) \stackrel{\text{def}}{=} \inf_{y\in\mathcal{Y}}\{c(x,y) - f(y)\}$	$f^C(x) \stackrel{\text{def}}{=} \inf_{\pi(\cdot\|x)\in\mathcal{P}(\mathcal{Y})}\{C(x,\pi(\cdot\|x)) - \mathbb{E}_{y\sim\pi(\cdot\|x)} f(y)\}$



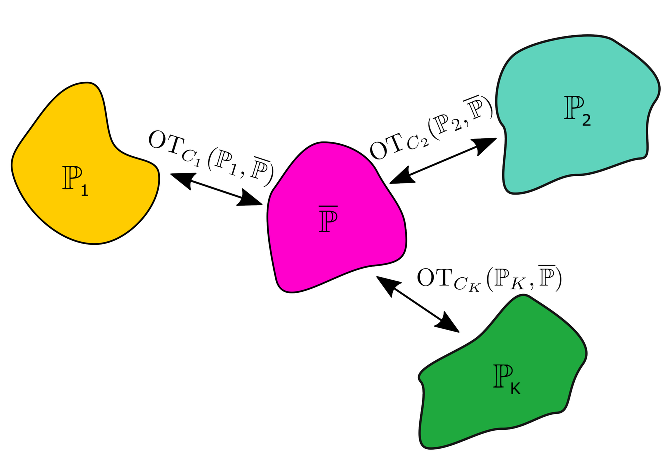


Weak OT barycenter

Let $\mathbb{P}_k \in \mathcal{P}(\mathcal{X}_k)$ be given distributions; 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 **OT barycenter problem** consists in finding a distribution $\overline{\mathbb{P}} = \mathbb{Q}^*$ that minimizes:

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



Our methodology

Step 1. Dual formulation of weak OT:

$$\text{OT}_C = \sup_{f\in\mathcal{C}(\mathcal{Y})} \inf_{\pi\in\Pi(\mathbb{P})} \big(\underbrace{\mathbb{E}_{x\sim\mathbb{P}}\{C(x,\pi(\cdot|x)) - \mathbb{E}_{y\sim\pi(\cdot|x)} f(y)\} + \mathbb{E}_{y\sim\mathbb{Q}} f(y)}_{\mathcal{H}(f,\pi,\mathbb{Q})}\big). \quad (1)$$

Step 2. Extending (1) to the barycenter objective: **min-max-min** problem:

$$\mathcal{L}^* = \inf_{\mathbb{Q}\in\mathcal{P}(\mathcal{Y})} \sum_{k=1}^{K} \sup_{f_k\in\mathcal{C}(\mathcal{Y})} \inf_{\pi_k\in\Pi(\mathbb{P}_k)} \lambda_k \mathcal{H}_k(f_k, \pi_k, \mathbb{Q}). \quad (2)$$

Step 3. Optimality: equating the **first variation** of (2) w.r.t the barycenter $\overline{\mathbb{P}}$ to 0:

$$\nabla_{\mathbb{Q}} \sum_{k=1}^{K} \lambda_k \mathbb{E}_{y\sim\mathbb{Q}} f_k^*(y)|_{\mathbb{Q}=\overline{\mathbb{Q}}} = 0.$$

Step 4. Obtaining **congruence condition**: $\nabla_{\mathbb{Q}}\mathbb{E}_{y\sim\mathbb{Q}} f(y) = f \Rightarrow \boxed{\sum_{k=1}^{K} \lambda_k f_k^* \equiv 0}$.

Optimization problem

Final optimization objective (combination of (2) with the congruence condition):

$$\mathcal{L}^* = \sup_{\sum_k \lambda_k f_k = 0} \inf_{\pi_k\in\Pi(\mathbb{P}_k)} \sum_{k=1}^{K} \lambda_k \mathbb{E}_{x_k\sim\mathbb{P}_k}\{C_k(x_k, \pi_k(\cdot|x_k)) - \mathbb{E}_{y\sim\pi_k(\cdot|x_k)} f_k(y)\}$$

Statement : Solutions π_k^* approximate the OT plans between $\mathbb{P}_k$ and barycenter $\overline{\mathbb{P}}$.

Transport plan parameterization with (stochastic) maps.

Basic idea: $d\pi_k(x,y) = d\mathbb{P}_k(x) d\pi_k(y|x)$;

$$\pi_k(\cdot|x) = T_k(x,\cdot)_{\#}\mathbb{S}.$$

- $\mathcal{S} \subset \mathbb{R}^{D_s}$ is an auxiliary space;
- $\mathbb{S} \in \mathcal{P}(\mathcal{S})$ is a distribution (e.g., Gaussian);
- T_k is a map $T_k: \mathcal{X}_k \times \mathcal{S} \to \mathcal{Y}$.

Particular case #1: **Gaussian model.**

$$\pi_k(\cdot|x) = \mathcal{N}(\cdot|\mu(x), \text{diag}(\sigma^2(x)))$$

Particular case #2: **Deterministic map.**

$$\pi_k(\cdot|x) = \delta_{T_k(x)}(\cdot)$$

Considered weak costs.

In what follows, $\nu \in \mathcal{P}(\mathcal{Y})$.

Classical

$$C(x,\nu) = \mathbb{E}_{y\sim\nu} c(x,y)$$

ϵ-KL

$$C(x,\nu) = \mathbb{E}_{y\sim\nu} c(x,y) + \epsilon\text{KL}(\nu\|\nu_0),$$

ν_0 is a given prior, e.g., Gaussian.

γ-Energy

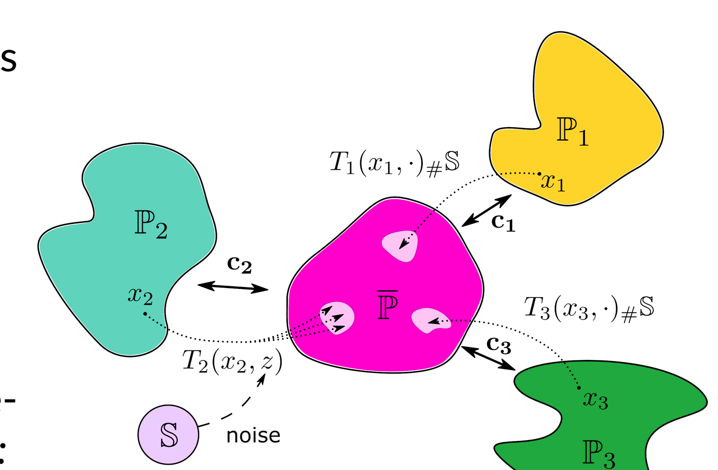
$$C(x,\nu) = \mathbb{E}_{y\sim\nu} c(x,y) + \gamma\mathcal{E}_\ell^2(\nu,\nu_0),$$

$\mathcal{E}_\ell^2$ is the (square) of energy distance.

$$\mathcal{L}^* = \sup_{\sum_k \lambda_k f_k = 0} \inf_{T_{1:K}} \sum_{k=1}^{K} \lambda_k \{\mathbb{E}_{x_k\sim\mathbb{P}_k} C_k(x_k, T_k(x_k,\cdot)\sharp\mathbb{S}) - \mathbb{E}_{x_k\sim\mathbb{P}_k}(\mathbb{E}_{s\sim\mathbb{S}} f_k(T_k(x_k,s)))\}$$

Method

We parameterize **conditional OT plans** $T_{1:K}$ as well as **potentials** $f_{1:K}$ with neural nets.



OT map parameterization:

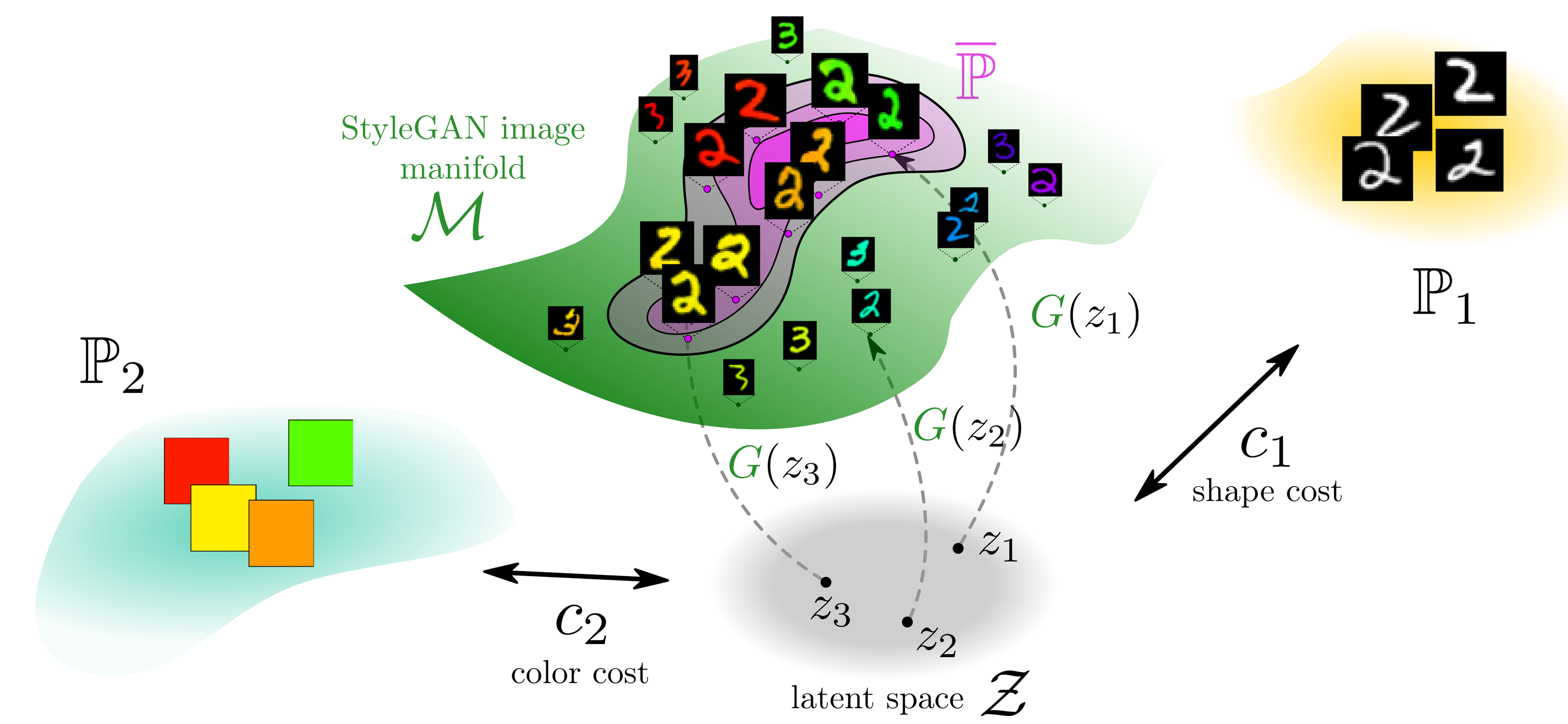
$$T_{1:K}: \forall k \quad T_{k,\phi}: \mathbb{R}^{D_k} \times \mathbb{R}^{D_s} \to \mathbb{R}^{D}.$$

OT Potential parameterization:
We introduce $g_k: \forall k \quad g_{k,\theta}: \mathbb{R}^D \to \mathbb{R}$ and represent potential $f_{k,\theta}$ through *congruence condition:*

$$f_{k,\theta} = g_{k,\theta} - \sum_{k'=1}^{K} \lambda_{k'} g_{k',\theta}.$$

Under appropriate stochastic maps parameterization, the **weak costs** (classical, ϵ-KL, γ-Energy) **permit sample estimates.**

Shape-Color Experiment



We consider ϵ-KL barycenter problem.

Shape distribution:
The distribution of gray-scale images of MNIST digits '2' on space $[0,1]^{32\times 32}$.

Color distribution:
The distribution of green, yellow and red HSV vectors on space $[0,1]^3$.

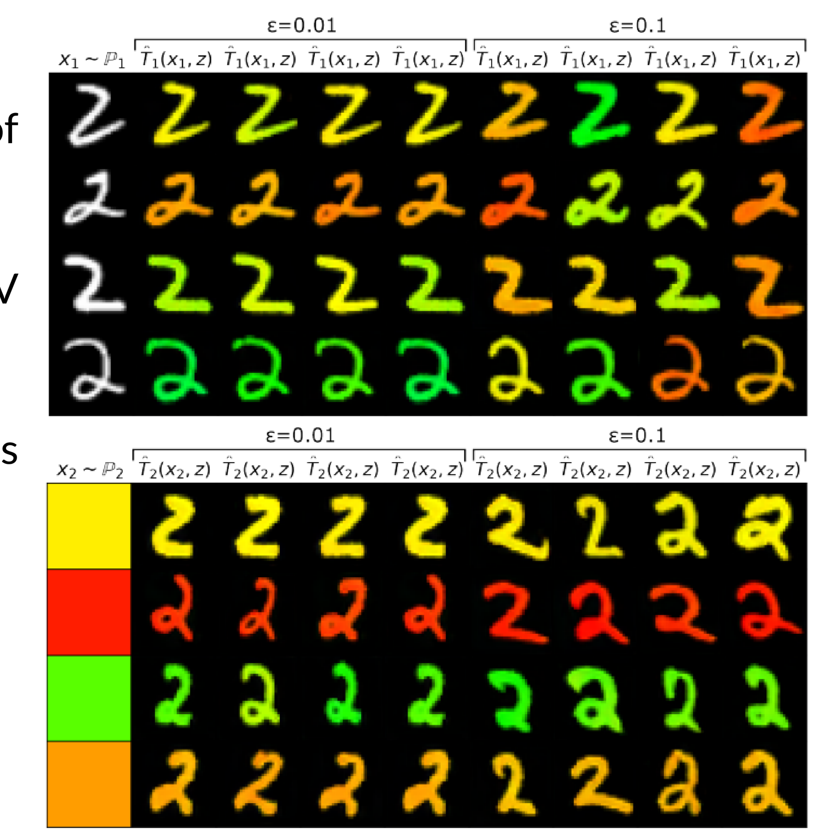
Manifold
It is represented by Style-GAN G that is trained on colored digits '**2**', '**3**' (all colors).

Transport costs:
Shape cost: $c_1(x_2,z) \stackrel{\text{def}}{=} \frac{1}{2}\|x_1 - H_g(G(z))\|_2^2$

Color cost: $c_2(x_2,z) \stackrel{\text{def}}{=} \frac{1}{2}\|x_2 - H_c(G(z))\|_2^2$

H_g(decolorization) : $\mathbb{R}^{3\times 32\times 32} \to \mathbb{R}^{32\times 32}$

H_c(defines HSV vector) : $\mathbb{R}^{3\times 32\times 32} \to \mathbb{R}^3$



Ave, CelebA Experiment

Barycenter distributions are transformed CelebA faces $\mathbb{P}_0, \mathbb{P}_1, \mathbb{P}_2$

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

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

FID:

Space	Solver	FID↓ $k=1$	$k=2$	$k=3$
Data space	SCWB	56.7	53.2	58.8
	WIN	49.3	46.9	61.5
	OURS	**39.0**	**38.6**	**39.8**
Mani-fold	EgBary	**8.4**	**8.7**	**10.2**
	OURS	30.7	31.0	31.7
	OURS (ϵ)	34.5	34.9	35.7
	OURS (γ)	38.3	37.8	37.6

Conclusion

We developed the pioneering **bi-level** (max-min) OT barycenter algorithm.

Our method is adopted for **general** cost functions.

Our method generates **meaningful** barycenters.

More details and code are represented in GitHub page.

GitHub

https://github.com/justkolesov/NOTBarycenters