Robust Barycenter Estimation using Semi-unbalanced Neural Optimal Transport

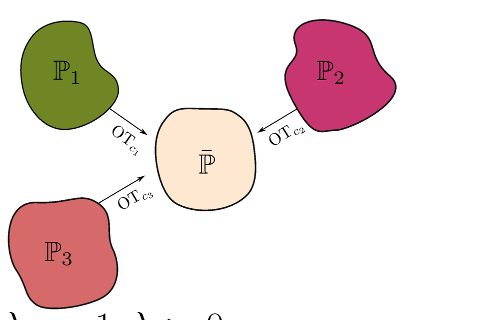
Milena Gazdieva[1,3], Jaemoo Choi[2], Alexander Kolesov[1,3], Jaewoong Choi[4], Petr Mokrov[1], Alexander Korotin[1,3]

[1]Skolkovo Institute of Science and Technology (Moscow, Russia)
[2]Georgia Institute of Technology (Atlanta, GA, USA)
[3]Artificial Intelligence Research Institute (Moscow, Russia)
[4]Sungkyunkwan University (Seoul, Korea)

OT barycenter problem

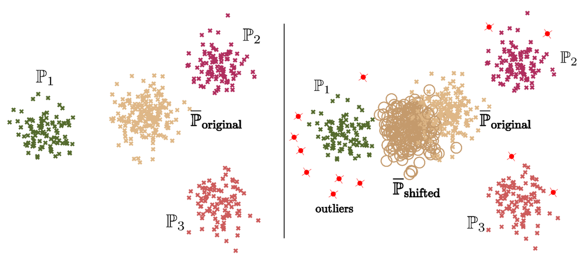
OT barycenter $\overline{\mathbb{P}}$ is the average of distributions $\{\mathbb{P}_k\}_{k=1}^K$ w.r.t. given transport cost functions c_k.

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



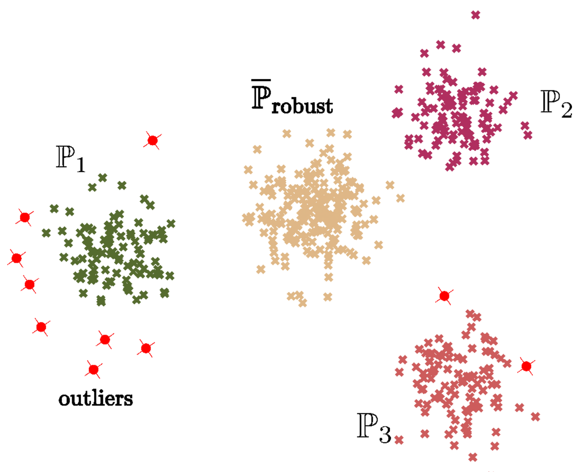
$$\overline{\mathbb{P}} = \arg\min_{\mathbb{Q}} \sum_{k=1}^K \lambda_k \mathbb{W}_2^2(\mathbb{P}_k,\mathbb{Q}) \text{ s.t. } \sum_{k=1}^K \lambda_k = 1, \lambda > 0.$$

Classic OT (e.g., $\mathbb{W}_2^2$) barycenters are sensitive to class imbalance and outliers in the input distributions.



Existing **OT barycenter** solvers lead to **biased** results in the case of outliers or class imbalance. They are **restricted** to dealing with *clean* datasets.

Question :
How to build **robust** barycenters?



Background on (unbalanced) OT

Classical OT

Transport cost: $c: \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$

Example: $c(x,y) = \frac{1}{2}\|x-y\|_2^2$

Conjugate : $f^c(x) \overset{\text{def}}{=} \inf_{y\in\mathcal{Y}} \{c(x,y) - f(y)\}$

Primal: $\text{OT}_c(\mathbb{P},\mathbb{Q}) = \inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{(x,y)\sim\pi} c(x,y)$

Dual: $\text{OT}_c(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{C}(\mathcal{Y})} \mathbb{E}_{x\sim\mathbb{P}} f^c(x) + \mathbb{E}_{y\sim\mathbb{Q}} f(y)$

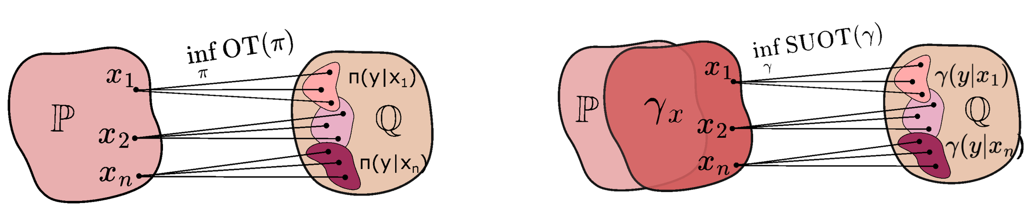
Semi-unbalanced OT

ψ-divergence between μ_1 and μ_2:

$$\text{D}_\psi(\mu_1\|\mu_2) \overset{def}{=} \int_{\mathcal{X}} \psi\left(\frac{\mu_1(x)}{\mu_2(x)}\right) d\mu_2(x).$$

Primal: $\text{SUOT}_{c,\psi}(\mathbb{P},\mathbb{Q}) = \inf_{\gamma\in\Pi(\mathbb{Q})} \mathbb{E}_{(x,y)\sim\gamma} c(x,y) + \text{D}_\psi(\gamma_x\|\mathbb{P})$

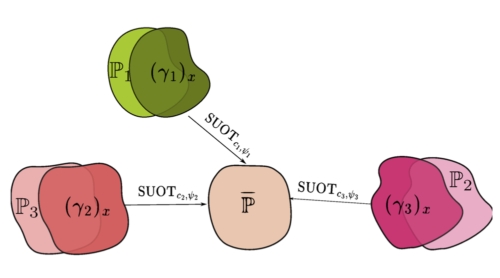
Dual: $\text{SUOT}_{c,\psi}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{C}(\mathcal{Y})} \mathbb{E}_{x\sim\mathbb{P}} -\overline{\psi}(-f^c)(x) + \mathbb{E}_{y\sim\mathbb{Q}} f(y)$



Semi-unbalanced OT barycenter

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

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



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

Our methodology

Step 1. Dual formulation of semi-unbalanced OT:

$$\text{SUOT}_{c,\psi} = \sup_{f\in\mathcal{C}(\mathcal{Y})} \inf_{\gamma(\cdot|x)\in\mathcal{P}(\mathcal{Y})} \Big(\mathcal{L}(f,\gamma,\mathbb{Q})\Big). \quad (1)$$

where $\mathcal{L}(f,\gamma,\mathbb{Q}) = \mathbb{E}_{x\sim\mathbb{P}} -\overline{\psi}\big(-\mathbb{E}_{x\sim\gamma(\cdot|x)}(c(x,y)-f(y)\big) + \mathbb{E}_{y\sim\mathbb{Q}} f(y)$

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

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

Step 3. Obtaining m-**congruence condition**:

$$\sum_{k=1}^K \lambda_k f_k^* \equiv m \text{ for some } m\in\mathbb{R}.$$

Optimization problem

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

$$\sup_{\substack{m\in\mathbb{R}, \\ \sum_{k=1}^K \lambda_k f_k \equiv m}} \inf_{\gamma(\cdot|x_k)\in\mathcal{P}(\mathcal{Y})} \sum_{k=1}^K \lambda_k\{-\mathbb{E}_{x_k\sim\mathbb{P}_k}\overline{\psi}_k(\mathbb{E}_{x_k\sim\gamma_k(\cdot|x_k)}(f_k(y)-c_k(x_k,y)) + m\}$$

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

Transport plan parameterization with (stochastic) maps.

Basic idea:

$d\gamma_k(x,y) = d\gamma_k(x)d\gamma_k(y|x)$;

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

- $\mathcal{S}\subset\mathbb{R}^{D_S}$ 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: **Deterministic map.**

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

Considered ψ-divergences.

Here $\overline{\psi}(\cdot)$ is a convex conjugate.

1. **Kullback-Leibler**
$\overline{\psi}_{\text{KL}}(t) = \exp(t) - 1$.
2. **Softplus**
$\overline{\psi}_{\text{Softplus}}(t) = \text{Softplus}(t)$.
3. **Identity (classic OT)**
$\overline{\psi}_{\text{Id}}(t) = t$.

Practice: **scaled** divergences τD_ψ; τ-**unbalancedness** parameter.

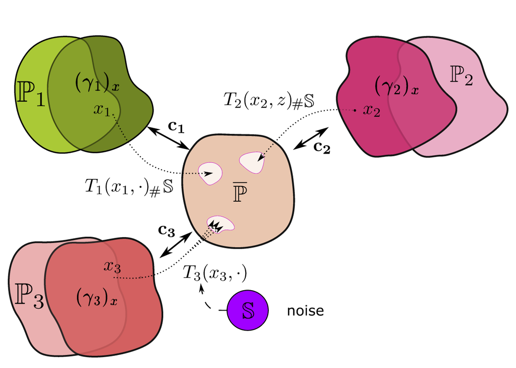
Empirical objective:

$$\sup_{\substack{m\in\mathbb{R} \\ \sum_{k=1}^K \lambda_k f_k \equiv m}} \inf_{T_{1:K}} \sum_{k=1}^K \lambda_k\{-\mathbb{E}_{x_k\sim\mathbb{P}_k}\overline{\psi}_k\Big(\mathbb{E}_{s\sim\mathbb{S}}\big(f_k(T_k(x_k,s)) - c_k(x_k,T_k(x_k,s))\big)\Big) + m\}.$$

Note: For $\psi_k = Id$, our solver reduces to classic OT barycenter solver (NOTB).

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:

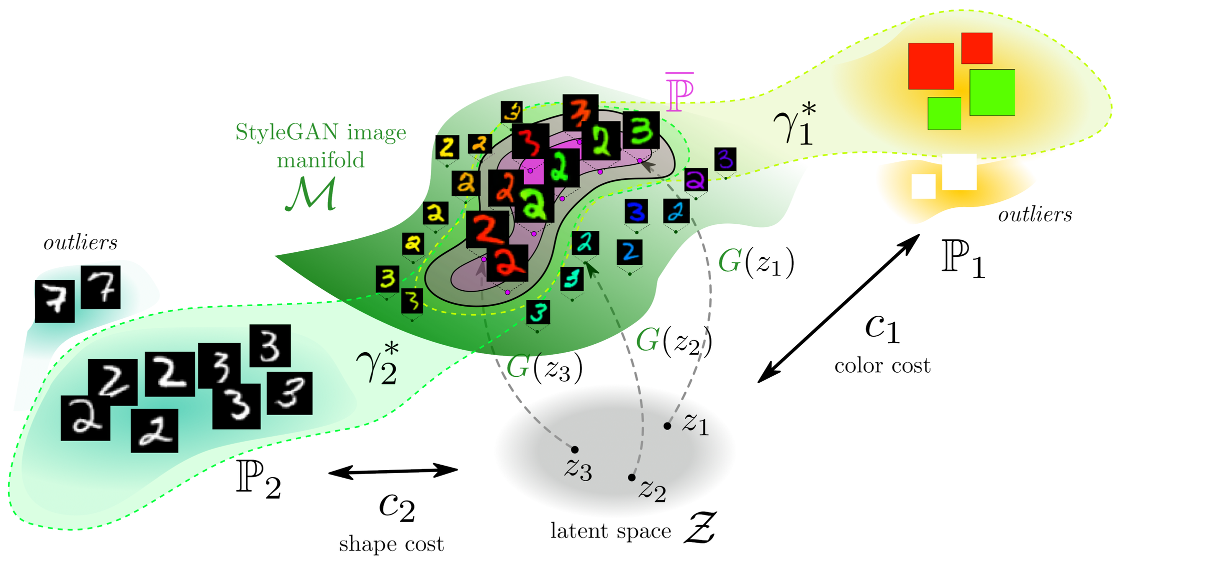
We introduce $g_k: \forall k \quad g_{k,\theta}: \mathbb{R}^D\to\mathbb{R}$ and represent potential $f_{k,\theta}$ through m-*congruence condition:*

$$f_{k,\theta} = g_{k,\theta} - \sum_{n\neq k} \frac{\lambda_n}{\lambda_k(K-1)} g_{n,\theta} + \frac{m}{K\lambda_k}.$$

Inference: input points should be sampled from $(\widehat{\gamma}_k)_x$. We approximate $\frac{d(\widehat{\gamma}_k)_x(x)}{d\mathbb{P}_k(x)} \approx \nabla\overline{\psi}_k(-(\widehat{f}_k)^c(x_k))$ where $\widehat{f}_k$ is the learned potential and apply *rejection sampling.*

Links

Shape-Color Experiment

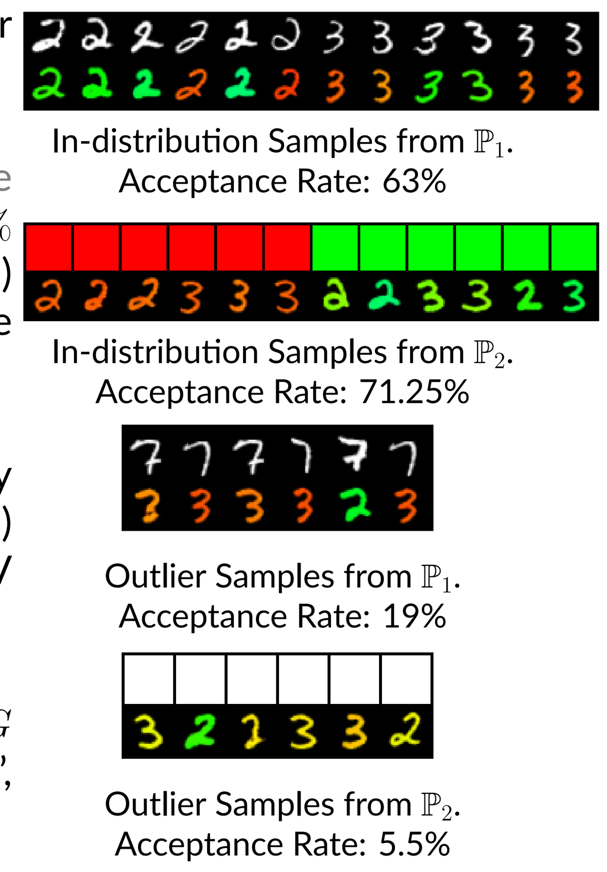


We consider SUOT barycenter problem with KL-divergence.

Shape distribution:
The distribution of gray-scale images of MNIST digits '2' (49% of training dataset), '3' (50%) and '7'(1% - *outliers*) on space $[0,1]^{32\times32}$.

Color distribution:
The distribution of red (probability mass $p_0 = 0.495$), green ($p_1 = 0.495$) and white($p_2 = 0.01$ - *outliers*) 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).



In-distribution Samples from $\mathbb{P}_1$. Acceptance Rate: 63%

In-distribution Samples from $\mathbb{P}_2$. Acceptance Rate: 71.25%

Outlier Samples from $\mathbb{P}_1$. Acceptance Rate: 19%

Outlier Samples from $\mathbb{P}_2$. Acceptance Rate: 5.5%

Transport costs:
Shape cost:

$$c_1(x_2,z) \overset{\text{def}}{=} \frac{1}{2}\|x_1 - H_g(G(z))\|_2^2$$

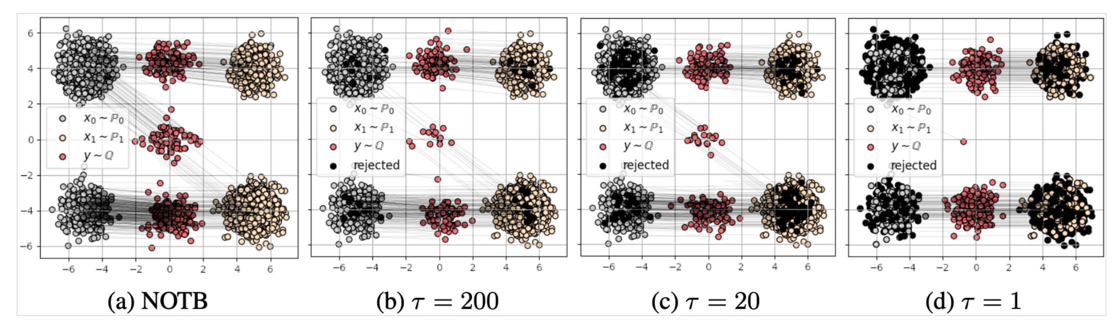
Color cost:

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

$H_g(\text{decolorization}): \mathbb{R}^{3\times32\times32}\to\mathbb{R}^{32\times32}$

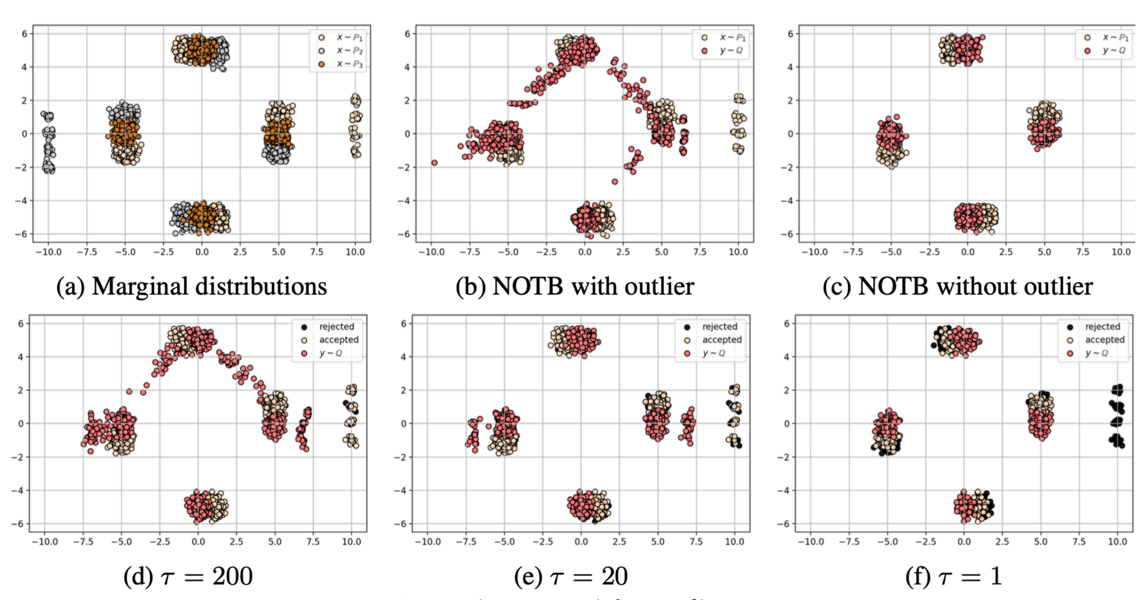
$H_c(\text{defines HSV vector}): \mathbb{R}^{3\times32\times32}\to\mathbb{R}^3$

Outlier & Class Imbalance Experiments



(a) NOTB (b) $\tau = 200$ (c) $\tau = 20$ (d) $\tau = 1$

Experiment with class imbalance



(a) Marginal distributions (b) NOTB with outlier (c) NOTB without outlier
(d) $\tau = 200$ (e) $\tau = 20$ (f) $\tau = 1$

Experiment with outliers

Imbalance distributions *(upper Fig.)*: Gaussian Mixtures $\mathbb{P}_0$ (gray), $\mathbb{P}_1$ (beige) with class imbalance.

Outliers distributions *(lower Figs.)*: Gaussian Mixtures $\mathbb{P}_0$ (gray, 5% outliers), $\mathbb{P}_1$ (beige, 5% outliers), $\mathbb{P}_2$ (brown).

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

Divergences: KL ($\tau\in[1,20,200]$) vs Id

Results: for small τ, our solver is robust to class imbalance and outliers issues. Increasing τ yields a more precise barycenter by incorporating all data points.