Continuous Wasserstein-2 Barycenter Estimation without Minimax Optimization

Alexander Korotin [1] Lingxiao Li [2] Justin Solomon [2] Evgeny Burnaev [1]

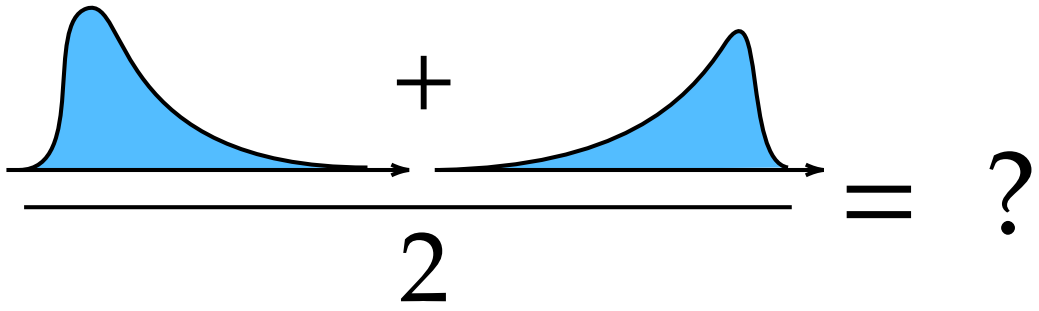
[1] Skolkovo Institute of Science and Technology
[2] Massachusetts Institute of Technology

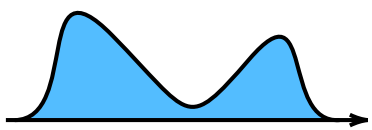
ICLR 2021

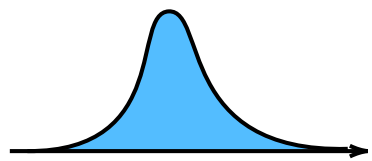
Skoltech
Skolkovo Institute of Science and Technology

MIT

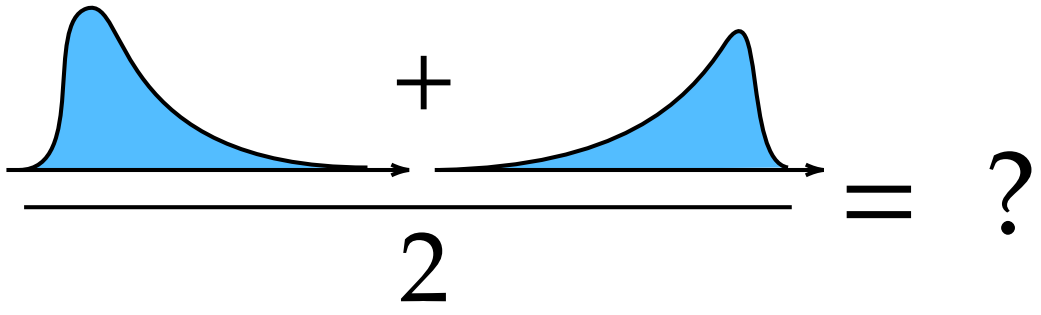
Motivation: Averaging Distributions

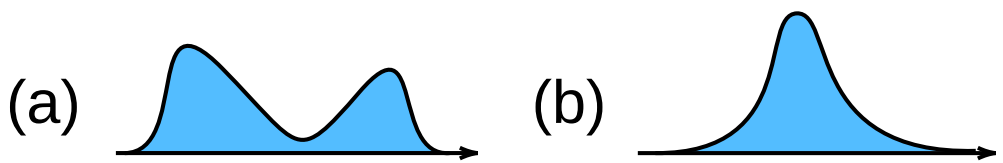


(a) 

(b) 

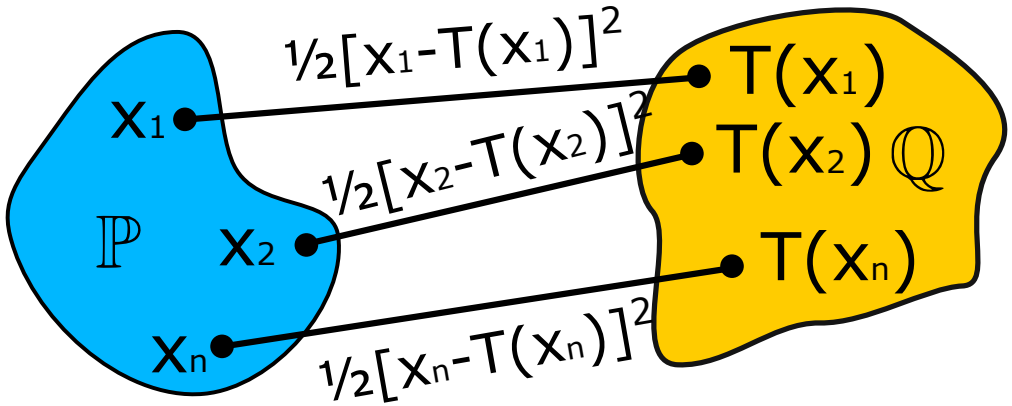
Motivation: Averaging Distributions





Need to capture the metric of the underlying space!

Wasserstein-2 Distance



The (squared) Wasserstein-2 distance between $\mathbb{P}$ and $\mathbb{Q}$ is defined by:

$$\mathbb{W}_2^2(\mathbb{P},\mathbb{Q}) \triangleq \min_{T:T_\sharp\mathbb{P}=\mathbb{Q}} \int_{\mathbb{R}^D} \frac{\|x-T(x)\|^2}{2} \mathrm{d}\mathbb{P}(x).$$

This is often referred to as the Monge form.

Wasserstein-2 Distance: Dual Form

With mild assumptions on $\mathbb{P}, \mathbb{Q}$, the (squared) Wasserstein-2 distance admits a dual form:

$$\mathbb{W}_2^2(\mathbb{P}, \mathbb{Q}) = \int_{\mathbb{R}^D} \frac{\|x\|^2}{2} \mathrm{d}\mathbb{P}(x) + \int_{\mathbb{R}^D} \frac{\|y\|^2}{2} \mathrm{d}\mathbb{Q}(y) - \min_{\psi \in \mathcal{C}} \left[\int_{\mathbb{R}^D} \psi(x) \mathrm{d}\mathbb{P}(x) + \int_{\mathbb{R}^D} \overline{\psi}(y) \mathrm{d}\mathbb{Q}(y) \right],$$

where

$$\mathcal{C} \stackrel{\triangle}{=} \{f : \mathbb{R}^D \to \mathbb{R} \cup \{\infty\} \mid f \text{ is convex}\}$$

$$\overline{\psi}(y) \stackrel{\triangle}{=} \sup_{x \in \mathbb{R}^D} \left[\langle x, y \rangle - \psi(x)\right].$$

Relating Optimal Transport Maps and Potentials

A key property of Wasserstein-2 distance is that the optimal transport map T^* and the optimal potential ψ^* can be recovered from one another:

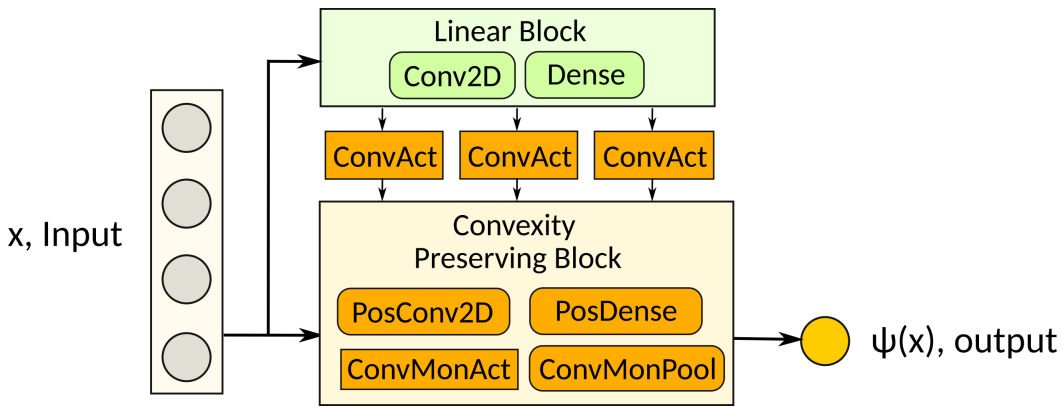
$$T^* = \nabla\psi^* \qquad \text{and} \qquad (T^*)^{-1} = \nabla\overline{\psi^*}.$$

Relating Optimal Transport Maps and Potentials

A key property of Wasserstein-2 distance is that the optimal transport map T^* and the optimal potential ψ^* can be recovered from one another:

$$T^* = \nabla \psi^* \qquad \text{and} \qquad (T^*)^{-1} = \nabla \overline{\psi^*}.$$

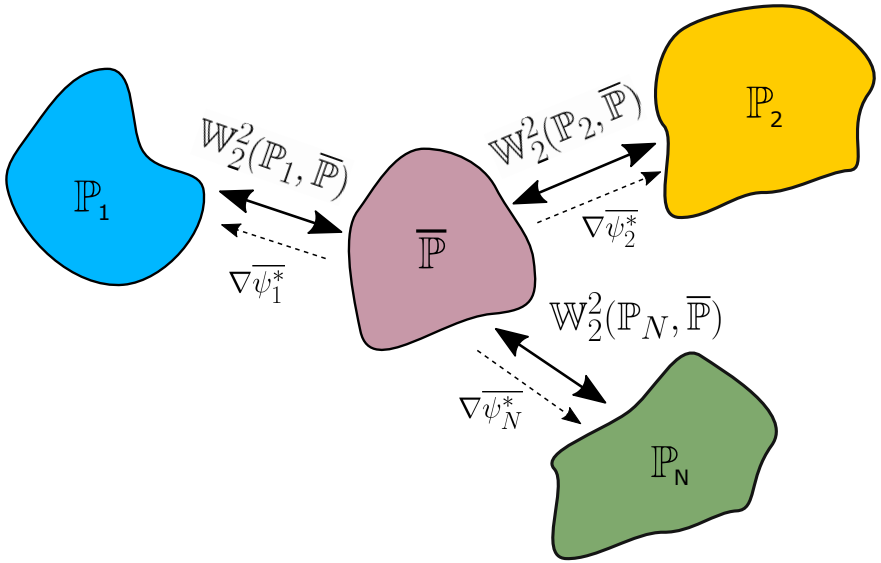
Strategy to compute Wasserstein-2 distance: parameterize the potential using Input Convex Neural Network (ICNN), and then recover the transport map by taking gradients.



Typical ICNN architecture[1].
Image source: *Wasserstein-2 Generative Networks*[2].

[1] Amos, Xu, and Kolter 2017.
[2] Korotin et al. 2019.

Wasserstein-2 Barycenter



The Wasserstein-2 barycenter of $\mathbb{P}_1, \ldots, \mathbb{P}_N$ with weights $\alpha_1, \ldots, \alpha_N$ is:

$$\overline{\mathbb{P}} \triangleq \operatorname*{argmin}_{\mathbb{P}} \sum_{n=1}^{N} \alpha_n \mathbb{W}_2^2(\mathbb{P}_n, \mathbb{P}).$$

Let ψ_n^* denote the optimal potential for $\mathbb{W}_2^2(\mathbb{P}_n, \mathbb{P})$, and $\overline{\psi_n^*}$ be its convex conjugate.

Wasserstein-2 Barycenter: Dual Form

The optimal potentials $\{\psi_n^*\}_{n=1}^N$ can be shown to be **congruent**, i.e.,

$$\sum_{n=1}^{N} \alpha_n \nabla \overline{\psi_n^*}(x) = x \qquad \text{and} \qquad \sum_{n=1}^{N} \alpha_n \overline{\psi_n^*}(x) \leq \frac{\|x\|^2}{2}.$$

Define **multiple correlations** to be

$$\text{MultiCorr}(\{\psi_n\}) \triangleq \sum_{n=1}^{N} \alpha_n \int_{\mathbb{R}^D} \psi_n(y) \mathrm{d}\mathbb{P}_n(y).$$

Then the optimal potentials $\{\psi_n^*\}_{n=1}^N$ of the barycenter problem can be found via

$$\{\psi_n^*\} = \min_{\{\psi_n\} \text{ convex, congruent}} \text{MultiCorr}(\{\psi_n\}).$$

Challenges

- How to keep track of the conjugate of a convex function?
  - ⋆ Parameterize both potential and its conjugate using two separate ICNNs $(\psi_n^{\dagger}, \overline{\psi_n^{\ddagger}})$, and then use a cycle-consistency regularizer[3]:

$$\mathcal{R}_2^{\mathbb{P}_n}(\psi_n^{\dagger}, \overline{\psi_n^{\ddagger}}) \triangleq \int_{\mathbb{R}^D} \left\| \nabla \overline{\psi_n^{\ddagger}} \circ \nabla \psi_n^{\dagger}(x) - x \right\|_2^2 \mathrm{d}\mathbb{P}_n(x).$$

[3] Korotin et al. 2019.

Challenges

- How to keep track of the conjugate of a convex function?

  ⋆ Parameterize both potential and its conjugate using two separate ICNNs $(\psi_n^\dagger, \overline{\psi_n^\ddagger})$, and then use a cycle-consistency regularizer[3]:

$$\mathcal{R}_2^{\mathbb{P}_n}(\psi_n^\dagger, \overline{\psi_n^\ddagger}) \triangleq \int_{\mathbb{R}^D} \left\| \nabla \overline{\psi_n^\ddagger} \circ \nabla \psi_n^\dagger(x) - x \right\|_2^2 \mathrm{d}\mathbb{P}_n(x).$$

- How to enforce congruency of potentials?

  ⋆ We propose to penalize non-congruence potentials as follows, for a given reference measure $\hat{\mathbb{P}}$ (alternatives[4] also exist):

$$\tau \mathcal{R}_1^{\hat{\mathbb{P}}}(\{\overline{\psi_n^\ddagger}\}) \triangleq \tau \int_{\mathbb{R}^D} \left[\sum_{n=1}^N \alpha_n \overline{\psi_n^\ddagger}(y) - \frac{\|y\|^2}{2} \right]_+ \mathrm{d}\hat{\mathbb{P}}(y).$$

Please refer to our paper for more details.

[3] Korotin et al. 2019.

[4] Li et al. 2020.

Non-Minimax Objective

With two regularizers, we obtain our final non-minimax objective:

$$\min_{\{\psi_n^\dagger, \overline{\psi_n^\ddagger}\} \subset \mathcal{C}} \left[\text{MultiCorr}(\{\psi_n^\dagger\}) + \lambda \sum_{n=1}^{N} \alpha_n \mathcal{R}_2^{\mathbb{P}_n}(\psi_n^\dagger, \overline{\psi_n^\ddagger}) + \tau \mathcal{R}_1^{\hat{\mathbb{P}}}(\{\overline{\psi_n^\ddagger}\}) \right].$$

We parameterize $\{\psi_n^\dagger, \overline{\psi_n^\ddagger}\}$ using ICNNs, and the optimization can be solved using SGD by sampling from $\mathbb{P}_n$'s and $\hat{\mathbb{P}}$.

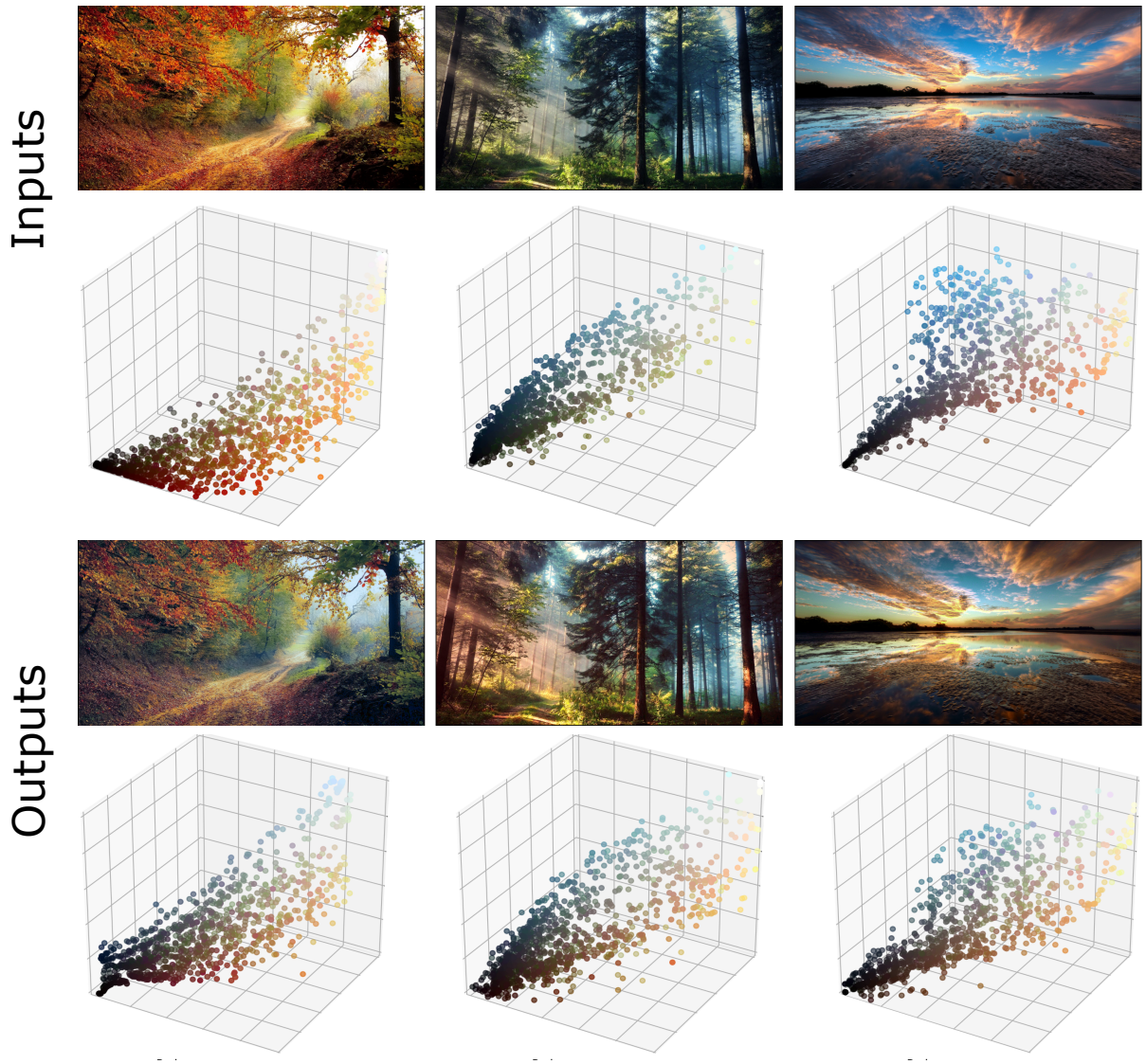
Guarantees

- If $\tau \cdot \hat{\mathbb{P}} \geq \overline{\mathbb{P}}$, then the regularized objective is a **tight** upper bound for the true MultiCorr. In particular, our regularization **does not introduce bias**.
- If the regularized objective is Δ-close to the true optimum, then for each n,

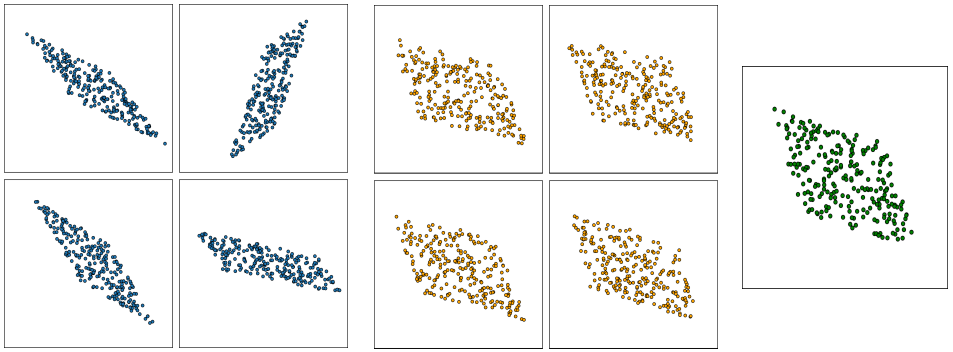
$$\mathbb{W}_2^2((\nabla\psi_n^\dagger)_\sharp \mathbb{P}_n, \overline{\mathbb{P}}) \leq O\left(\frac{\Delta}{\alpha_n}\right).$$

  This suggests $(\nabla\psi_n^\dagger)_\sharp \mathbb{P}_n$ are good candiates for the barycenter.

Application: Averaging Color Palettes



Quantitative Results: Location-Scatter Families in High Dimensions



Inputs Outputs Ground Truth

location scatter family generated by Gaussian distributions

Metric	Method	D=2	4	8	16	32	64	128	256
$B\mathbb{W}_2^2$-UVP, %	[FC$\mathbb{W}$B], Cuturi & Doucet (2014)	0.7	0.68	1.41	3.87	8.85	14.08	18.11	21.33
$\mathcal{L}_2$-UVP, % (potentials)	[SC$\mathbb{W}_2$B], (Fan et al., 2020)	0.07	0.09	0.16	0.28	0.43	0.59	1.28	2.85
		0.08	0.10	0.17	0.29	0.47	0.63	1.14	1.50
	[CR$\mathbb{W}$B], (Li et al., 2020)	0.99	2.52	8.62	22.23	67.01	>100		
	[C$\mathbb{W}_2$B], **ours**	0.06	0.05	0.07	0.11	0.19	0.24	0.42	0.83

location scatter family generated the Uniform distribution on a cube

Metric	Method	D=2	4	8	16	32	64	128	256
$B\mathbb{W}_2^2$-UVP, %	[FC$\mathbb{W}$B], Cuturi & Doucet (2014)	0.64	0.77	1.22	3.75	8.92	14.3	18.46	21.64
$\mathcal{L}_2$-UVP, % (potentials)	[SC$\mathbb{W}_2$B], (Fan et al., 2020)	0.12	0.10	0.19	0.29	0.46	0.6	1.38	2.9
		0.17	0.12	0.2	0.31	0.47	0.62	1.21	1.52
	[CR$\mathbb{W}$B], (Li et al., 2020)	0.58	1.83	8.09	21.23	55.17	> 100		
	[C$\mathbb{W}_2$B], **ours**	0.17	0.08	0.06	0.1	0.2	0.25	0.42	0.82

Metric: Unexplained Variance Percentage $\mathrm{UVP}(\tilde{\mathbb{P}}) = 100 \frac{\mathbb{W}_2^2(\tilde{\mathbb{P}}, \overline{\mathbb{P}})}{1/2 \mathrm{Var}(\overline{\mathbb{P}})} \%$

Thank You!

We will be at Poster Session 8, ID: 2976, 9-11am on May 5 (PDT).

ArXiv: `https://arxiv.org/abs/2102.01752`.

Code + Poster:
`https://github.com/iamalexkorotin/Wasserstein2Barycenters`.