Parametric Methods for Computing Optimal Transport Maps, Distances, Barycenters

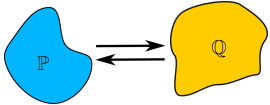
Evgeny Burnaev
Prof., Head of Skoltech Applied AI Center
Head of AIRI Research Group

Skoltech

Handling Distributions in Machine Learning Tasks

Mapping/comparing

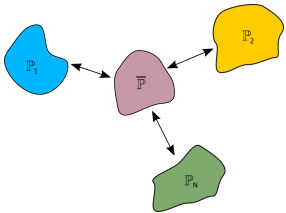
2 distributions



generative modeling, domain adaptation, image manipulation: enhancement, super-resolution, style transfer, etc.

Averaging

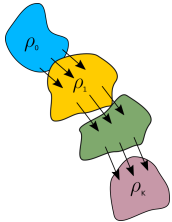
N distributions



shape interpolation, texture mixing, aggregating probabilistic forecasts, etc.

Modeling dynamics

Sequence of distributions



inference of diffusion processes appearing in machine learning, economics, physics, etc.

This presentation

presents scalable neural methods to solve these problems based on the **Optimal Transport** (OT) theory

Overview

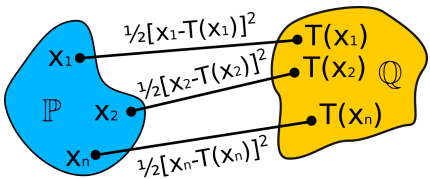
1. Introduction to Optimal Transport
2. Existing OT Methods: Challenges and Limitations
3. Overview of presented results
4. A Continuous $\mathbb{W}_2$ Benchmark [Problem 0]
5. A Non-Minimax Algorithm to Compute OT Maps for $\mathbb{W}_2$ [Problem A]
6. A Non-Minimax Algorithm for Continuous $\mathbb{W}_2$ Barycenters [Problem B]
7. A Neural Implementation of the JKO scheme for $\mathbb{W}_2$ Gradient Flows [Problem C]
8. Optimal Transport Modeling
9. Summary: Publications, Presentations

1. Introduction to Optimal Transport

Monge's Optimal Transport for the Quadratic Cost[1]

The (square of) the **Wasserstein-2 distance** between $\mathbb{P} \in \mathcal{P}_2(\mathbb{R}^D)$ and $\mathbb{Q} \in \mathcal{P}_2(\mathbb{R}^D)$ is

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



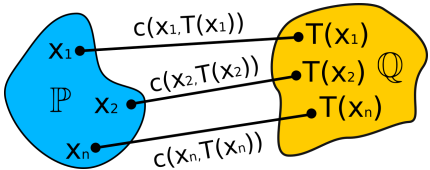
The map T^* attaining the minimum is called the **optimal transport map**.

Problem A: Computing the OT map & distance.

[1] Cédric Villani (2008). *Optimal transport: old and new.* Vol. 338. Springer Science & Business Media.

Monge's Formulation of Optimal Transport[2]

Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be a cost function, e.g., $c(x, y) = \frac{\|x-y\|^2}{2}$.



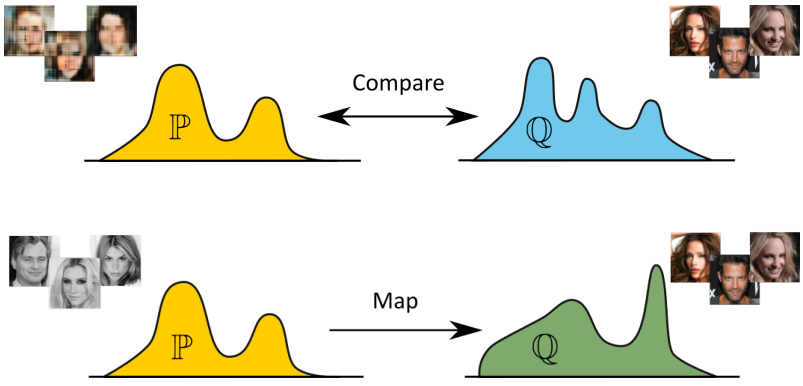
The optimal transport **cost** between measures $\mathbb{P}$ and $\mathbb{Q}$ is

$$\text{Cost}(\mathbb{P}, \mathbb{Q}) = \min_{T\sharp\mathbb{P}=\mathbb{Q}} \int_{\mathbb{R}^D} c\big(x, T(x)\big) d\mathbb{P}(x).$$

The map T^* attaining the minimum is called
the optimal **transport map** between $\mathbb{P}$ and $\mathbb{Q}$.

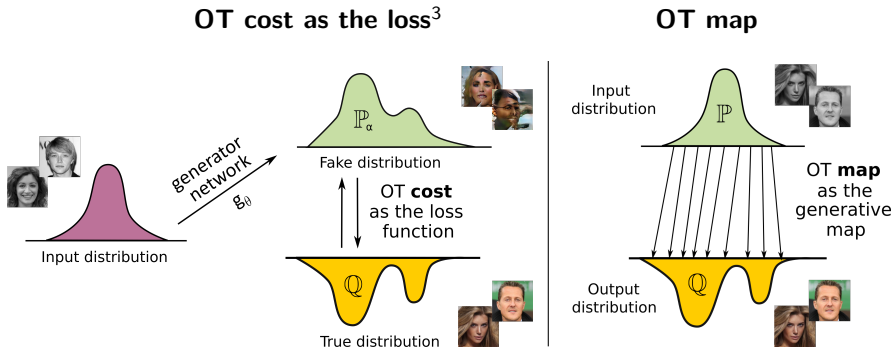
[2] Cédric Villani (2008). *Optimal transport: old and new.* Vol. 338. Springer Science & Business Media.

Optimal Transport in Machine Learning Tasks



Problem A: Computing the OT map & distance.

Approaches to Use OT in Large-Scale Generative Models

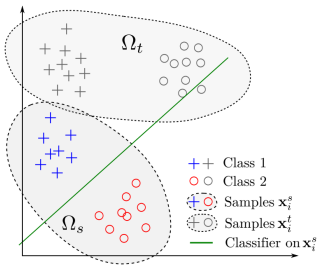


[3]Martin Arjovsky, Soumith Chintala, and Léon Bottou (2017). "Wasserstein generative adversarial networks". In: *International conference on machine learning*. PMLR, pp. 214–223.

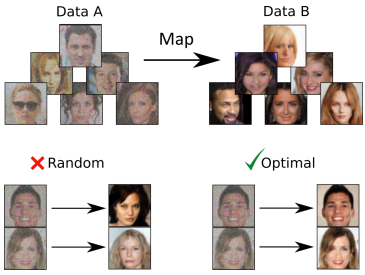
Why Do We Need Optimal Maps?

Practical tasks[4][5]

Domain Adaptation



Unpaired Learning



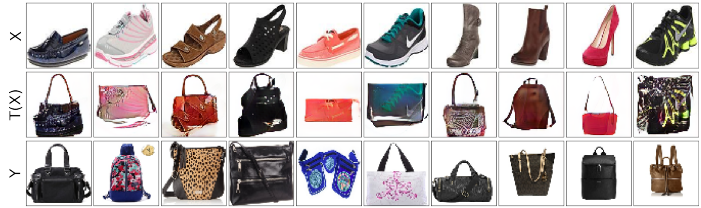
[4]Nicolas Courty et al. (2016). "Optimal transport for domain adaptation". In: *IEEE transactions on pattern analysis and machine intelligence* 39.9, pp. 1853–1865.

[5]Yujia Xie et al. (2019). "On scalable and efficient computation of large scale optimal transport". In: *International Conference on Machine Learning*. PMLR, pp. 6882–6892.

Examples (Unpaired Style Transfer)

Handbags $\longrightarrow$ shoes (64×64)

Shoes $\longrightarrow$ handbags (64×64)



Examples (Unpaired Style Transfer)

Outdoor $\longrightarrow$ churches (64×64)

Faces $\longrightarrow$ anime (64×64)

Examples (Unpaired Style Transfer)

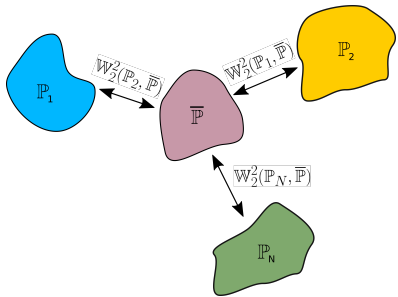
Handbags $\longrightarrow$ shoes (128×128)

Faces $\longrightarrow$ anime (128×128)

Barycenters in the Wasserstein-2 Space[6]

The **Wasserstein-2 barycenter** $\overline{\mathbb{P}}$ of distributions $\mathbb{P}_1, \mathbb{P}_2, \cdots, \mathbb{P}_N \in \mathcal{P}_2(\mathbb{R}^D)$ w.r.t. weights $\alpha_1, \ldots, \alpha_N \geq 0$ ($\sum_{n=1}^N \alpha_n = 1$) is defined by

$$\overline{\mathbb{P}} = \operatorname*{arg\,min}_{\mathbb{P} \in \mathcal{P}_2(\mathbb{R}^D)} \sum_{n=1}^{N} \alpha_n \mathbb{W}_2^2(\overline{\mathbb{P}}, \mathbb{P}_n).$$



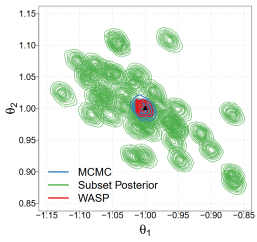
Problem B: Computing the Wasserstein-2 barycenter.

[6] Martial Agueh and Guillaume Carlier (2011). "Barycenters in the Wasserstein space". In: *SIAM Journal on Mathematical Analysis* 43.2, pp. 904–924.

Applications of Wasserstein-2 Barycenters

Practical tasks[7][8]

Subset Posterior Aggregation



$$p(\theta|D) \approx \text{Barycenter}\big[p(\theta|D_1), \ldots, p(\theta|D_N))\big]$$

Color/Style Mixing

[7] Sanvesh Srivastava, Cheng Li, and David B Dunson (2018). "Scalable Bayes via barycenter in Wasserstein space". In: *The Journal of Machine Learning Research* 19.1, pp. 312–346.

[8] Youssef Mroueh (2020). "Wasserstein Style Transfer". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 842–852.

Wasserstein-2 Gradient Flows[9]

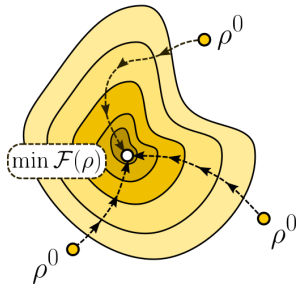
Let $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^D) \to \mathbb{R}$ be a **functional** on the space of probability distributions with the finite second moment.

Let $\frac{\delta \mathcal{F}}{\delta \rho}(\rho) : \mathbb{R}^D \to \mathbb{R}$ be its **flat derivative** at a point ρ.

The **Wasserstein-2 gradient flow** of $\mathcal{F}$ which starts at a point $\rho^0 \in \mathcal{P}_2(\mathbb{R}^D)$ is the continuous sequence of probability distributions $\rho_t \in \mathcal{P}_2(\mathbb{R}^D)$ satisfying the following PDE:

$$\frac{\partial \rho_t}{\partial t} - \underbrace{\nabla \cdot \left(\rho_t \nabla_x \frac{\delta \mathcal{F}}{\delta \rho}(\rho_t)\right)}_{-\nabla_{\mathbb{W}_2} \mathcal{F}(\rho_t)} = 0 \qquad \text{s.t. } \rho_0 = \rho^0.$$

$$\frac{\partial \rho_t}{\partial t} = -\nabla_{\mathbb{W}_2} \mathcal{F}(\rho_t)$$

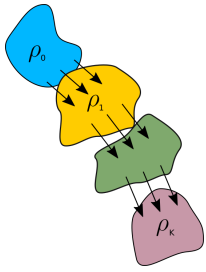


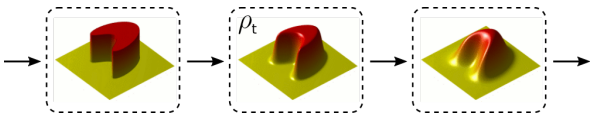
[9] Filippo Santambrogio (2017). "{Euclidean, metric, and Wasserstein} gradient flows: an overview". In: *Bulletin of Mathematical Sciences* 7.1, pp. 87–154.

Common PDEs and Flow Functionals[10][11]

Certain **well-celebrated PDEs** are the Wasserstein-2 gradient flows:

Class	**PDE** $\frac{\partial \rho_t}{\partial t} =$	**Flow functional** $\mathcal{F}(\rho) =$
Heat Equation	$\Delta\rho$	$\int_{\mathbb{R}^D} \log \frac{d\rho}{dx} d\rho(x)$
Advection	$\nabla \cdot (\rho \nabla V)$	$\int_{\mathbb{R}^D} V(x) d\rho(x)$
Fokker-Plank	$\nabla \cdot (\rho \nabla V) + \Delta\rho$	$\int_{\mathbb{R}^D} V(x) d\rho(x) + \int_{\mathbb{R}^D} \log \frac{d\rho}{dx} d\rho(x)$





[10] David Alvarez-Melis, Yair Schiff, and Youssef Mroueh (2022). "Optimizing Functionals on the Space of Probabilities with Input Convex Neural Networks". In: *Transactions on Machine Learning Research*. URL: https://openreview.net/forum?id=dpOYN7o8Jm.

[11] Image source: https://en.wikipedia.org/wiki/Heat_equation

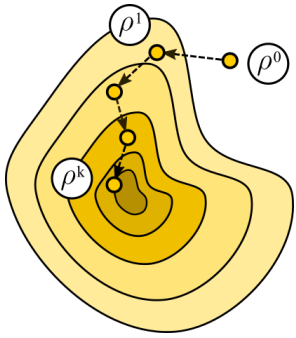
Computing the Wasserstein-2 Gradient Flows[12]

Jordan Kinderlehrer and Otto proposed to compute a **discrete** sequence $\rho^1, \rho^2, \ldots$ given by

$$\rho^{k+1} \leftarrow \underset{\rho \in \mathcal{P}_2(\mathbb{R}^D)}{\arg\min} \left[\mathcal{F}(\rho) + \frac{1}{\tau} \mathbb{W}_2^2(\rho^k, \rho) \right].$$

For $\tau \to 0$, it holds that $\rho^k \approx \rho_{\tau \cdot k}$, i.e., time-discretized gradient flow converges to the true continuous flow.

The practical implementation of the JKO scheme is **non-trivial** as it requires computing the $\mathbb{W}_2$ distance term.



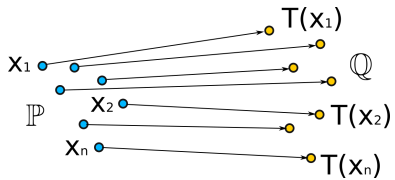
Problem C: Computing the Wasserstein-2 gradient flow.

[12] Richard Jordan, David Kinderlehrer, and Felix Otto (1998). "The variational formulation of the Fokker-Planck equation". In: *SIAM journal on mathematical analysis* 29.1, pp. 1–17.

2. Existing OT Methods: Challenges and Limitations

Types of Optimal Transport Methods[13]

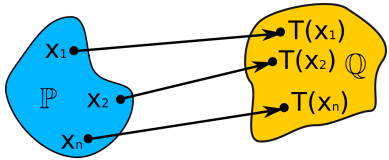
Discrete



+ Convex optimization;
+ Strong theoretical guarantees;
- Poor scalability;
- No out-of-support estimates;

Continuous

(Focus of the presentation)



± Neural networks;
± Limited guarantees;
+ Good scalability;
+ Out-of-sample estimation

[13] Gabriel Peyré, Marco Cuturi, et al. (2019). "Computational optimal transport: With applications to data science". In: *Foundations and Trends® in Machine Learning* 11.5-6, pp. 355–607.

Dual Form Methods

Primal form

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

Dual form[14]

$$\mathbb{W}_2^2(\mathbb{P}, \mathbb{Q}) = \max_{f,g} \left[\int_{\mathbb{R}^D} f(x) d\mathbb{P}(x) + \int_{\mathbb{R}^D} g(y) d\mathbb{Q}(y) \right],$$

where $f, g \in \mathcal{L}^1(\mathbb{P}), \mathcal{L}^1(\mathbb{Q})$ satisfy $f(x) + g(y) \leq \frac{\|x-y\|^2}{2}$ for $x, y \in \mathbb{R}^D$.

Dual form (c-transform)

$$\mathbb{W}_2^2(\mathbb{P}, \mathbb{Q}) = \max_{f} \left[\int_{\mathbb{R}^D} f(x) d\mathbb{P}(x) + \int_{\mathbb{R}^D} f^c(y) d\mathbb{Q}(y) \right],$$

where $f^c(y) \stackrel{def}{=} \min_{x \in \mathbb{R}^D} \left[\frac{1}{2}\|x - y\|^2 - f(x) \right]$ is called the c-transform of f.

[14] Cédric Villani (2008). *Optimal transport: old and new.* Vol. 338. Springer Science & Business Media.

Approaches to Solve the Dual Form

PRIMAL-DUAL RELATION: extract T^* from f^*

$$T^*(x) = x - \nabla f^*(x)$$

Most existing continuous methods solve the dual problem, i.e., obtain the potential f^*, and then recover the primal solution from it, i.e., the transport map $T^* = x - \nabla f^*(x)$.

Existing dual-form methods are of **two types**:

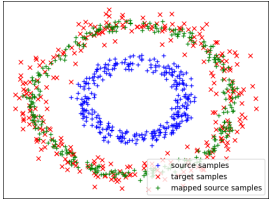
1. (**Regularized OT**) Soft penalization of potentials f, g for disobeying $f \oplus g \leq \frac{1}{2}\|\cdot\|^2$.
   Such methods recover **biased** solutions;
2. (**Maximin OT**) Maximin optimization of potentials f, g via the c-transform.
   Such methods suffer from **training instabilities**.

Entropy-regularized Optimal Transport[15]

Regularized dual form

$$\mathbb{W}_{2,\epsilon}^2(\mathbb{P},\mathbb{Q}) = \max_{f,g}\left[\int_{\mathbb{R}^D} f(x)d\mathbb{P}(x) + \int_{\mathbb{R}^D} g(y)d\mathbb{Q}(y) - \mathcal{R}_{\text{Ent}}^{\epsilon}(f,g)\right]$$

$$\mathcal{R}_{\text{Ent}}^{\epsilon}(f,g) = \epsilon \int_{\mathbb{R}^D\times\mathbb{R}^D} \exp\frac{f(x)+g(y)-\frac{\|x-y\|^2}{2}}{\epsilon} d\big(\mathbb{P}\times\mathbb{Q}\big)$$



Problems: highly biased for $\epsilon \gg 0$, unstable for $\epsilon \to 0$.

[15] Vivien Seguy et al. (2018). "Large Scale Optimal Transport and Mapping Estimation". In: *International Conference on Learning Representations*.

Brenier Optimal Transport with the Quadratic Cost

$$\mathbb{W}_2^2(\mathbb{P},\mathbb{Q}) = -\min_{\psi\in\text{Conv}}\Bigg[\overbrace{\int\psi(x)d\mathbb{P}(x)+\int\overline{\psi}(y)d\mathbb{Q}(y)}^{\text{Corr}(\mathbb{P},\mathbb{Q}|\psi,\overline{\psi})}\Bigg]+\text{Const}(\mathbb{P},\mathbb{Q})$$

$$\overline{\psi}(y)=\max_x\left(\langle x,y\rangle-\psi(x)\right)$$

Minimax Approach[16]

$$\min_{\psi\in\text{Convex}}\left[\int\psi(x)d\mathbb{P}(x)+\int\overline{\psi}(y)d\mathbb{Q}(y)\right]=$$

$$\min_{\psi\in\text{Conv}}\max_{\phi\in\text{Conv}}\Bigg[\underbrace{\int\psi(x)d\mathbb{P}(x)+\int\left[\langle\nabla\phi(y),y\rangle-\psi(\nabla\phi(y))\right]d\mathbb{Q}(y)}_{\text{Corr}(\mathbb{P},\mathbb{Q}|\psi,\phi)}\Bigg]$$

Relation to the optimal Kantorovich potential f^*:

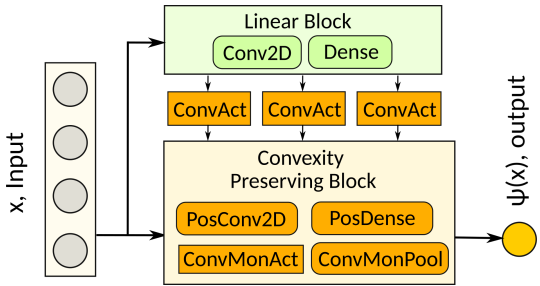
$$\psi^*(x)=\frac{\|x\|^2}{2}-f^*(x)\qquad\text{and}\qquad\nabla\psi^*(x)=x-\nabla f^*(x)=T^*(x)$$

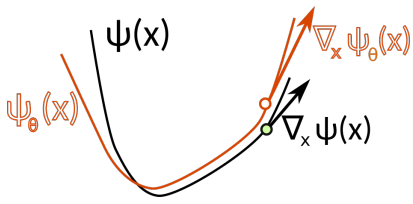
[16] Ashok Makkuva et al. (2020). "Optimal transport mapping via input convex neural networks". In: *International Conference on Machine Learning*. PMLR, pp. 6672–6681.

Input Convex Neural Networks

Approximate convex function $\psi(x) : \mathbb{R}^D \to \mathbb{R}$ by neural nets!

- $\psi_\theta : \mathbb{R}^D \to \mathbb{R}$ – deep input convex neural network[17] (ICNN);
- $T_\theta = \nabla_x \psi_\theta : \mathbb{R}^D \to \mathbb{R}^D$ - transport map.

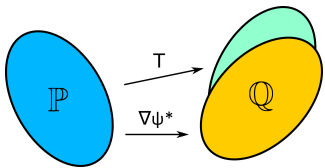




[17] Brandon Amos, Lei Xu, and J Zico Kolter (2017). "Input convex neural networks". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, pp. 146–155.

How to Assess the Results?

Problem: How to understand whether continuous OT methods recover the OT map T^* well?



- There are **no benchmarks** in the field of continuous OT.
- There exists a **limited** amount of pairs $(\mathbb{P}, \mathbb{Q})$ with analytically known **ground truth** T^*; these pairs are rather trivial and not representative, e.g., 1-dimensional, Gaussian, etc.
- New methods are tested on a restricted set of self-generated ad-hoc examples;

Problem 0: Benchmarking continuous OT methods.

3. Overview of presented results

Main results

1. (**Problem 0**). The novel methodology to construct **benchmark** pairs of continuous distributions with analytically known OT maps and distances for $\mathbb{W}_2$ between them. The methodology enables quantitative evaluation of existing methods for $\mathbb{W}_2$.
2. (**Problem A**). Algorithm for computing the $\mathbb{W}_2$ **map** between continuous distributions;
3. (**Problem B**). Algorithm for computing $\mathbb{W}_2$ **barycenters** of continuous distributions.
4. (**Problem C**). Algorithm for computing the $\mathbb{W}_2$ **gradient flows**.

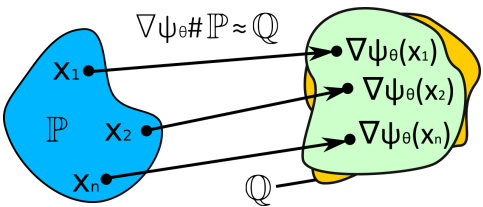
4. A Continuous $\mathbb{W}_2$ Benchmark [Problem 0]

Do Neural Optimal Transport Solvers Work? A Continuous Wasserstein-2 Benchmark (NeurIPS 2021, A*)

https://openreview.net/forum?id=CI0T_3l-n1

Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, Alexander Filippov, Evgeny Burnaev

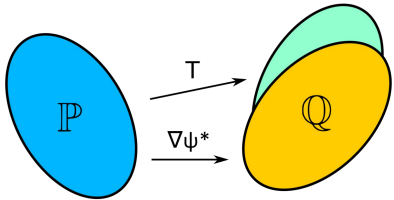
A generic methodology to construct pairs of continuous distributions with analytically-known optimal transport (OT) solutions for the quadratic cost to test continuous OT methods.



Existing Metrics for OT Solvers

1. **Indirect Metrics** use an OT solver as a component in a larger pipeline, using end-to-end performance as a proxy for solver quality.
   - ☺ A lot of metrics exist, e.g., FID, Inception scores for GANs.
   - ☹ Do not provide understanding about the quality of the solver itself

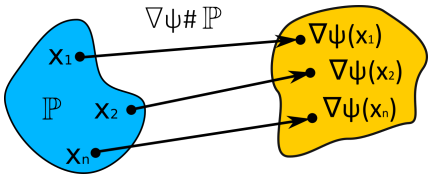
2. **Direct metics** compare computed $\hat{T}$ with the true T^*;
   - ☹ Limited number of pairs $(\mathbb{P}, \mathbb{Q})$ with known T^*.
   - ☹ Limited amount of metrics.



Benchmark Key Idea

Let $\mathbb{P}$ be a continuous measure with finite second moment on $\mathbb{R}^D$.

Let $\nabla\psi\sharp\mathbb{P}$ be its pushforward with some convex $\psi : \mathbb{R}^D \to \mathbb{R}$.



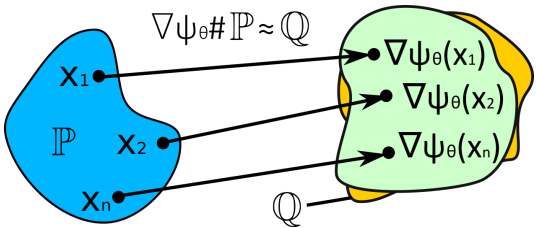
Then $\nabla\psi$ is the OT map from $\mathbb{P}$ to $\nabla\psi\sharp\mathbb{P}$ (**Brenier's Theorem**).[18][19]

[18] Robert J McCann et al. (1995). "Existence and uniqueness of monotone measure-preserving maps". In: *Duke Mathematical Journal* 80.2, pp. 309–324.

[19] Yann Brenier (1991). "Polar factorization and monotone rearrangement of vector-valued functions". In: *Communications on pure and applied mathematics* 44.4, pp. 375–417.

Approximating Arbitrary Pairs

Let $(\mathbb{P}, \mathbb{Q})$ be a pair of continuous measures on $\mathbb{R}^D$.

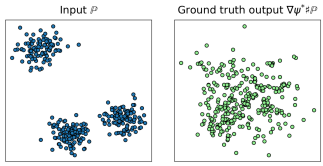


1. We approximate the OT map from $\mathbb{P}$ to $\mathbb{Q}$ by an ICNN $\nabla\psi_\theta$, i.e. we train ψ_θ by the OT solver with ICNN parametrization.
2. We use the *approximate* pair $(\mathbb{P}, \nabla\psi_\theta\sharp\mathbb{P}) \approx (\mathbb{P}, \mathbb{Q})$ as a pair of benchmark measures with known OT solution.

Developed Benchmark Pairs

HIGH-DIMENSIONAL benchmark distributions

$D = 2, 4, 8, \ldots, 256$



IMAGES benchmark distributions (based on Celeba[20] faces)

3 pairs, $D = 12288$

[20] Ziwei Liu et al. (2015). "Deep Learning Face Attributes in the Wild". In: *Proceedings of International Conference on Computer Vision (ICCV)*.

Evaluation Metrics

- $\mathcal{L}^2$-**unexplained variance percentage**[21] of the computed map $\hat{T}$

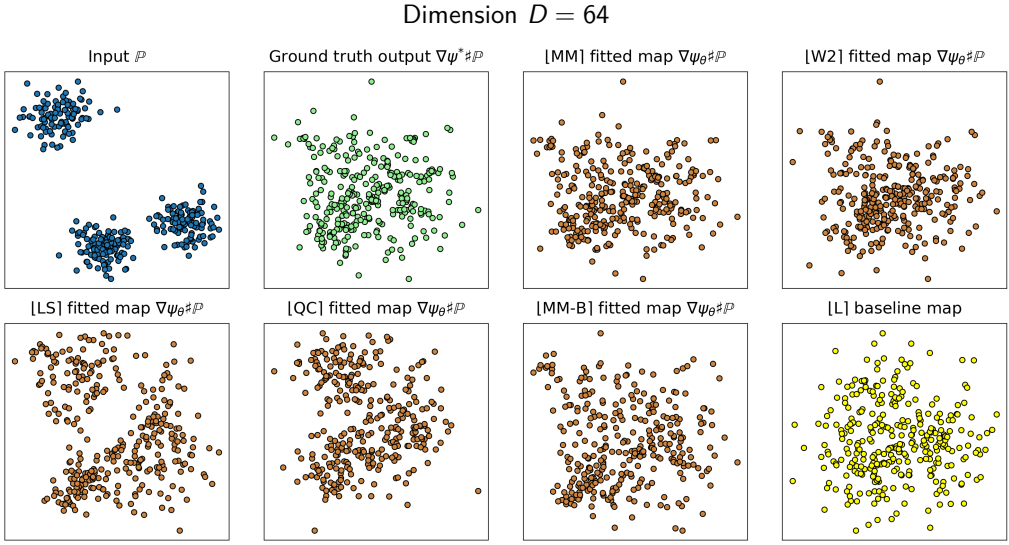
$$\mathcal{L}^2\text{-UVP}(\hat{T}, T^*) = 100\% \times \frac{\|\hat{T} - T^*\|^2_{\mathcal{L}^2(\mathbb{P})}}{\text{Var}(\mathbb{Q})}$$

- **Cosine similarity** of $\nabla\hat{f} : x \mapsto x - \hat{T}(x)$ and the ground truth ∇f^*

$$\cos(\underbrace{\text{id} - \hat{T}}_{\nabla\hat{f}}, \underbrace{\text{id} - T^*}_{\nabla f^*}) \stackrel{\text{def}}{=} \frac{\langle \nabla f^*, \nabla\hat{f} \rangle_{\mathcal{L}^2(\mathbb{P})}}{\|\nabla f^*\|_{\mathcal{L}^2(\mathbb{P})} \cdot \|\nabla\hat{f}\|_{\mathcal{L}^2(\mathbb{P})}} \in [-1, 1]$$

[21] Alexander Korotin, Vage Egiazarian, Arip Asadulaev, et al. (2021). "Wasserstein-2 Generative Networks". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=bEoxzW_EXsa.

High-dimensional Benchmark Pairs: Qualitative Results



High-dimensional Benchmark Pairs: Quantitative Results

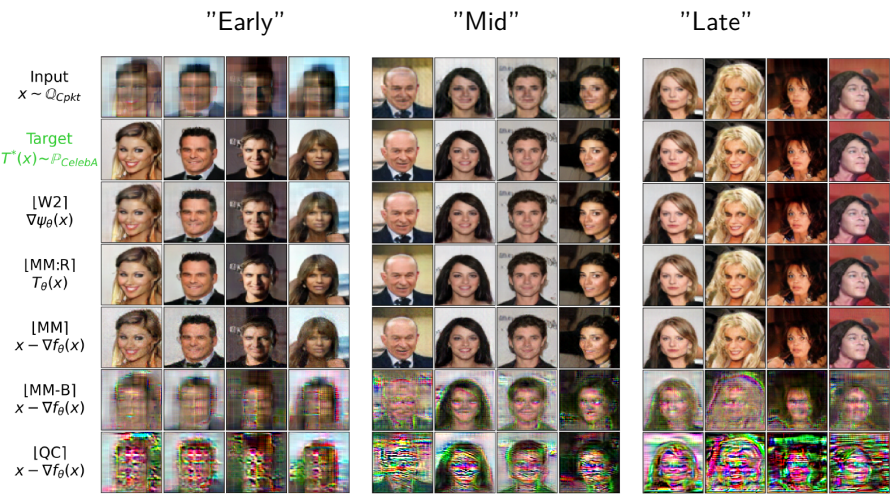
Dim	2	4	8	16	32	64	128	256
⌊W2⌉ **(Ours)**	0.1	0.7	2.6	3.3	6.0	7.2	2.0	2.7
⌊W2:R⌉	0.2	0.9	4.0	5.3	5.2	7.0	2.0	2.7
⌊MMv1⌉	0.2	1.0	1.8	1.4	6.9	8.1	2.2	2.6
⌊MM⌉	0.1	0.3	0.9	2.2	4.2	3.2	3.1↯	4.1↯
⌊MM:R⌉	0.1	0.3	0.7	1.9	2.8	4.5	↛	↛
⌊MMv2⌉	0.1	0.68	2.2	3.1	5.3	10.1↯	3.2↯	2.7↯
⌊MMv2:R⌉	0.1	0.7	4.4	7.7	5.8	6.8	2.1	2.8
⌊MM-B⌉	0.1	0.7	3.1	6.4	12.0	13.9	19.0	22.5
⌊LS⌉	5.0	11.6	21.5	31.7	42.1	40.1	46.8	54.7
⌊L⌉	14.1	14.9	27.3	41.6	55.3	63.9	63.6	67.4
⌊QC⌉	1.5	14.5	28.6	47.2	64.0	75.2	80.5	88.2
⌊C⌉	100	100	100	100	100	100	100	100
⌊ID⌉	32.7	42.0	58.6	87	121	137	145	153

Dimensions $D = 2, 2^2, \ldots, 2^8$. Orange highlights $\mathcal{L}^2$-UVP $> 10\%$.
Red indicates performance worse than ⌊L⌉ baseline.

The experimental results confirm the issues of existing solvers:

1. Minimax methods ⌊MM⌉ are unstable and sometimes diverge (↯);
2. Entropic OT ⌊LS⌉ is notably biased in high-dimensions;
3. The proposed method ⌊W2⌉ is stable and performs well.

Images Benchmark Pairs: Qualitative Results



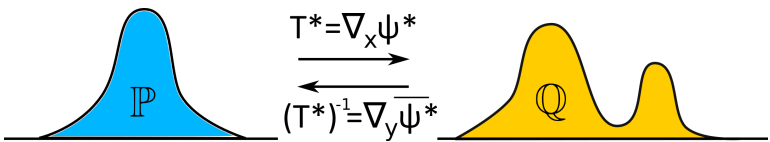
5. A Non-Minimax Algorithm to Compute OT Maps for $\mathbb{W}_2$ [Problem A]

Wasserstein-2 Generative Networks (ICLR 2021, A*)

https://openreview.net/forum?id=bEoxzW_EXsa

Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, Evgeny Burnaev

A novel end-to-end parametric method to compute optimal transport maps between continuous distributions without introducing bias or resorting to minimax optimization.



Proposed Optimization Objective

MINIMAX APPROACH (prior art)[22]

$$\min_{\psi \in \text{Convex}} \left[\int \psi(x) d\mathbb{P}(x) + \int \overline{\psi}(y) d\mathbb{Q}(y) \right] =$$

$$\min_{\psi \in \text{Conv}} \max_{\phi \in \text{Conv}} \Bigg[\underbrace{\int \psi(x) d\mathbb{P}(x) + \int \left[\langle \nabla \phi(y), y \rangle - \psi\left(\nabla \phi(y)\right) \right] d\mathbb{Q}(y)}_{\text{Corr}(\mathbb{P}, \mathbb{Q} | \psi, \phi)} \Bigg]$$

NON-MINIMAX APPROACH (this presentation): *cycle consistency* regularizer ($\lambda > 0$)

$$\min_{\psi, \phi \in \text{Conv}} \text{Corr}(\mathbb{P}, \mathbb{Q} \mid \psi, \phi; \lambda) \stackrel{def}{=} \min_{\psi, \phi \in \text{Conv}} \Bigg[\text{Corr}(\mathbb{P}, \mathbb{Q} | \psi, \phi) + \frac{\lambda}{2} \underbrace{\int_{\mathcal{Y}} \| \nabla \psi(\nabla \phi(y)) - y \|^2 d\mathbb{Q}(y)}_{\text{Cycle Reg.}} \Bigg].$$

[22] Ashok Makkuva et al. (2020). "Optimal transport mapping via input convex neural networks". In: *International Conference on Machine Learning*. PMLR, pp. 6672–6681.

Theoretical Results[23]

Theorem (primal-dual relation)

For $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_{2,ac}(\mathbb{R}^D)$, with mild assumptions on $\lambda > 0$ and convex $\hat{\psi}, \hat{\phi}$ the statements

1. $Corr(\mathbb{P}, \mathbb{Q}|\psi^*, \overline{\psi^*}) \leq Corr(\mathbb{P}, \mathbb{Q} \mid \hat{\psi}, \hat{\phi}; \lambda) \leq Corr(\mathbb{P}, \mathbb{Q}|\psi^*, \overline{\psi^*}) + O(\epsilon)$ *#dual;*
2. $\|\nabla\hat{\psi} - \nabla\psi^*\|^2_{\mathcal{L}^2(\mathbb{P})} \leq O(\epsilon)$ *and* $\|\nabla\hat{\phi} - \nabla\phi^*\|^2_{\mathcal{L}^2(\mathbb{Q})} \leq O(\epsilon)$ *#primal;*

are equivalent. Here $\nabla\psi^*, \nabla\phi^*$ are the forward and inverse OT maps.

Theorem (optimization over restricted sets of functions)

For $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_{2,ac}(\mathbb{R}^D)$, consider two sets of convex functions Φ, Ψ with mild assumptions on them. Assume that $\exists\hat{\psi} \in \Psi$ and $\exists\hat{\phi} \in \Phi$ satisfying the #primal condition. Let

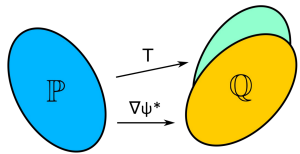
$$(\psi^\dagger, \phi^\dagger) \stackrel{def}{=} \min_{\psi\in\Psi, \phi\in\Phi} Corr(\mathbb{P}, \mathbb{Q} \mid \psi, \phi; \lambda).$$

Then it holds that $\|\nabla\psi^\dagger - \nabla\psi^*\|^2_{\mathcal{L}^2(\mathbb{P})} \leq O(\epsilon)$ and $\|\nabla\phi^\dagger - \nabla\phi^*\|^2_{\mathcal{L}^2(\mathbb{Q})} \leq O(\epsilon)$.

[23] In the theorem formulations, some technical assumptions are skipped for the simplicity.

Experiments: Gaussian Optimal Transport

Gaussian Setting: $\mathbb{P}, \mathbb{Q} = \mathcal{N}(\mu_{\mathbb{P}}, \Sigma_{\mathbb{P}}), \mathcal{N}(\mu_{\mathbb{Q}}, \Sigma_{\mathbb{Q}})$



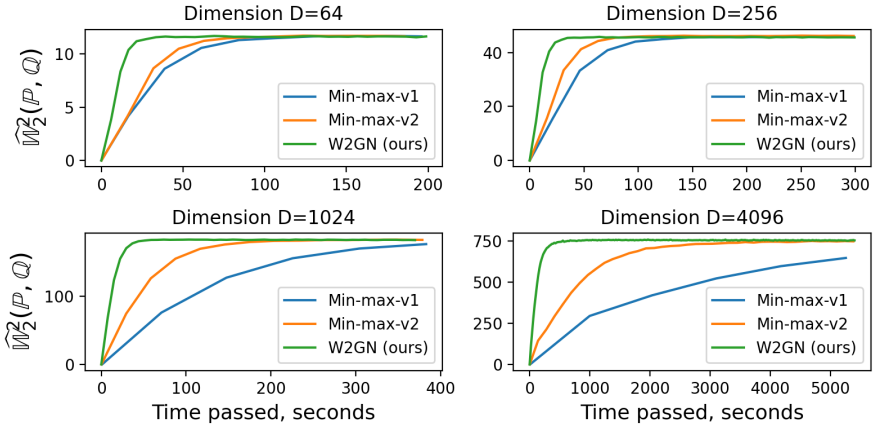
Metric:

$$\mathcal{L}^2\text{-UVP}(T) = 100 \cdot \frac{\|T - \nabla\psi^*\|^2_{\mathbb{P}}}{\text{Var}(\mathbb{Q})}\%$$

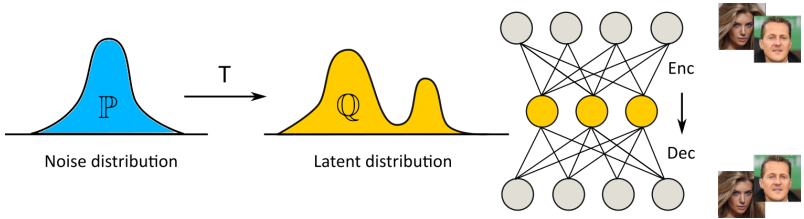
Dim	*2*	*4*	*8*	*16*	*32*	*64*	*128*	*256*	*512*	*1024*	*2048*	*4096*
LSOT	< 1	3.7	7.5	14.3	23	34.7	46.9	> 50				
MM-1	< 1	< 1	< 1	< 1	< 1	1.2	1.4	1.3	1.5	1.6	1.8	2.7
MM-2	< 1	< 1	< 1	< 1	< 1	< 1	1	1.1	1.2	1.3	1.5	2.1
W2GN (ours)	< 1	< 1	< 1	< 1	< 1	< 1	1	1.1	1.3	1.3	1.8	1.5

Experiments: Gaussian Optimal Transport

Comparison of convergence of MM-1, MM-2 and W2GN (ours) methods in dimensions $D = 64, 256, 1024, 4096$.



Experiments: Latent Space Optimal Transport

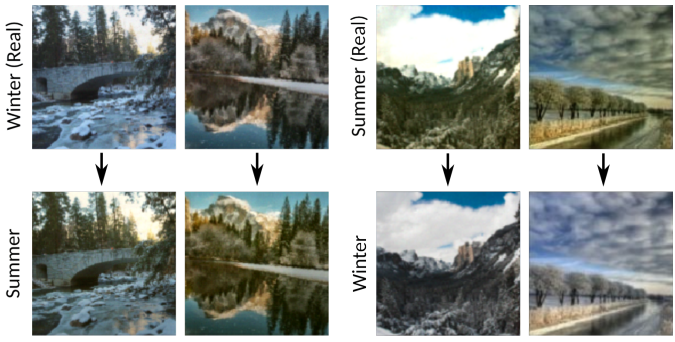


Method	FID
AE: $Dec(Enc(X))$	7.5
AE Raw Decode: $Dec(Z)$	31.81
W2GN+AE: $Dec(g^{\dagger}(Z))$	17.21
WGAN-QC : $Gen(Z)$	14.41

Experiments: Unpaired Image-to-image Style Transfer

128×128 image crops

Winter2SummerYosemite dataset[24]



[24] Jun-Yan Zhu et al. (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232.

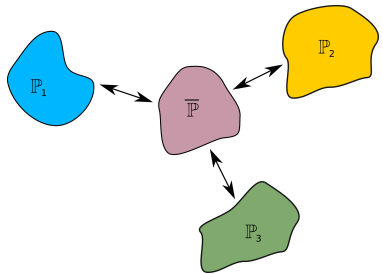
6. A Non-Minimax Algorithm for Continuous $\mathbb{W}_2$ Barycenters [Problem B]

Continuous Wasserstein-2 Barycenter Estimation without Minimax Optimization (ICLR 2021, A*)

https://openreview.net/forum?id=3tFAs5E-Pe

Alexander Korotin, Lingxiao Li, Justin Solomon, Evgeny Burnaev

A new algorithm to compute Wasserstein-2 barycenters powered by *input convex neural networks* and a straightforward optimization procedure without introducing bias.



Primal and Dual Formulations of the Barycenter Problem

Primal Form

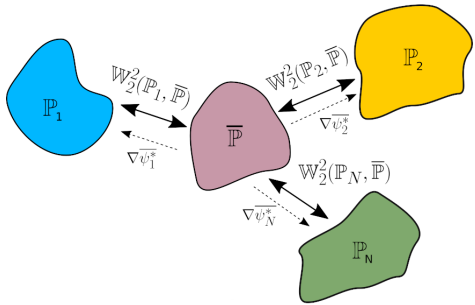
$$\overline{\mathbb{P}} = \underbrace{\arg\min_{\mathbb{P}\in\mathcal{P}_2(\mathbb{R}^D)} \sum_{n=1}^{N} \alpha_n \mathbb{W}_2^2(\overline{\mathbb{P}}, \mathbb{P}_n)}_{\mathcal{B}(\mathbb{P})}.$$

Dual form

$$\mathcal{B}(\overline{\mathbb{P}}) = \mathsf{Const} - \min_{\{\phi_n\}\ \text{congr.}} \underbrace{\sum_{n=1}^{N} \alpha_n \int_{\mathbb{R}^D} \psi_n(y) d\mathbb{P}_n(y)}_{\mathsf{MultiCorr}(\{\alpha_n,\mathbb{P}_n\}|\{\psi_n\})},$$

where **convex** functions ψ_n are **congruent**, i.e.,

$$\forall x \in \mathbb{R}^D: \qquad \sum_{n=1}^{N} \alpha_n \overline{\psi}_n(x) = \frac{\|x\|^2}{2}.$$



Primal-dual relation

$$\nabla\psi_n^* \sharp \mathbb{P}_n = \overline{\mathbb{P}}$$

allows to recover the barycenter $\overline{\mathbb{P}}$ from dual solutions ψ_n^*.

Proposed Barycenter Optimization Objective

Consider the following optimization ($\lambda, \tau > 0$) over $2N$ convex functions $\{\psi_n, \phi_n\}$:

$$\min_{\{\psi_n,\phi_n\}} \overbrace{\sum_{n=1}^{N}\left[\alpha_n \int_{\mathbb{R}^D}[\langle x, \nabla\psi_n(x)\rangle - \phi_n(\nabla\psi_n(x))]d\mathbb{P}_n(x)\right]}^{\text{Approximate multiple correlation}} + \underbrace{\tau\cdot\mathcal{R}_1^{\widehat{\mathbb{P}}}(\{\phi_n\})}_{\text{Congruence reg.}} + \underbrace{\lambda\sum_{n=1}^{N}\alpha_n\mathcal{R}_2^{\mathbb{P}_n}(\psi_n,\phi_n)}_{\text{Cycle regularizer}}.$$

Here $\mathcal{R}_2^{\mathbb{P}_n}(\psi_n, \phi_n)$ is the proposed cycle regularizer and $\mathcal{R}_1^{\widehat{\mathbb{P}}}(\{\phi_n\})$ is the novel proposed **congruence regularizer** (with the prior $\hat{\mathbb{P}}$):

$$\mathcal{R}_1^{\widehat{\mathbb{P}}}(\{\phi_n\}) = \int_{\mathbb{R}^D}\left[\sum_{n=1}^{N}\alpha_n\phi_n(y) - \frac{\|y\|^2}{2}\right]_+ d\widehat{\mathbb{P}}.$$

In practice, we optimize the objective by approximating $\{\psi_n, \phi_n\}$ with ICNNs whose parameters are trained by the stochastic gradient-descent on random batches from distributions $\mathbb{P}_n, \hat{\mathbb{P}}$.

Theorem (Primal-dual relation)

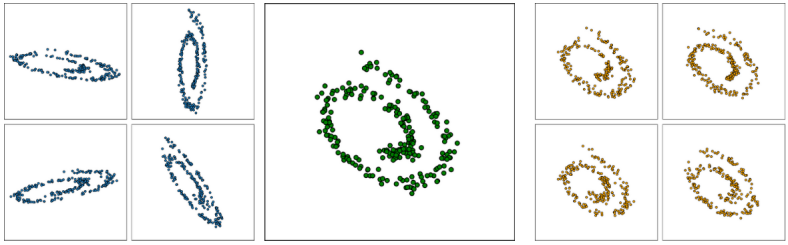
Let $\overline{\mathbb{P}} \in \mathcal{P}_{2,ac}(\mathbb{R}^D)$ *be the barycenter of* $\mathbb{P}_1, \dots, \mathbb{P}_N \in \mathcal{P}_{2,ac}(\mathbb{R}^D)$ *w.r.t. weights* $\alpha_1, \dots, \alpha_N$. *Let* $\{\psi_n^*\}$ *be the optimal congruent potentials of the barycenter problem. With mild assumptions on 2N convex functions* $\hat{\psi}_n, \hat{\phi}_n$, *parameters* $\tau \geq 1, \lambda > 0$ *and the prior distribution* $\hat{\mathbb{P}}$, *it holds that*

$$\epsilon = \mathit{MultiCorr}\big(\{\alpha_n, \mathbb{P}_n\} \mid \{\hat{\psi}_n\}, \{\hat{\phi}_n\}; \tau, \widehat{\mathbb{P}}, \lambda\big) - \mathit{MultiCorr}\big(\{\alpha_n, \mathbb{P}_n\} \mid \{\psi_n^*\}\big) \geq 0,$$

and for all $n \in \{1, \dots, N\}$, *we have*

$$\mathbb{W}_2^2\big(\nabla\hat{\psi}\sharp\mathbb{P}_n, \overline{\mathbb{P}}\big) \leq \|\nabla\hat{\psi} - \nabla\psi^*\|^2_{\mathcal{L}^2(\mathbb{P}_n)} \leq O(\epsilon).$$

Comparison in the Location-Scatter Case[25]



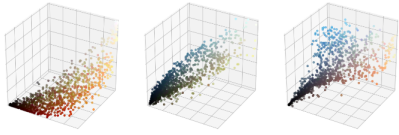
Metric	Method	D=2	4	8	16	32	64	128	256
$\mathbb{BW}_2^2$-UVP, %	[FC$\mathbb{W}$B], Cuturi and Doucet 2014	0.64	0.77	1.22	3.75	8.92	14.3	18.46	21.64
	[SC$\mathbb{W}_2$B], Fan, Taghvaei, and Chen 2021	0.12	0.10	0.19	0.29	0.46	0.6	1.38	2.9
$\mathcal{L}_2$-UVP, % (potentials)		0.17	0.12	0.2	0.31	0.47	0.62	1.21	1.52
	[CR$\mathbb{W}$B], L. 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

Table 1: Comparison of UVP metric for the location-scatter family (cube uniform), $N = 4$.

[25] Pedro C Álvarez-Esteban et al. (2016). "A fixed-point approach to barycenters in Wasserstein space". In: *Journal of Mathematical Analysis and Applications* 441.2, pp. 744–762.

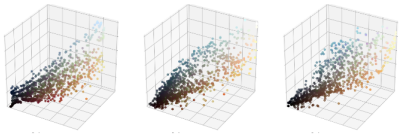
Application: Color Palette Averaging

(a) Original images $\{\mathcal{I}_n\}$.



(b) Color palettes $\{\mathbb{P}_n\}$ of original images.

(c) Images with averaged color palette $\{\nabla\psi_n^\dagger \sharp \mathcal{I}_n\}$.



(d) Barycenter palettes $\{\nabla\psi_n^\dagger \sharp \mathbb{P}_n\}$.

Figure 2: Results of our method applied to averaging color palettes of images.

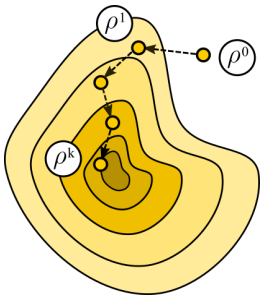
7. A Neural Implementation of the JKO scheme for $\mathbb{W}_2$ Gradient Flows [Problem C]

Large-Scale Wasserstein Gradient Flows (NeurIPS 2021, A*)

A neural algorithm to compute the Wasserstein-2 Gradient Flows via JKO scheme.

https://openreview.net/forum?id=nlLjIuHsMHp

Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, Evgeny Burnaev



An Algorithm to Compute The Wasserstein-2 Gradient Flow

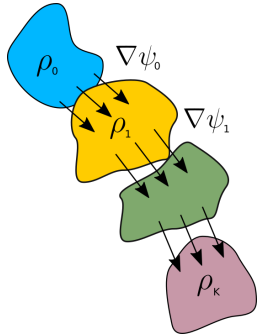
Recall the JKO gradient step:

$$\rho^{k+1} \leftarrow \underset{\rho \in \mathcal{P}_2(\mathbb{R}^D)}{\arg\min} \left[\mathcal{F}(\rho) + \frac{1}{2\tau} \mathbb{W}_2^2(\rho^k, \rho) \right].$$

We **replace** the optimization over distributions with the practically feasible optimization over convex functions ψ:

$$\psi_k \leftarrow \underset{\psi \in \text{Conv}}{\arg\min} \left[\mathcal{F}(\nabla\psi \sharp \rho^k) + \frac{1}{\tau} \int_{\mathbb{R}^D} \frac{1}{2} \|x - \nabla\psi(x)\|^2 d\rho^k(x) \right].$$

The approach yields $\rho^k = \nabla\psi_{k-1} \sharp \left[\nabla\psi_{k-2} \sharp \left[\dots \nabla\psi_0 \sharp \rho^0 \right] \right]$.



In practice, to compute the JKO steps, we learn a successively sequence of ICNNs as $\{\psi_k\}$ and optimize their parameters via the stochastic gradient descent by using random batches from ρ^k.

Connection between Wasserstein gradient flows and SDEs

Consider an $\mathbb{R}^D$-valued stochastic process $\{X_t\}_{t\geq 0}$, governed by the following Ito SDE:

$$dX_t = -\nabla V(X_t)dt + \sqrt{2\beta^{-1}}dW_t, \quad \text{s.t. } X_0 \sim \rho^0,$$

where $V : \mathbb{R}^D \to \mathbb{R}$ is the potential function, W_t is the standard Wiener process

The marginal measure ρ_t of X_t satisfies the Fokker-Planck equation

$$\frac{\partial \rho_t}{\partial t} = \operatorname{div}\left(\nabla V(x)\rho_t\right) + \beta^{-1}\Delta\rho_t, \text{ s.t. } \rho_0 = \rho^0$$

This equation is the Wasserstein gradient flow for

$$\mathcal{F}_{FP}(\rho) = \int_{\mathbb{R}^D} V(x)dx + \beta^{-1}\int_{\mathbb{R}^D} \log\frac{d\rho}{dx}d\rho(x)$$

Theoretical Results - Estimator for the Fokker-Plank Functional

Consider the Fokker-Plank free energy functional $\mathcal{F} = \mathcal{F}_{\text{FP}}$:

$$\mathcal{F}_{FP}(\rho) \stackrel{def}{=} \int_{\mathbb{R}^D} V(x) d\rho(x) + \int_{\mathbb{R}^D} \log \frac{d\rho(x)}{dx} d\rho(x).$$

How to estimate $\mathcal{F}(\underbrace{\nabla\psi}_{T} \sharp \rho)$?

Theorem (Stochastic Estimator of $\mathcal{F}_{\text{FP}}$)

Let $\rho \in \mathcal{P}_{2,ac}(\mathbb{R}^D)$ and $T : \mathbb{R}^D \to \mathbb{R}^D$ be a diffeomorphism. For a random batch $x_1, \dots, x_N \sim \rho$, the expression

$$\frac{1}{N}\sum_{n=1}^{N} V\big(T(x_n)\big) - \frac{1}{N}\sum_{n=1}^{N} \log |\det \nabla T(x_n)|,$$

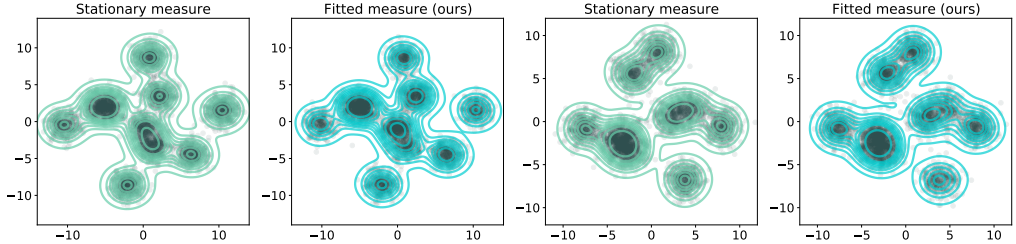
is a consistent estimator of $\mathcal{F}_{FP}(T\sharp\rho)$ up to constant (w.r.t. T) shift.

Evaluation: Convergence to the Stationary Distribution

Starting from an arbitrary initial ρ^0, the gradient flow of $\mathcal{F}_{\text{FP}}$ converges to[26]

$$\frac{d\rho^*}{dx}(x) = Z^{-1} \exp(-V(x)),$$

where $Z = \int_{\mathbb{R}^D} \exp(-V(x))dx$ is the normalization constant.



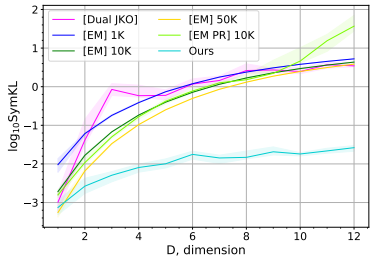
PCA components of the true stationary and the learned distributions in $D = 13$ (left) and $D = 32$ (right).

[26] Hannes. Risken (1996). *The Fokker-Planck Equation: Methods of Solution and Applications (Springer Series in Synergetics)*.. Springer,

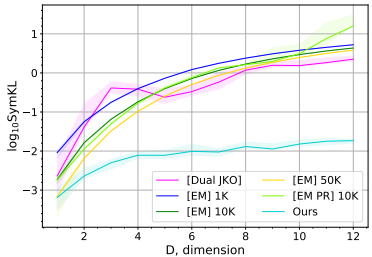
Ornstein-Uhlenbeck processes

When $V(x) = \frac{1}{2}(x-b)^T A(x-b)$ with $A \succeq 0$ and the initial distribution ρ^0 is Gaussian, the gradient flow of $\mathcal{F}_{\text{FP}}$ admits closed form at every time point $t \geq 0$.[27]

We compare our recovered flow with the ground truth by using SymKL metric.



$t = 0.5$



$t = 0.8$

[27] Pat Vatiwutipong and Nattakorn Phewchean (2019). "Alternative way to derive the distribution ofthe multivariate Ornstein-Uhlenbeck process". In: *Advances in Difference Equations* 276.

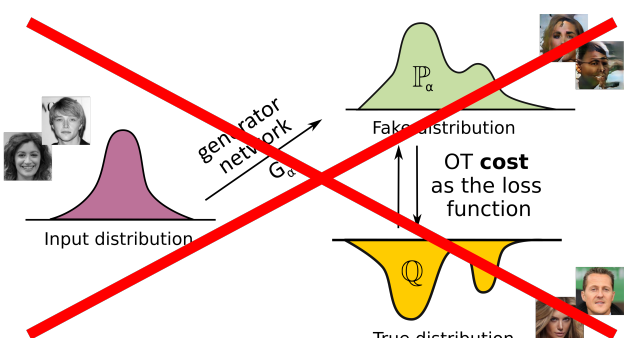
8. Optimal Transport Modeling

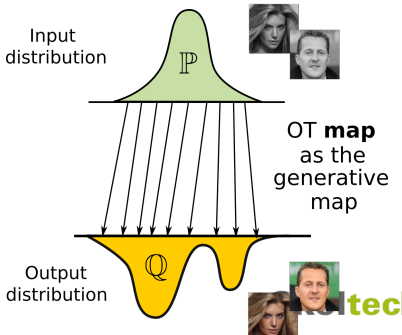
Generative Modeling with Optimal Transport Maps (ICLR 2022, A*)

https://arxiv.org/abs/2110.02999

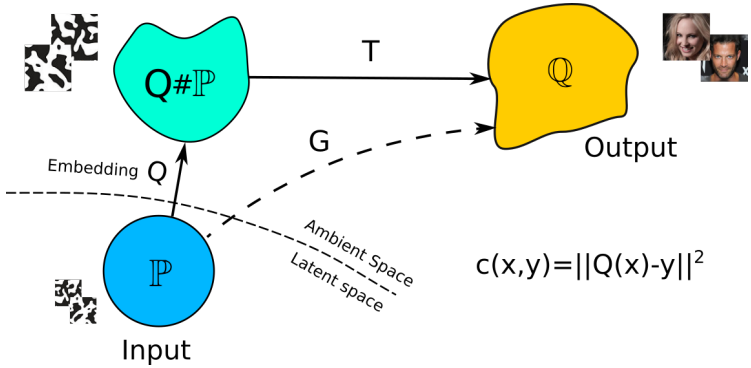
Litu Rout, Alexander Korotin, Evgeny Burnaev

While the optimal transport cost serves as the loss for popular generative models, we demonstrate that the optimal transport map can be used as the generative model itself.





Optimal Transport Modeling



$$\mathcal{L}(\psi, G) = \min_{\psi} \max_{G} \left\{ \int \left\{ \langle Q(x), G(x) \rangle - \psi\big(G(x)\big) \right\} d\mathbb{P}(x) + \int \psi(y) d\mathbb{Q}(y) \right\}$$

In the optimal pair (ψ^*, G^*),

G^* is the map from $\mathbb{P}$ to $\mathbb{Q}$ for the Q-embedded quadratic cost.

Theoretical Guarantees

$$\mathcal{L}(\psi, G) = \min_{\psi} \max_{G} \left\{ \int \left\{ \langle Q(x), G(x) \rangle - \psi\big(G(x)\big) \right\} d\mathbb{P}(x) + \int \psi(y) d\mathbb{Q}(y) \right\}$$

Theorem [Non detailed]. For an approximate solution $(\hat{\psi}, \hat{G})$ define

$$\epsilon_1 = \max_{G} \mathcal{L}(\hat{\psi}, G) - \mathcal{L}(\hat{\psi}, \hat{G}) \quad \text{and} \quad \epsilon_2 = \max_{G} \mathcal{L}(\hat{\psi}, G) - \min_{\psi} \max_{G} \mathcal{L}(\psi, G)$$

Then the following bound holds true for the OT map G^* from $\mathbb{P}$ to $\mathbb{Q}$:

$$\frac{\text{FID}(\hat{G}_{\#}\mathbb{P}, \mathbb{Q})}{L^2} \leq 2 \cdot \mathcal{W}_2^2(\hat{G}_{\#}\mathbb{P}, \mathbb{Q}) \leq \int \|\hat{G}(x) - G^*(x)\|^2 d\mathbb{P}(x) \leq O(\epsilon_1 + \epsilon_2),$$

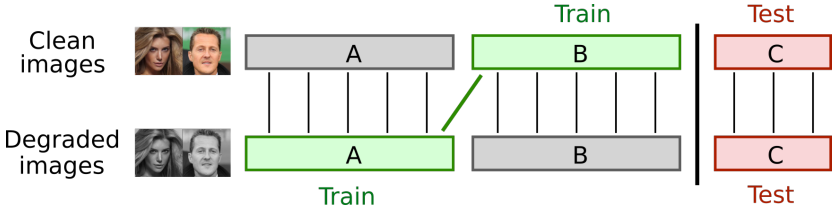
where L is the Lipschitz constant of InceptionV3[28].

[28] Christian Szegedy et al. (2016). "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826.

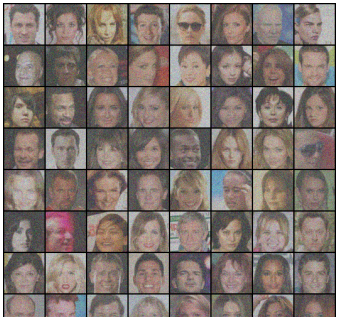
Celeba Faces Generation

Model	FID ↓
DCGAN	52.0
DRAGAN	42.3
BEGAN	38.9
NVAE	13.4
NCP-VAE	5.2
WGAN	41.3
WGAN-GP	30.0
WGAN-QC	12.9
AE-OT	28.6
W2GN+AE	17.2
AE-OT-GAN	7.8
OTM (Ours)	7.7

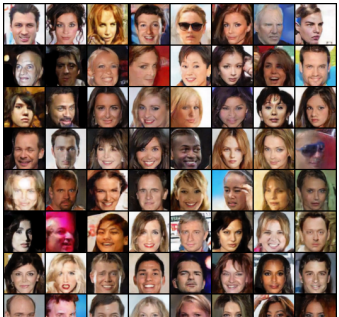
Unpaired Image Denoising



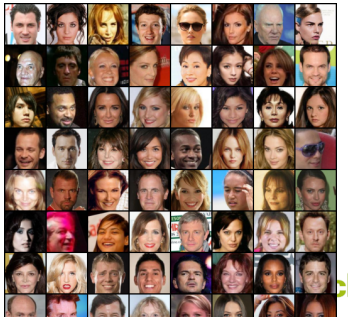
Test C (degraded)



Test C (mapped)



Test C (clean)



Extension to Arbitrary Transport Costs $c(\cdot, \cdot)$

(1) Solver for the **quadratic cost** $c(x, y) = \frac{\|x-y\|^2}{2} \Leftrightarrow c(x, y) = -\langle x, y \rangle$:

$$\mathcal{L}(\psi, T) = \min_{\psi} \max_{T} \left\{ \int \left\{ \langle x, T(x) \rangle - \psi\big(T(x)\big) \right\} d\mathbb{P}(x) + \int \psi(y) d\mathbb{Q}(y) \right\}$$

(2) Solver for the **arbitrary cost** $c(x, y)$:

$$\mathcal{L}(f, T) = \max_{f} \min_{T} \left\{ \int \left\{ c(x, T(x)) - f\big(T(x)\big) \right\} d\mathbb{P}(x) + \int f(y) d\mathbb{Q}(y) \right\}$$

Example: $c(x, y) = dist\big(\text{color\_palette}(x), \text{color\_palette}(y)\big)$

[29] Guansong Lu et al. (2019). "Guiding the one-to-one mapping in cyclegan via optimal transport". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 4432–4439.

Issues, open questions

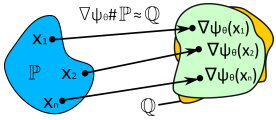
1. Algorithmic issues (solved in our recent papers!)
   - Not all solutions T^* are OT maps
   - Algorithm searches only for dereministic solutions (maps, not plans)
     - Application: one-to-many problems, such as image colorization.
2. Statistical issues
   - Absolutely no results
3. Approximation issues (with neural networks)
4. Optimization issues
   - This min-max problem has not been studied yet
   - Disconverges near the optimum
5. Extensions
   - Multi-marginal problems, variational problems (barycenters), etc.
   - Gradient Flows

9. Summary: Publications, Presentations

Summary: 4 OT Problems = 4 Solutions

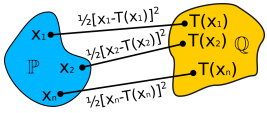
[Problem 0]

Benchmarking continuous OT methods



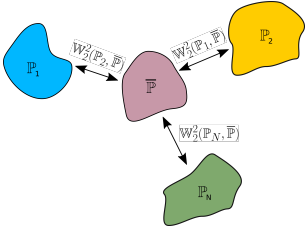
[Problem A]

Computation of $\mathbb{W}_2$ OT maps



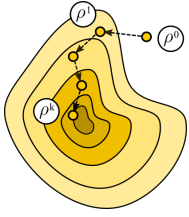
[Problem B]

Computation of $\mathbb{W}_2$ barycenters



[Problem C]

Computation of $\mathbb{W}_2$ gradient flows



References

[1] Alexander Korotin, Vage Egiazarian, Arip Asadulaev, et al. (2021). "Wasserstein-2 Generative Networks". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=bEoxzW_EXsa **[ICLR 2021]**

[2] Alexander Korotin, Lingxiao Li, Justin Solomon, et al. (2021). "Continuous Wasserstein-2 Barycenter Estimation without Minimax Optimization". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=3tFAs5E-Pe**[ICLR 2021]**

[3] Alexander Korotin, Lingxiao Li, Aude Genevay, et al. (2021). "Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark". In: *Advances in Neural Information Processing Systems* 34, pp. 14593–14605 **[NeurIPS 2021]**

[4] Petr Mokrov et al. (2021). "Large-scale wasserstein gradient flows". In: *Advances in Neural Information Processing Systems* 34, pp. 15243–15256 **[NeurIPS 2021]**

[5] Litu Rout, Alexander Korotin, and Evgeny Burnaev (2021). "Generative Modeling with Optimal Transport Maps". In: *International Conference on Learning Representations* **[ICLR 2022]**

References

[6] Alexander Korotin, Alexander Kolesov, and Evgeny Burnaev (2022). "Kantorovich Strikes Back! Wasserstein GANs are not Optimal Transport?". In: *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. URL: https://openreview.net/forum?id=VtEEpi-dGlt **[NeurIPS 2022]**

[7] Alexander Korotin, Vage Egiazarian, Lingxiao Li, et al. (2022). "Wasserstein iterative networks for barycenter estimation". In: *arXiv preprint arXiv:2201.12245* **[NeurIPS 2022]**

[8] Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev (2022b). "Neural optimal transport". In: *arXiv preprint arXiv:2201.12220*

[9] Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev (2022a). "Kernel Neural Optimal Transport". In: *arXiv preprint arXiv:2205.15269*

[10] Milena Gazdieva et al. (2022). "Unpaired image super-resolution with optimal transport maps". In: *arXiv preprint arXiv:2202.01116*

[11] Arip Asadulaev et al. (2022). "Neural Optimal Transport with General Cost Functionals". In: *arXiv preprint arXiv:2205.15403*