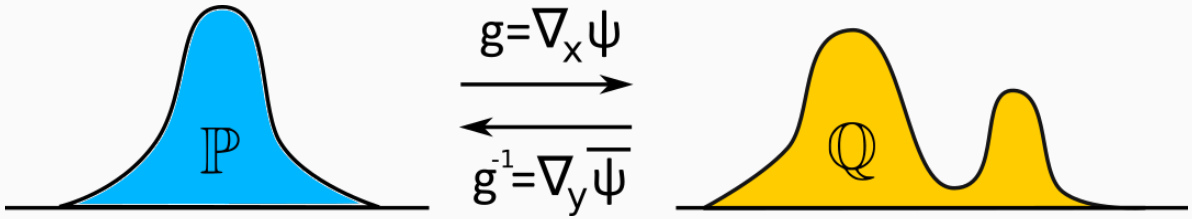
Wasserstein-2 Generative Networks

Alexander Korotin[1] & Vage Egiazarian[1] & Arip Asadulaev[2]
Alexander Safin[1] & Evgeny Burnaev[1]

[1]Skolkovo Institute of Science and Technology (Moscow, Russia)
[2]Information Technologies, Mechanics and Optics University (Saint Petersburg, Russia)

Wasserstein-2 Generative Networks



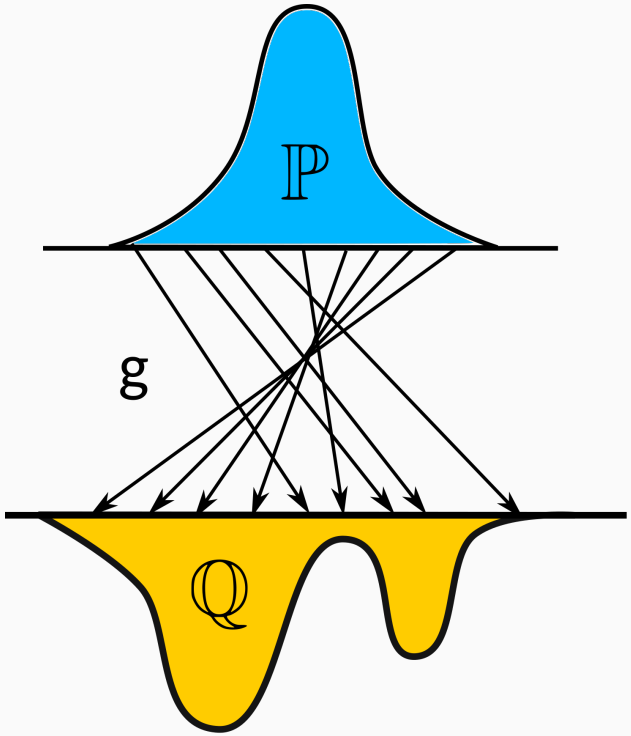
Alexander Korotin et al. (2019). "Wasserstein-2 Generative Networks". In: *arXiv preprint arXiv:1909.13082*

`github.com/iamalexkorotin/Wasserstein2GenerativeNetworks`

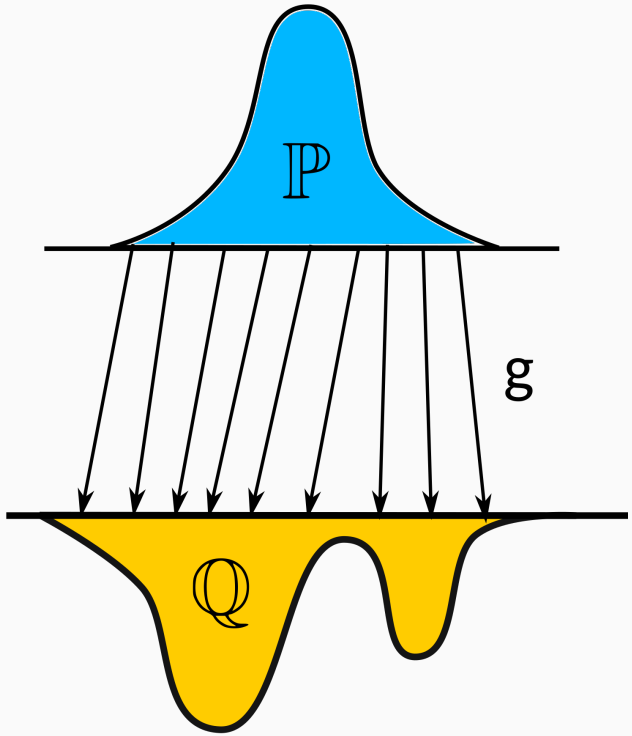
Table of Contents

Preliminaries: Well-structured Mappings

Optimal Mappings?



Not optimal



Optimal

What is **optimal** generative mapping?

Optimal = Simple, Intuitive, Not Overparametrized

Cyclically Monotone Mapping

Assume $\mathcal{X} = \mathcal{Y} = \mathbb{R}^D$.

- **Monotonicity** for $D = 1$: for all $x, x' \in \mathcal{X}$

$$\big(g(x) - g(x')\big) \cdot \big(x - x'\big) \geq 0;$$

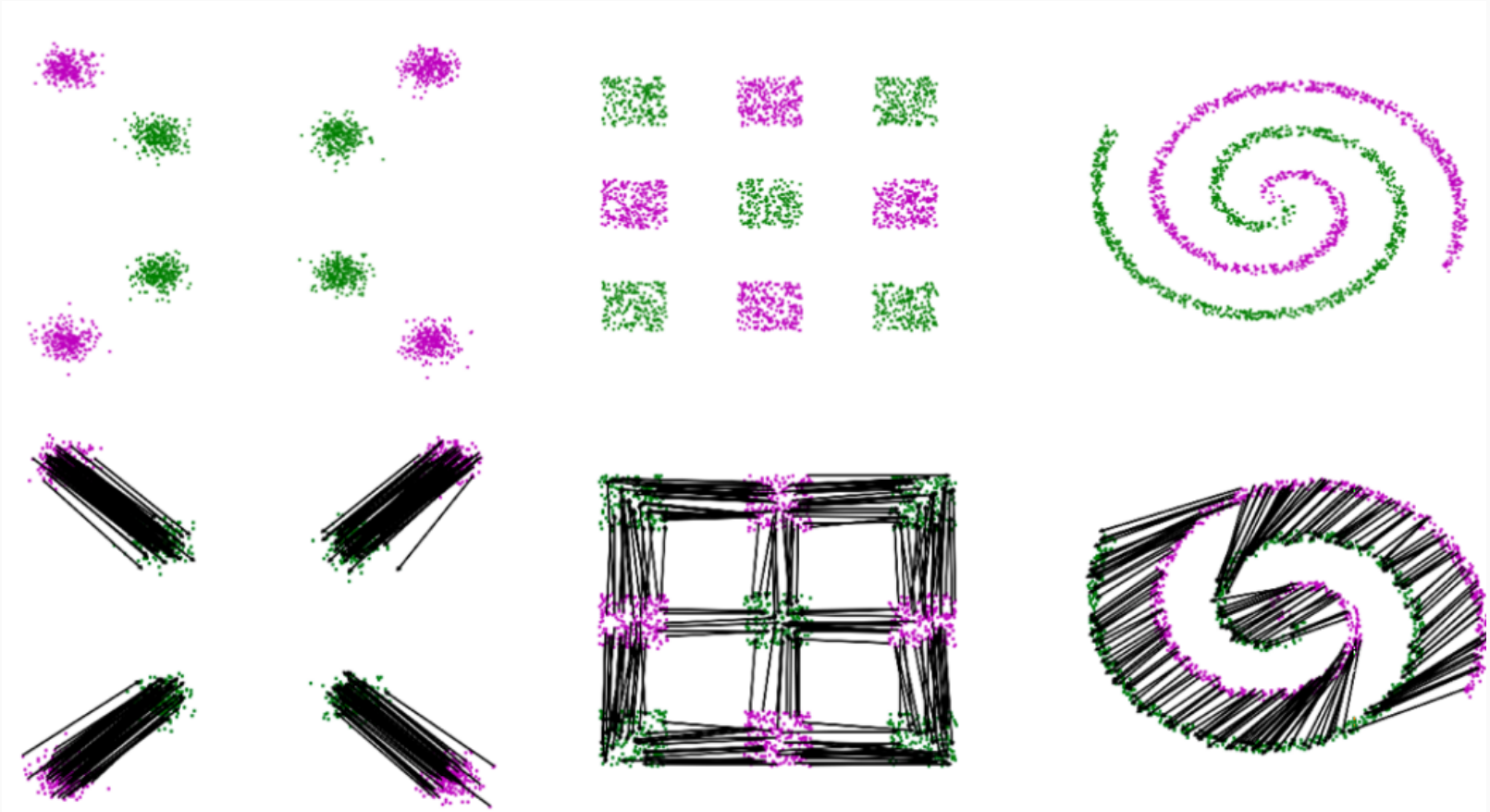
- **Cycle Monotonity**[1] for $D \geq 1$: for all distinct $x_1, \ldots x_N \in \mathcal{X}$

$$\sum_{n=1}^{N} \langle g(x_n), x_n - x_{n+1} \rangle \geq 0.$$

Well-structured and (usually) **invertible** generative mapping g!

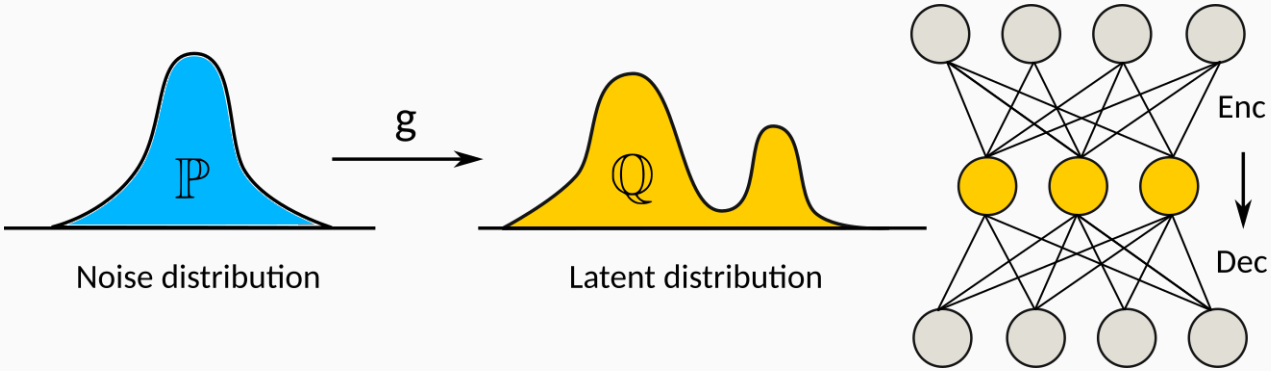
[1] Cédric Villani (2008). *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media.

Cycle Monotone Maps: 2D Examples

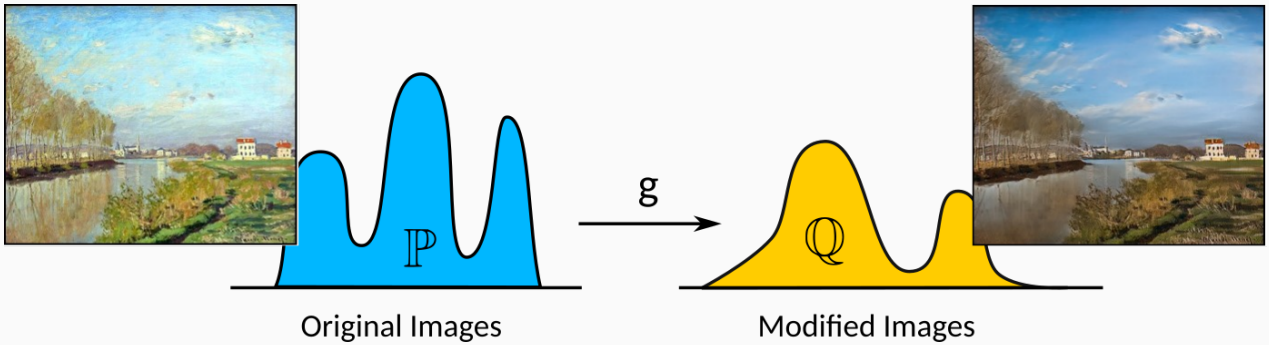


When $\mathcal{X} = \mathcal{Y}$?

- Latent Space Generative Mapping



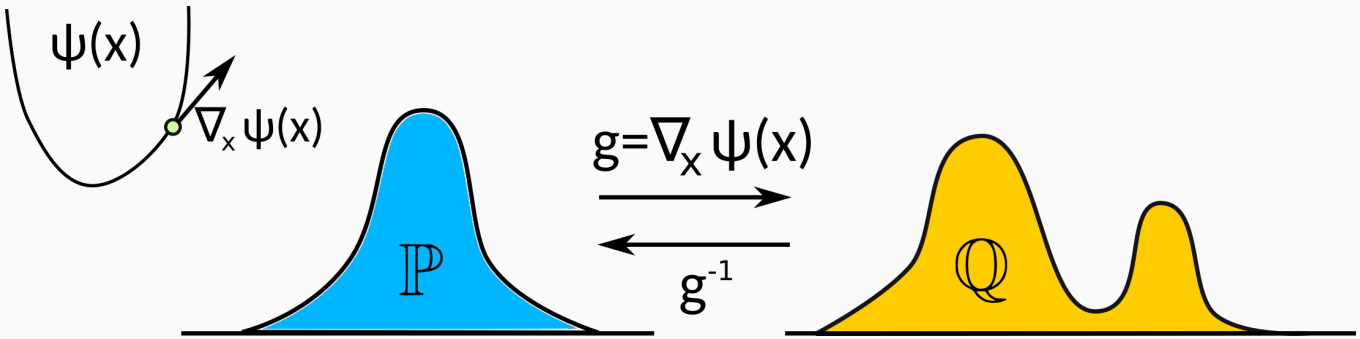
- Image-To-Image Translation



Existince and Uniqueness

Theorem[2][3]. Consider distributions $\mathbb{P}, \mathbb{Q}$ on $\mathcal{X} = \mathcal{Y} = \mathbb{R}^D$. If $\mathbb{P}$ does not give mass to small sets, then cyclically monotone generator $g \circ \mathbb{P} = \mathbb{Q}$

- Exists and is Unique [up to the values outside of Supp $\mathbb{P}$]!
- Is a gradient $g = \nabla_x \psi(x)$ of a convex function $\psi : \mathbb{R}^D \to \mathbb{R}$!

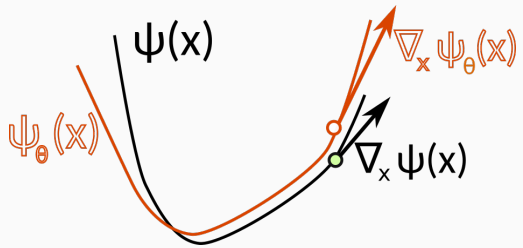


[2]Yann Brenier (1991). "Polar factorization and monotone rearrangement of vector-valued functions". In: *Communications on pure and applied mathematics* 44.4, pp. 375–417.

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

The Idea

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 (ICNN);
- $g_\theta = \nabla_x \psi_\theta : \mathbb{R}^D \to \mathbb{R}^D$ - generative mapping.

Questions

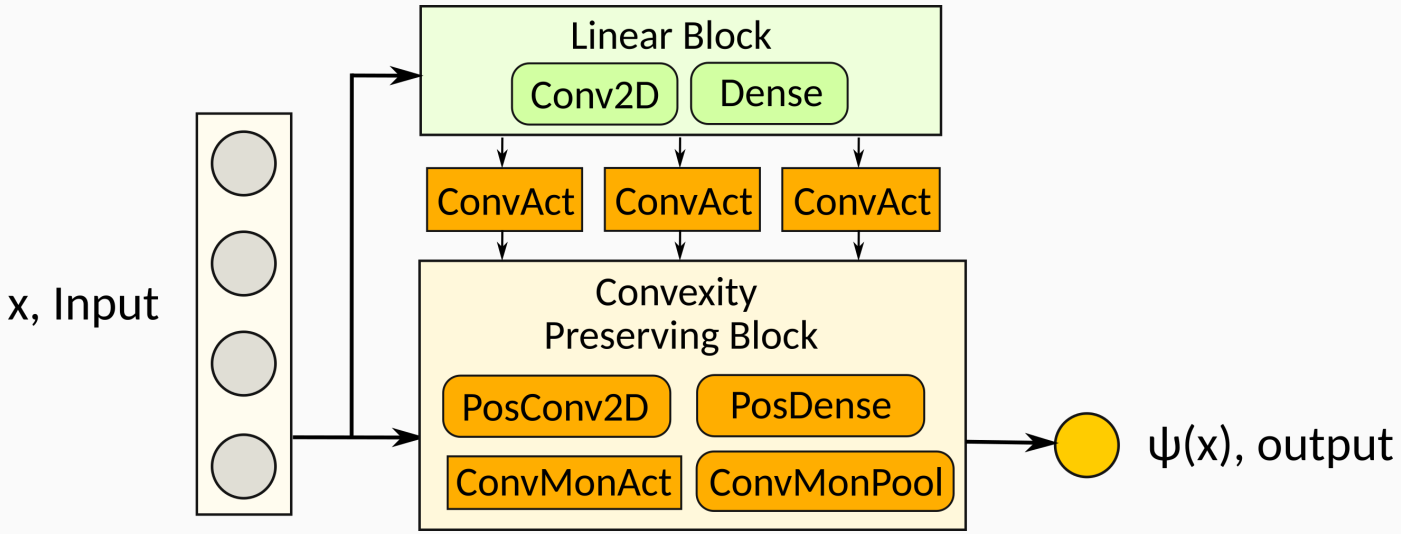
1. How to approximate convex functions by nets? - next slide
2. How to find the generator? - GANs, normalizing flows[4], etc.

$$\min_\theta \Big[\text{Loss}\big(\nabla\psi_\theta \circ \mathbb{P}, \mathbb{Q}\big)\Big]$$

[4] Chin-Wei Huang et al. (2020). *Convex Potential Flows: Universal Probability Distributions with Optimal Transport and Convex Optimization*. arXiv: 2012.05942 [cs.LG].

Input Convex Neural Networks

Based on the sequential variant for ICNN[5] $\psi_\theta(x) : \mathbb{R}^D \to \mathbb{R}$.



Inside CP Block: only positive weights (except biases) in linear layers and only convex & monotone activations (e.g. ReLU, CELU).

Computational complexity of $\nabla_x \psi_\theta(x)$ comparable to that of $\psi_\theta(x)$!

[5] 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.

Optimal Transport with the Quadratic Cost

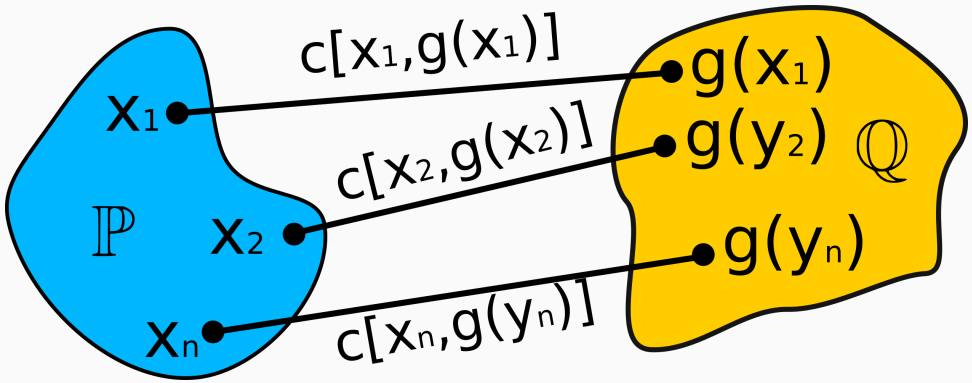
Wasserstein-2 Distance

The cycle monotone mapping

$$g^* = \underset{g \circ \mathbb{P} = \mathbb{Q}}{\operatorname{argmin}} \int_{\mathcal{X}} \frac{\|x - g(x)\|^2}{2} d\mathbb{P}(x)$$

attains the optimal quadratic transport cost (**Wasserstein-2 distance**)

$$\mathbb{W}_2^2(\mathbb{P}, \mathbb{Q}) = \min_{g \circ \mathbb{P} = \mathbb{Q}} \int_{\mathcal{X}} \frac{\|x - g(x)\|^2}{2} d\mathbb{P}(x)$$



Dual Form of Wasserstein-2 distance

$$\mathbb{W}_2^2(\mathbb{P},\mathbb{Q}) = -\overbrace{\min_{\psi\in\text{Conv}}\left[\int_{\mathcal{X}}\psi(x)d\mathbb{P}(x)+\int_{\mathcal{Y}}\overline{\psi}(y)d\mathbb{Q}(y)\right]}^{\text{Corr}(\mathbb{P},\mathbb{Q})}+\text{Const}(\mathbb{P},\mathbb{Q})$$

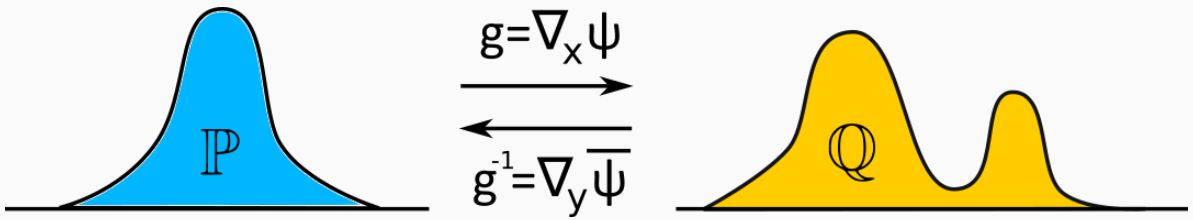
$$\overline{\psi}(y) = \max_x\left(\langle x,y\rangle - \psi(x)\right)$$

Important! (Properties of convex conjugates)

$$\overline{\psi}(y) = \langle\nabla\psi^{-1}(y),y\rangle - \psi\big((\nabla\psi)^{-1}(y)\big) \qquad \text{and} \qquad (\nabla\psi)^{-1} = \nabla\overline{\psi}$$

Very Important! (Properties of the optimal potentials)

Optimal Convex **Potential** ψ^* $\Leftrightarrow$ Optimal **Generator** $g^* = \nabla_x\psi^*(x)$



Optimization of Correlation: Prior Art

Idea: optimize correlation to obtain the cyclically monotone map!

$$\mathrm{Corr}(\mathbb{P},\mathbb{Q}) = \min_{\psi\in\mathrm{Conv}} \left[\int_{\mathcal{X}} \psi(x)d\mathbb{P}(x) + \int_{\mathcal{Y}} \overline{\psi}(y)d\mathbb{Q}(y)\right]$$

Use neural network ψ_θ with weights θ to approximate ψ:

$$\min_\theta \mathrm{Corr}(\mathbb{P},\mathbb{Q}|\psi_\theta) = \min_\theta \left[\int_{\mathcal{X}} \psi_\theta(x)d\mathbb{P}(x) + \int_{\mathcal{Y}} \overline{\psi_\theta}(y)d\mathbb{Q}(y)\right]$$

Problem: how to back-propagate through conjugate function $\overline{\psi_\theta}(y)$?

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

Solution 1: Entropy or Quadratic Regularization[6] [LSOT]

Assume that ψ, ϕ are convex conjugates, i.e.

$$\psi(x) = \max_y \left[\langle x, y\rangle - \phi(y)]\right] \qquad \text{and} \qquad \phi(y) = \max_x \left[\langle x, y\rangle - \psi(x)\right]$$

We can derive the necessary condition for conjugacy of two functions:

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y} : \qquad \psi(x) + \phi(y) \geq \langle x, y\rangle.$$

Use two networks ψ_θ, ϕ_ω. Penalize potentials for being non-conjugate!

$$\min_{\theta,\omega} \left(\left[\int_{\mathcal{X}} \psi_\theta(x) d\mathbb{P}(x) + \int_{\mathcal{Y}} \phi_\omega(y) d\mathbb{Q}(y) \right] + \mathcal{R}_{\mathbb{P},\mathbb{Q}}(\psi_\theta, \phi_\omega) \right)$$

[6] Vivien Seguy et al. (2017). "Large-scale optimal transport and mapping estimation". In: *arXiv preprint arXiv:1711.02283*.

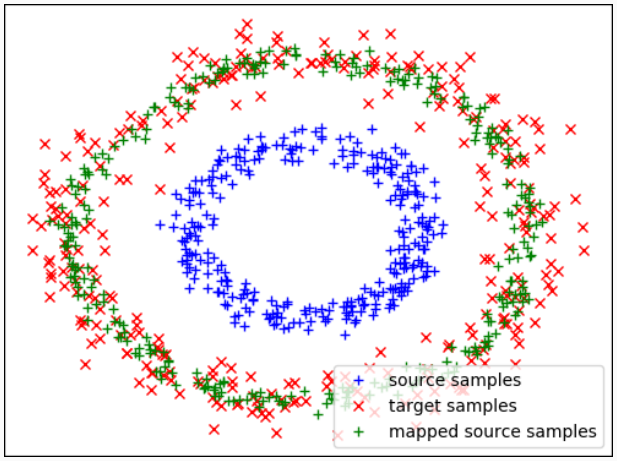
Solution 1: Entropy or Quadratic Regularization [LSOT]

Entropic Regularizer $(\epsilon \geq 0)$

$$\mathcal{R}_{\mathbb{P},\mathbb{Q}}(\psi_\theta, \phi_\omega) = \epsilon \int_{\mathcal{X}\times\mathcal{Y}} \exp \frac{\langle x, y\rangle - \psi_\theta(x) - \phi_\omega(y)}{\epsilon} d\big(\mathbb{P}\times\mathbb{Q}\big)$$

Quadratic Regularizer $(\epsilon \geq 0)$

$$\mathcal{R}_{\mathbb{P},\mathbb{Q}}(\psi_\theta, \phi_\omega) = \frac{1}{4\epsilon} \int_{\mathcal{X}\times\mathcal{Y}} \big(\langle x, y\rangle - \psi_\theta(x) - \phi_\omega(y)\big)_+^2 d\big(\mathbb{P}\times\mathbb{Q}\big)$$



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

Solution 2: Explicit Backpropagation [MM-1]

$$\min_{\theta} \mathrm{Corr}(\mathbb{P}, \mathbb{Q}|\psi_\theta) = \min_{\theta} \left[\int_{\mathcal{X}} \psi_\theta(x) d\mathbb{P}(x) + \int_{\mathcal{Y}} \overline{\psi_\theta}(y) d\mathbb{Q}(y) \right]$$

Lemma (a variant of[7])

The derivative of $Corr(\mathbb{P}, \mathbb{Q} \mid \psi_\theta)$ *w.r.t.* θ *is given by*

$$\frac{\partial Corr(\mathbb{P}, \mathbb{Q}|\psi_\theta)}{\partial \theta} = \int_{\mathcal{X}} \frac{\partial \psi_\theta(x)}{\partial \theta} d\mathbb{P}(x) - \int_{\mathcal{Y}} \frac{\partial \psi_\theta(\hat{x})}{\partial \theta} d\mathbb{Q}(y),$$

where $\hat{x}$ *satisfies* $y = \nabla \psi_\theta(\hat{x})$ *or* $\hat{x} = \nabla \overline{\psi_\theta}(y)$*, equivalently,*

$$\hat{x} = \arg\max_{x} \left(\langle x, y \rangle - \psi_\theta(x) \right).$$

Additional inner (but convex) optimization subproblem appears!

[7] Amirhossein Taghvaei and Amin Jalali (2019). "2-Wasserstein Approximation via Restricted Convex Potentials with Application to Improved Training for GANs". In: *arXiv preprint arXiv:1902.07197*.

Solution 3: Min-max Approach[8][9] [MM-2]

$$\min_\theta \text{Corr}(\mathbb{P}, \mathbb{Q}|\psi_\theta) = \min_\theta \left[\int_{\mathcal{X}} \psi_\theta(x) d\mathbb{P}(x) + \int_{\mathcal{Y}} \overline{\psi_\theta}(y) d\mathbb{Q}(y) \right] =$$

$$\min_\theta \left[\int_{\mathcal{X}} \psi_\theta(x) d\mathbb{P}(x) + \int_{\mathcal{Y}} \max_x \left(\langle x, y \rangle - \psi_\theta(x) \right) d\mathbb{Q}(y) \right] =$$

$$\min_\theta \left[\int_{\mathcal{X}} \psi_\theta(x) d\mathbb{P}(x) + \max_{T \in \mathcal{Y}^{\mathcal{X}}} \int_{\mathcal{Y}} \left[\langle T(y), y \rangle - \psi(T(y)) \right] d\mathbb{Q}(y) \right]$$

Remark: For each θ, we have $T^* = \nabla \overline{\psi_\theta}$!

Idea 1: approximate functions T by neural networks $T_\omega : \mathbb{R}^D \to \mathbb{R}^D$!

Idea 2: define T as the gradient of of ICNN $\nabla \overline{\psi_\omega}$, i.e. $T_\omega := \nabla \overline{\psi_\omega}$!

---

[8] Quan Hoang Nhan Dam et al. (2019). "Threeplayer wasserstein gan via amortised duality". In: *Proc. of the 28th Int. Joint Conf. on Artificial Intelligence (IJCAI)*.

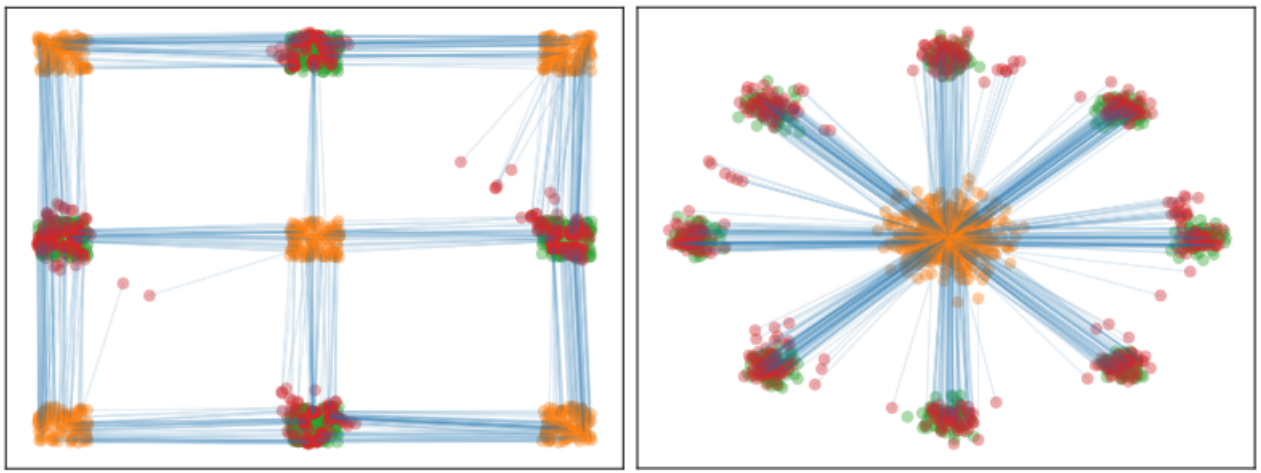
[9] Ashok Makkuva et al. (2020). "Optimal transport mapping via input convex neural networks". In: *International Conference on Machine Learning*. PMLR, pp. 6672–6681.

Solution 3: Min-max Approach [MM-2]

Use two ICNNs $\psi_\theta, \overline{\psi_\omega}$ and solve the following min-max problem:

$$\min_\theta \max_\omega \mathrm{Corr}(\mathbb{P}, \mathbb{Q} \mid \psi_\theta, \overline{\psi_\omega}) =$$

$$\min_\theta \left[\int_{\mathcal{X}} \psi_\theta(x) d\mathbb{P}(x) + \max_\omega \int_{\mathcal{Y}} \left[\langle \nabla \overline{\psi_\omega}(y), y \rangle - \psi_\theta(\nabla \overline{\psi_\omega}(y)) \right] d\mathbb{Q}(y) \right]$$



Wasserstein-2 Generative Networks

Removing Minimaxity from the Optimization

$$\min_{\theta}\left[\int_{\mathcal{X}}\psi_\theta(x)d\mathbb{P}(x)+\max_{\omega}\int_{\mathcal{Y}}\left[\langle\nabla\overline{\psi_\omega}(y),y\rangle-\psi_\theta(\nabla\overline{\psi_\omega}(y))\right]d\mathbb{Q}(y)\right]$$

Idea (revisited): Penalize potentials for being non-conjugate!

- The necessary condition for ψ,ϕ to be conjugate:

$$\forall(x,y)\in\mathcal{X}\times\mathcal{Y}:\qquad\psi(x)+\phi(y)\geq\langle x,y\rangle.$$

The condition requires considering pairs from $(x,y)\in\mathcal{X}\times\mathcal{Y}$!

- More feasible necessary condition for ψ,ϕ to be conjugate:

$$\nabla\psi=(\nabla\phi)^{-1}\qquad\Leftrightarrow\qquad\nabla\psi\circ\nabla\phi=\mathrm{id}_{\mathcal{Y}}$$

Cycle-Consistency Regularization

BEFORE: Min-max optimization

$$\min_{\theta} \max_{\omega} \mathrm{Corr}(\mathbb{P}, \mathbb{Q} \mid \psi_\theta, \overline{\psi_\omega}) =$$

$$\min_{\theta} \left[\int_{\mathcal{X}} \psi_\theta(x) d\mathbb{P}(x) + \max_{\omega} \int_{\mathcal{Y}} \left[\langle \nabla \overline{\psi_\omega}(y), y \rangle - \psi_\theta(\nabla \overline{\psi_\omega}(y)) \right] d\mathbb{Q}(y) \right]$$

AFTER: Non-minimax optimization (our proposal)

$$\min_{\theta,\omega} \mathrm{Corr}(\mathbb{P}, \mathbb{Q} \mid \psi_\theta, \overline{\psi_\omega}; \lambda) =$$

$$\min_{\theta,\omega} \left[\left(\int_{\mathcal{X}} \psi_\theta(x) d\mathbb{P}(x) + \int_{\mathcal{Y}} \left[\langle \nabla \overline{\psi_\omega}(y), y \rangle - \psi_\theta(\nabla \overline{\psi_\omega}(y)) \right] d\mathbb{Q}(y) \right) + \frac{\lambda}{2} \underbrace{\int_{\mathcal{Y}} \| \nabla \psi_\theta \circ \nabla \overline{\psi_\omega}(y) - y \|^2 d\mathbb{Q}(y)}_{\text{Cycle-consistency regularizer } \mathcal{R}_{\mathcal{Y}}(\theta,\omega)} \right].$$

The Algorithm

Algorithm 1: Numerical Procedure for Optimizing Correlations

Input: Distributions $\mathbb{P}, \mathbb{Q}$ with sample access; cycle-consistency regularizer coefficient $\lambda > 0$;

a pair of input-convex neural networks ψ_θ and $\overline{\psi_\omega}$; batch size $K > 0$;

for $t = 1, 2, \ldots$ **do**

1. Sample batches $X \sim \mathbb{P}$ and $Y \sim \mathbb{Q}$;
2. Compute the Monte-Carlo estimate of the correlations:
$$\mathcal{L}_{\text{Corr}} = \frac{1}{K}\left[\sum_{x \in X} \psi_\theta(x) + \sum_{y \in Y}\left[\langle \nabla\overline{\psi_\omega}(y), y\rangle - \psi_\theta\left(\nabla\overline{\psi_\omega}(y)\right)\right]\right];$$
3. Compute the Monte-Carlo estimate of the cycle-consistency regularizer:
$$\mathcal{L}_{\text{Cycle}} := \frac{1}{K}\sum_{y \in Y} \|\nabla\psi_\theta \circ \nabla\overline{\psi_\omega}(y) - y\|_2^2;$$
4. Compute the total loss $\mathcal{L}_{\text{Total}} := \mathcal{L}_{\text{Corr}} + \frac{\lambda}{2} \cdot \mathcal{L}_{\text{Cycle}}$;
5. Perform a gradient step over $\{\theta, \omega\}$ by using $\frac{\partial \mathcal{L}_{\text{Total}}}{\partial\{\theta,\omega\}}$;

end

Theoretical guarantees (supershort!)

Informal formulation:

approximation of correlation $\equiv$ searching for a generative map

$$\mathsf{Corr}(\mathbb{P}, \mathbb{Q} \mid \psi^{\dagger}, \psi^{\ddagger}; \lambda) \leq \mathsf{Corr}(\mathbb{P}, \mathbb{Q}) + \Theta(\epsilon)$$

is equivalent to

$$\mathbb{W}_2^2(\nabla\psi^{\dagger} \circ \mathbb{P}, \mathbb{Q}) \leq \Theta(\epsilon) \qquad \bigwedge \qquad \mathbb{W}_2^2(\nabla\overline{\psi^{\ddagger}} \circ \mathbb{Q}, \mathbb{P}) \leq \Theta(\epsilon)$$

Theoretical Guarantees: Part 1

Theorem (Generative Property for Approximators of Correlations)
Let $\mathbb{P}, \mathbb{Q}$ be two continuous probability distributions on $\mathcal{X} = \mathcal{Y} = \mathbb{R}^D$ with finite second moments. Let $\psi^* : \mathcal{X} \to \mathbb{R}$ be the optimal convex potential:

$$\psi^* = \underset{\psi \in \text{Convex}}{\arg\min}\ \text{Corr}(\mathbb{P}, \mathbb{Q}|\psi) = \underset{\psi \in \text{Convex}}{\arg\min} \left[\int_{\mathcal{X}} \psi(x) d\mathbb{P}(x) + \int_{\mathcal{Y}} \overline{\psi}(y) d\mathbb{Q}(y) \right].$$

Let two differentiable convex functions $\psi^\dagger : \mathcal{X} \to \mathbb{R}$ and $\overline{\psi^\ddagger} : \mathcal{Y} \to \mathbb{R}$ satisfy for some $\epsilon \in \mathbb{R}$:

$$\text{Corr}\left(\mathbb{P}, \mathbb{Q} \mid \psi^\dagger, \overline{\psi^\ddagger}; \lambda\right) - \text{Corr}(\mathbb{P}, \mathbb{Q} \mid \psi^*) = \epsilon. \tag{1}$$

Assume that $\psi^\dagger$ is $\beta^\dagger$-strongly convex ($\beta^\dagger > \frac{1}{\lambda} > 0$) and $\mathcal{B}^\dagger$-smooth ($\mathcal{B}^\dagger \geq \beta^\dagger$). Assume that $\overline{\psi^\ddagger}$ has bijective gradient $\nabla \overline{\psi^\ddagger}$. Then the following inequalities hold true:

Theoretical Guarantees: Part 1

1. **Correlation Upper Bound**
   (regularized correlations dominate over the true ones)

$$\mathrm{Corr}\big(\mathbb{P},\mathbb{Q}\mid \psi^{\dagger},\overline{\psi^{\ddagger}};\lambda\big)\geq \mathrm{Corr}\big(\mathbb{P},\mathbb{Q}\big)\qquad (\text{i.e. } \epsilon\geq 0);$$

2. **Forward Generative Property**
   (mapping $g^{\dagger}=\nabla\psi^{\dagger}$ pushes $\mathbb{P}$ to be $O(\epsilon)$-close to $\mathbb{Q}$)

$$\mathbb{W}_2^2(g^{\dagger}\circ\mathbb{P},\mathbb{Q})=\mathbb{W}_2^2(\nabla\psi^{\dagger}\circ\mathbb{P},\mathbb{Q})\leq \frac{(\mathcal{B}^{\dagger})^2\cdot\epsilon}{\lambda\beta^{\dagger}-1}\cdot\Big[\frac{1}{\sqrt{\beta^{\dagger}}}+\sqrt{\lambda}\Big]^2;$$

3. **Inverse Generative Property**
   (mapping $(g^{\ddagger})^{-1}=\nabla\overline{\psi^{\ddagger}}$ pushes $\mathbb{Q}$ to be $O(\epsilon)$-close to $\mathbb{P}$)

$$\mathbb{W}_2^2\big((g^{\ddagger})^{-1}\circ\mathbb{Q},\mathbb{P}\big)=\mathbb{W}_2^2(\nabla\overline{\psi^{\ddagger}}\circ\mathbb{Q},\mathbb{P})\leq\frac{\epsilon}{\beta^{\dagger}-\frac{1}{\lambda}}.$$

Theoretical Guarantees: Part 2

Theorem (Approximability of Correlations)
Let $\mathbb{P}, \mathbb{Q}$ be two continuous probability distributions on $\mathcal{X} = \mathcal{Y} = \mathbb{R}^D$ with finite second moments. Let $\psi^* : \mathcal{Y} \to \mathbb{R}$ be the optimal convex potential.

Let $\Psi_{\mathcal{X}}, \overline{\Psi}_{\mathcal{Y}}$ be classes of differentiable convex functions $\mathcal{X} \to \mathbb{R}$ and $\mathcal{Y} \to \mathbb{R}$ respectively and

1. *$\exists \psi^{\mathcal{X}} \in \Psi_{\mathcal{X}}$ with $\epsilon_{\mathcal{X}}$-close gradient to the forward mapping $\nabla\psi^*$ in $\mathcal{L}^2(\mathcal{X} \to \mathbb{R}^D, \mathbb{P})$ sense:*

$$\|\nabla\psi^{\mathcal{X}} - \nabla\psi^*\|^2_{\mathbb{P}} \stackrel{def}{=} \int_{\mathcal{X}} \|\nabla\psi^{\mathcal{X}}(y) - \nabla\psi^*(y)\|^2 d\mathbb{P}(y) \leq \epsilon_{\mathcal{X}},$$

   and $\psi^{\mathcal{X}}$ is $\mathcal{B}^{\mathcal{X}}$-smooth;

2. *$\exists \overline{\psi^{\mathcal{Y}}} \in \overline{\Psi_{\mathcal{Y}}}$ with $\epsilon_{\mathcal{Y}}$-close gradient to the inverse mapping $\nabla\overline{\psi^*}$ in $\mathcal{L}^2(\mathcal{Y} \to \mathbb{R}^D, \mathbb{Q})$ sense:*

$$\|\nabla\overline{\psi^{\mathcal{Y}}} - \nabla\overline{\psi^*}\|^2_{\mathbb{Q}} \stackrel{def}{=} \int_{\mathcal{Y}} \|\nabla\overline{\psi^{\mathcal{Y}}}(y) - \nabla\overline{\psi^*}(y)\|^2 d\mathbb{Q}(y) \leq \epsilon_{\mathcal{Y}}.$$

Theoretical Guarantees: Part 2

Let $(\psi^\dagger, \overline{\psi^\ddagger})$ be the minimizers of the regularized correlations within $\Psi_{\mathcal{X}} \times \overline{\Psi_{\mathcal{Y}}}$:

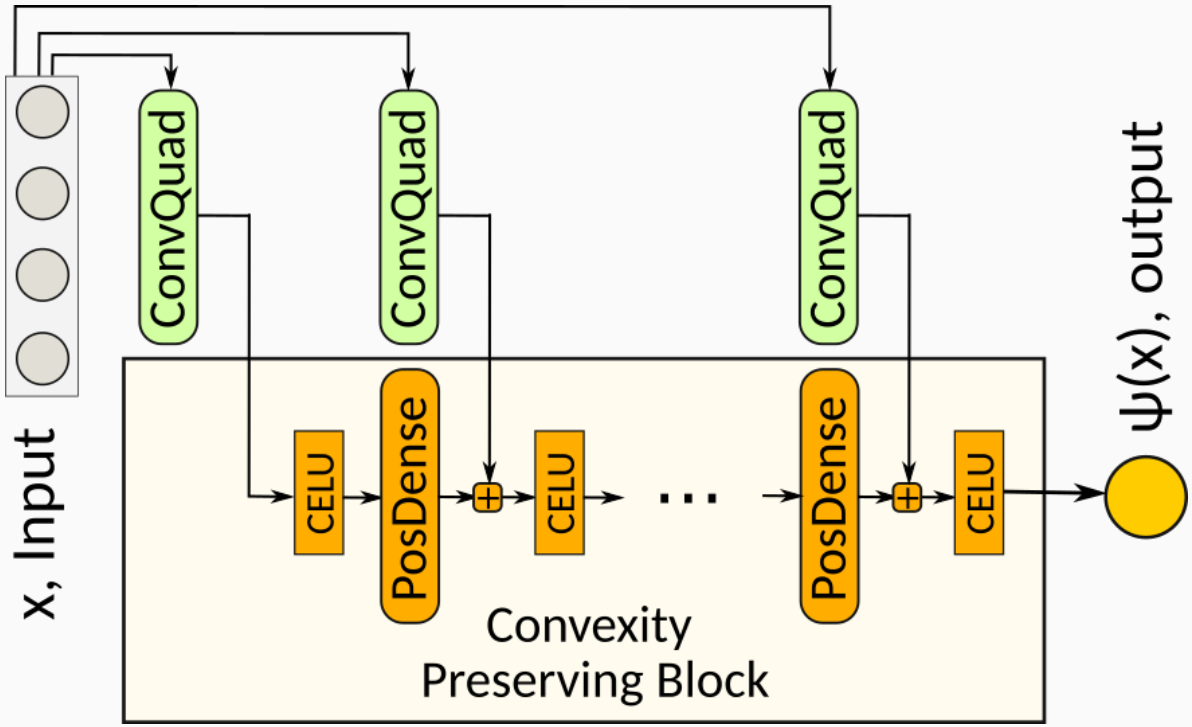
$$(\psi^\dagger, \overline{\psi^\ddagger}) = \underset{\psi \in \Psi_{\mathcal{X}}, \overline{\psi'} \in \overline{\Psi_{\mathcal{Y}}}}{\arg\min} \mathsf{Corr}\big(\mathbb{P}, \mathbb{Q} \mid \psi, \overline{\psi'}; \lambda\big).$$

Then the regularized correlations for $(\psi^\dagger, \overline{\psi^\ddagger})$ satisfy the following inequality:

$$\mathsf{Corr}\big(\mathbb{P}, \mathbb{Q} \mid \psi^\dagger, \overline{\psi^\ddagger}; \lambda\big) \leq \mathsf{Corr}(\mathbb{P}, \mathbb{Q}) + \Big[\frac{\lambda}{2}(\mathcal{B}^{\mathcal{X}}\sqrt{\epsilon_{\mathcal{Y}}} + \sqrt{\epsilon_{\mathcal{X}}})^2 + (\mathcal{B}^{\mathcal{X}}\sqrt{\epsilon_{\mathcal{Y}}} + \sqrt{\epsilon_{\mathcal{X}}}) \cdot (\sqrt{\epsilon_{\mathcal{Y}}}) + \frac{\mathcal{B}^{\mathcal{X}}}{2}\epsilon_{\mathcal{Y}}\Big],$$

i.e. regularized correlations do not exceed true correlations plus $O(\epsilon_{\mathcal{X}} + \epsilon_{\mathcal{Y}})$ term.

Dense Input Convex Neural Network Architecture



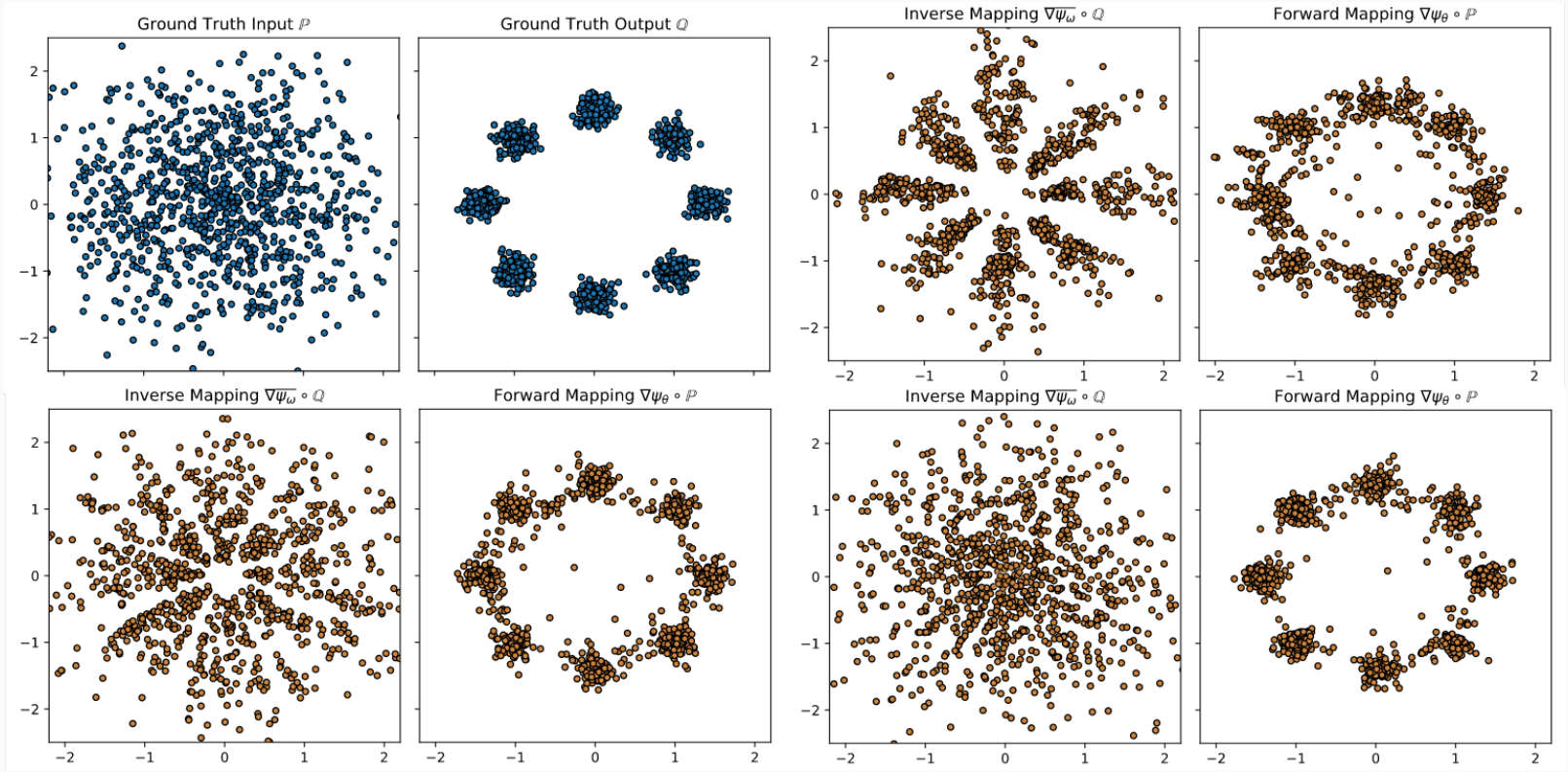
Convex Quadratic Layer (n-th neuron):

$$\mathrm{cq}_n(x) = \langle x, A_n x \rangle + \langle b_n, x \rangle + c_n, \qquad A_n = F_n^T F_n$$

Toy Experiments (2D)

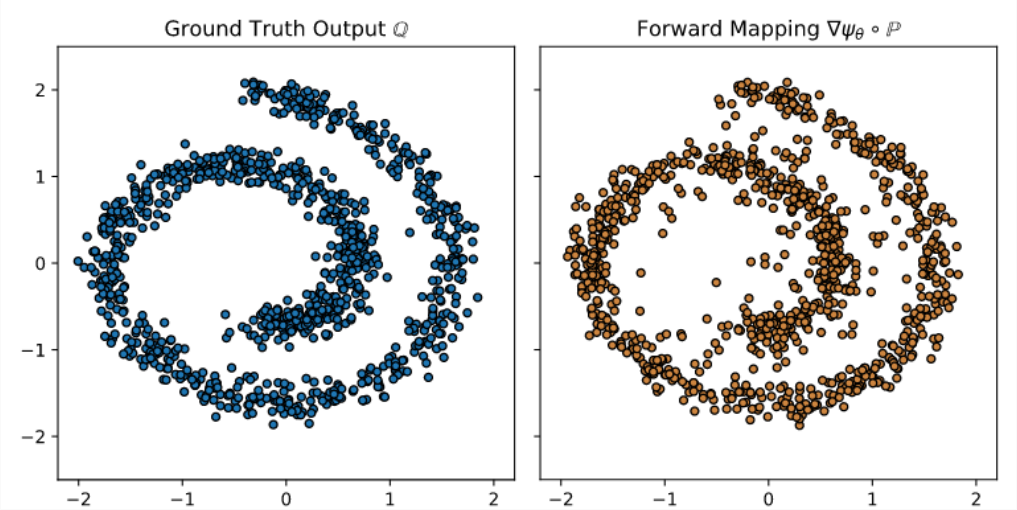
Fitting the optimal transport maps between Gaussian and Gaussian Mix

Convergence stages: Ground Truth, 200 steps, 2000 steps, 30000 steps.

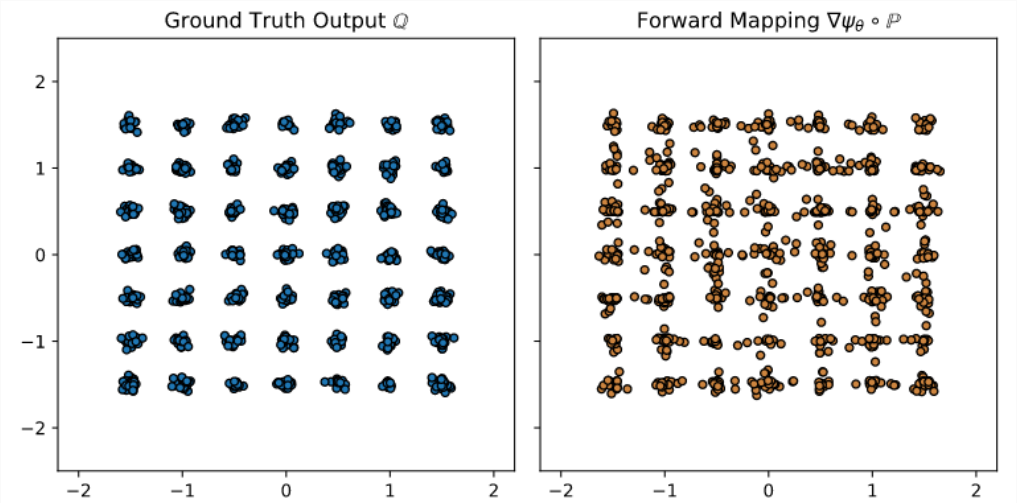


Toy Experiments (2D)

Fitting Swiss Roll

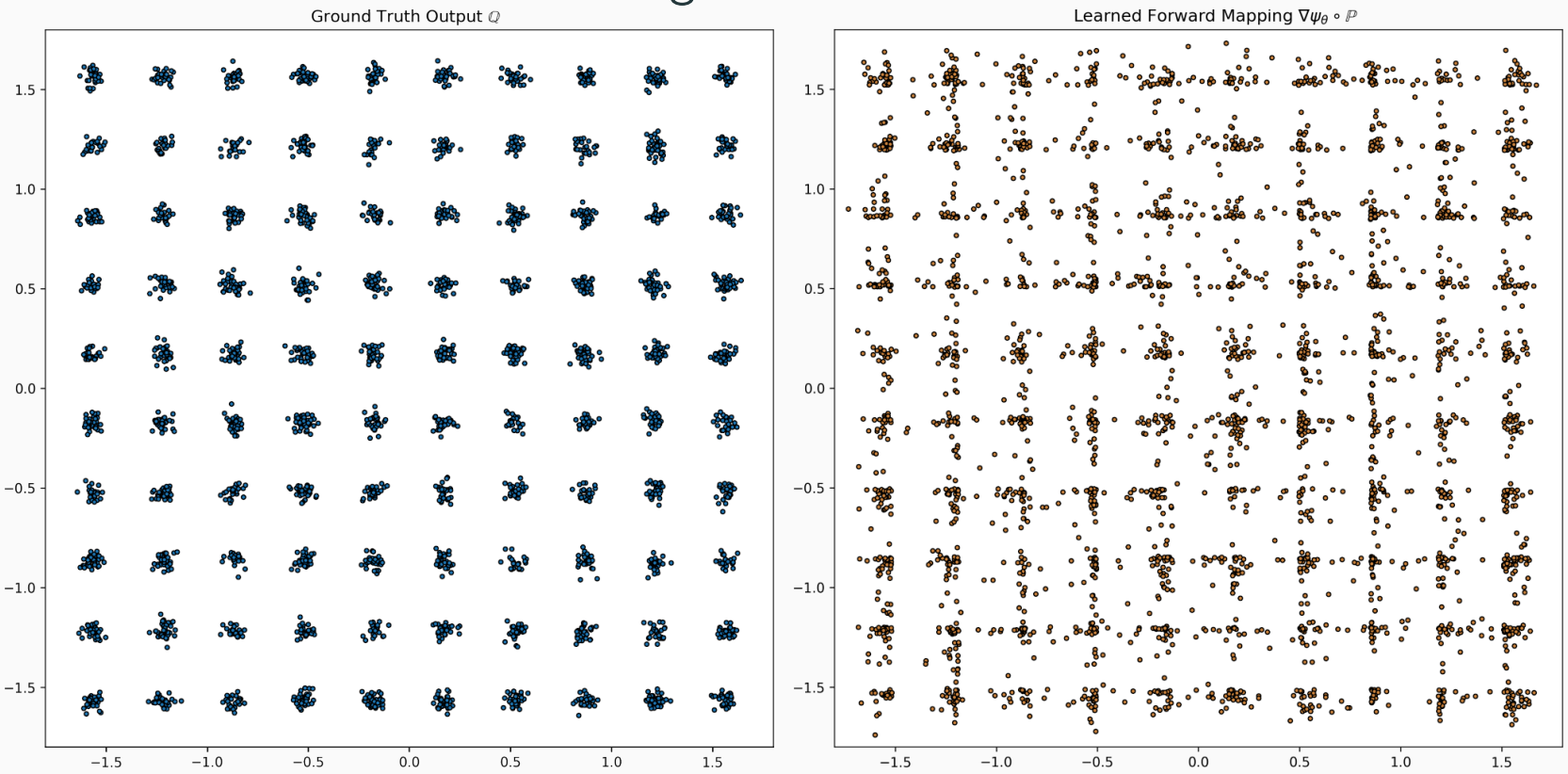


Fitting 49 Gaussians



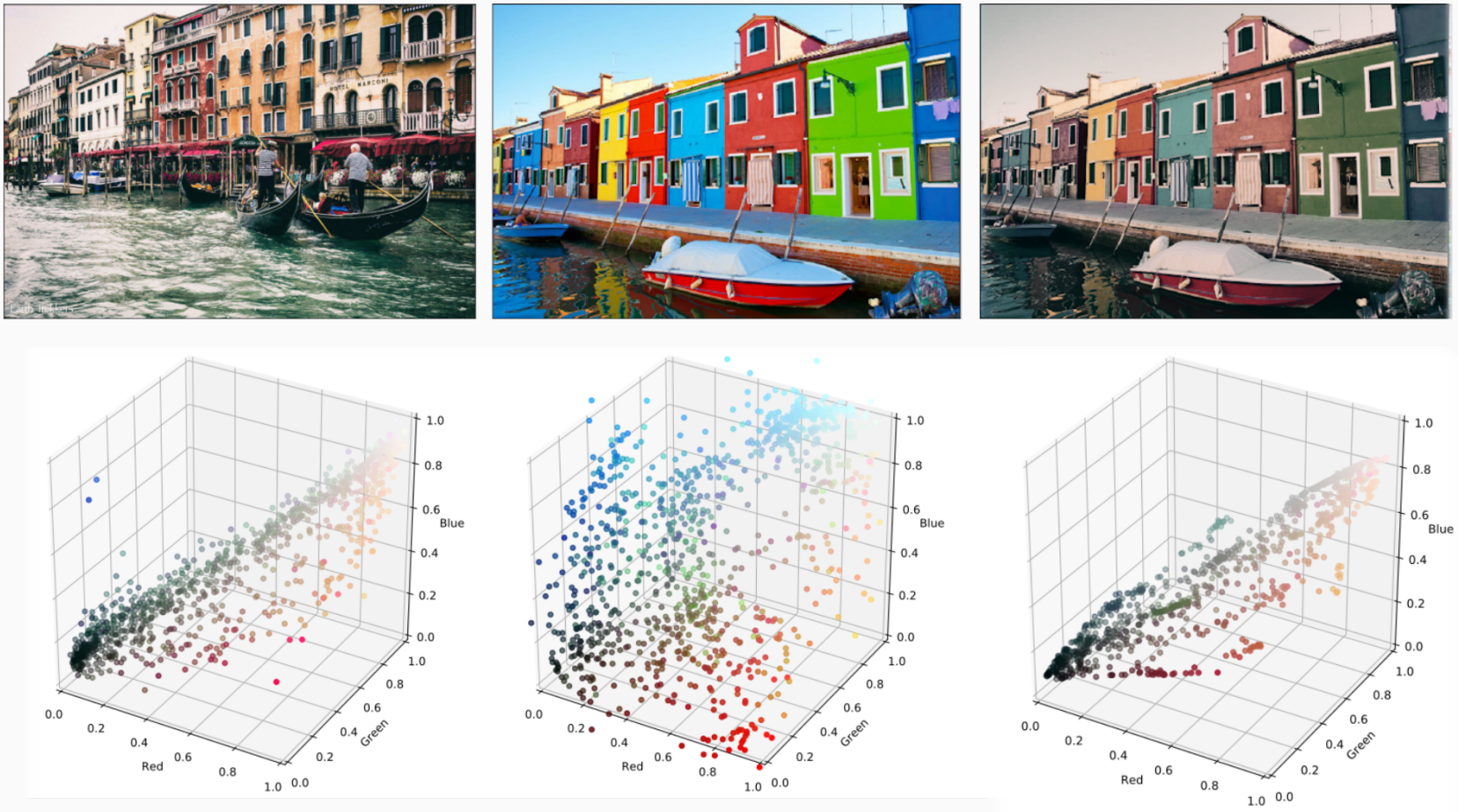
Toy Experiments (2D)

Fitting 100 Gaussians



Color Transfer (3D)

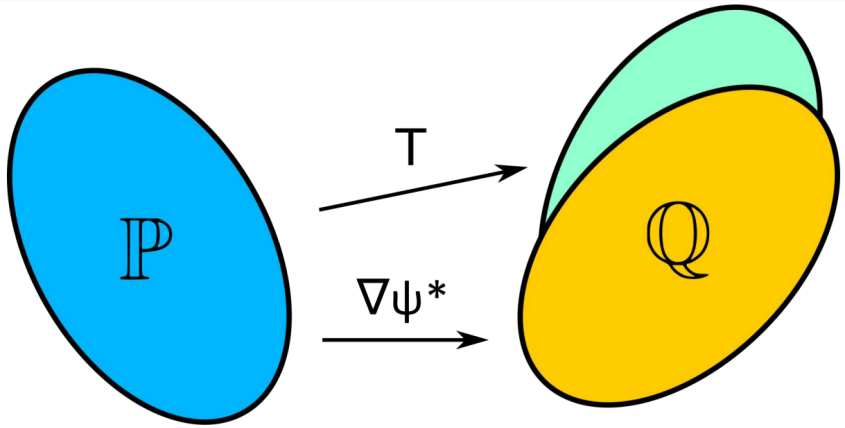
Cycle monotone map in $\mathbb{R}^3$ RGB color space



High-dimensional Gaussian Optimal Transport

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

Optimal transport map is known in a closed form!



Metric: $\mathcal{L}^2$ unexplained variance percentage

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

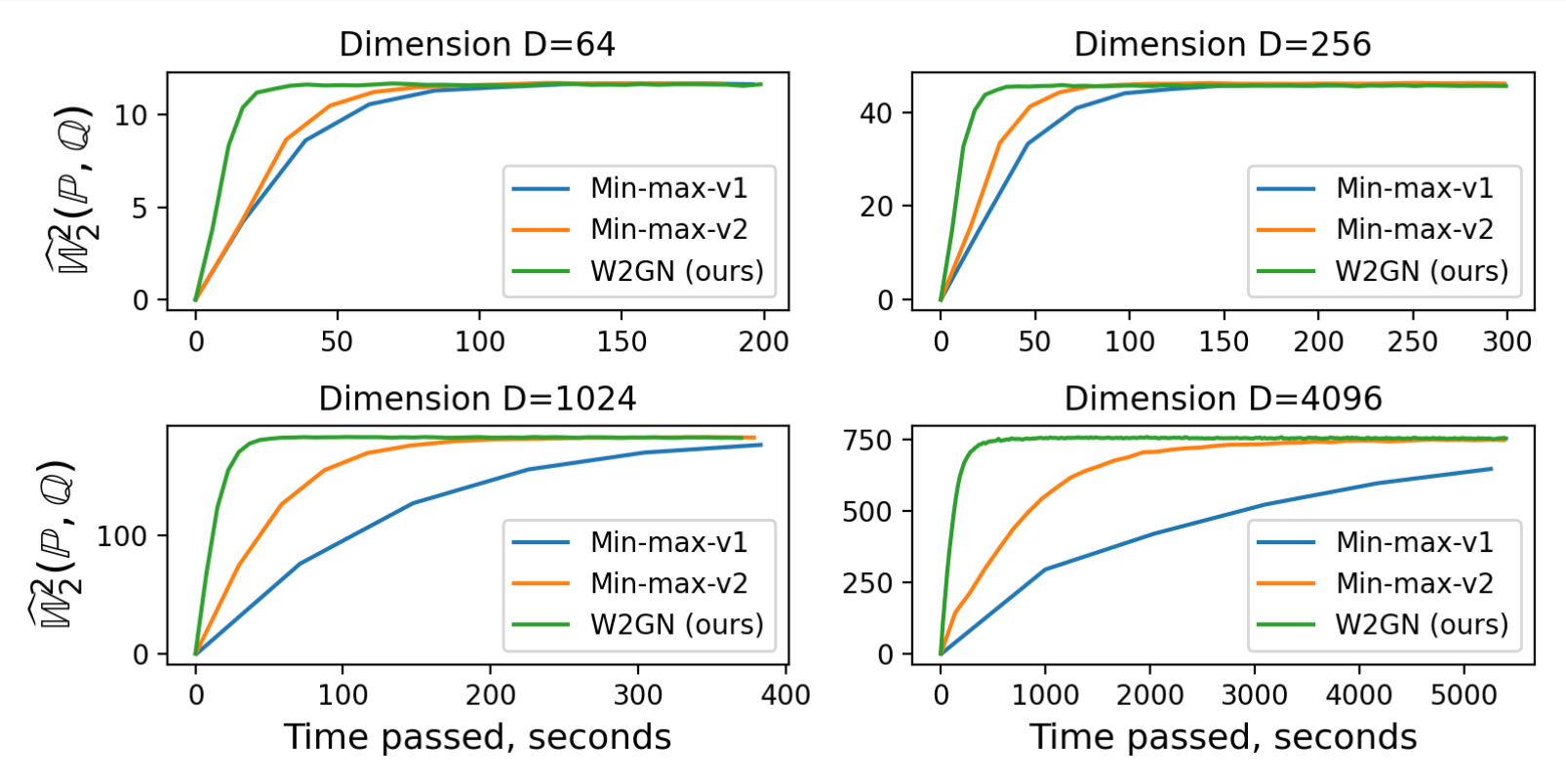
High-Dimensional Gaussian Optimal Transport

Comparison of $\mathcal{L}^2$-UVP (%) for LSOT, MM-1, MM-2 and W2GN (ours) methods in dimensions $D = 2, 4, \ldots, 2^{12}$.

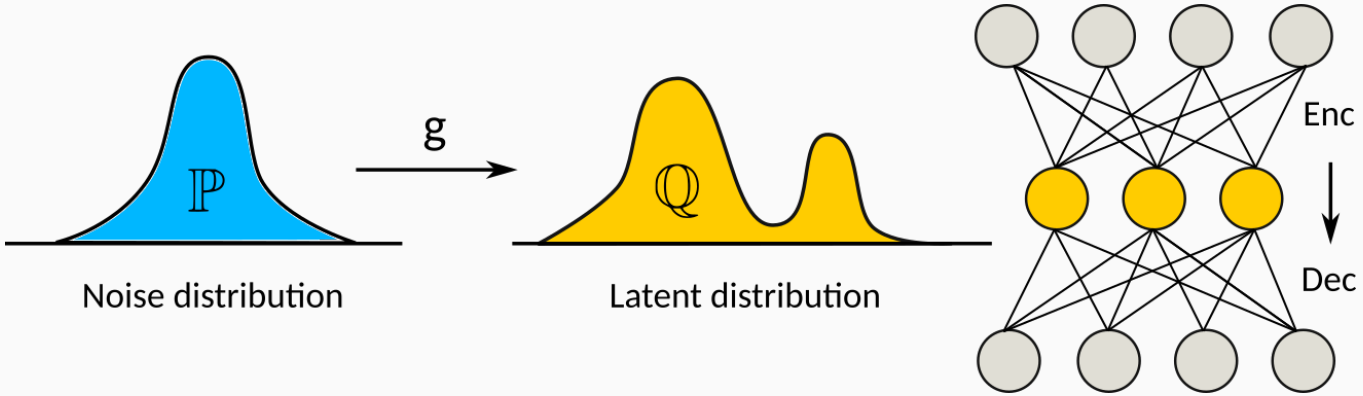
Dim	*2*	*4*	*8*	*16*	*32*	*64*	*128*	*256*	*512*	*1024*	*2048*	*4096*
LSOT	< 1	3.7	7.5	14.3	23	34.7	46.9	> 50				
MM-1	< 1	< 1	< 1	< 1	< 1	1.2	1.4	1.3	1.5	1.6	1.8	2.7
MM-2	< 1	< 1	< 1	< 1	< 1	< 1	1	1.1	1.2	1.3	1.5	2.1
W2GN	< 1	< 1	< 1	< 1	< 1	< 1	1	1.1	1.3	1.3	1.8	1.5

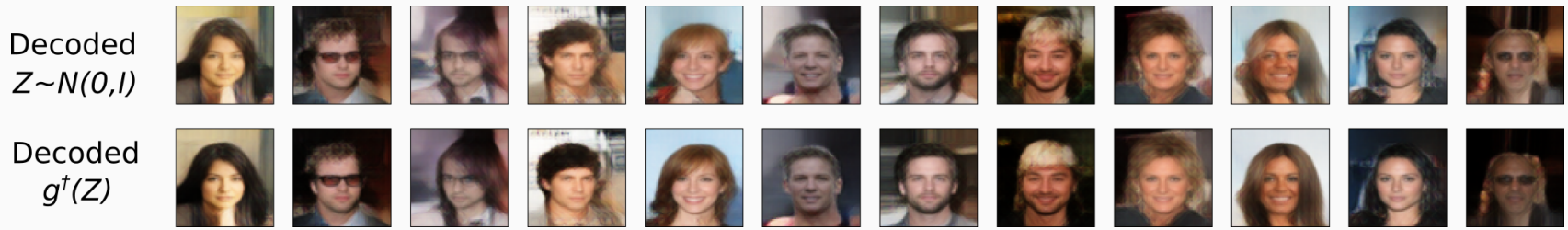
High-Dimensional Gaussian Optimal Transport

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



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

Latent Space Optimal Transport: Theory

Theorem (Decoding Theorem)
Let $\mathbb{S}$ be the real data distribution on $\mathcal{S} \subset \mathbb{R}^K$. Let $u : \mathcal{S} \to \mathcal{Y} = \mathbb{R}^D$ be the encoder and $v : \mathcal{Y} \to \mathbb{R}^K$ be L-Lipschitz decoder.

Assume that a latent space generative model has fitted a map $g^\dagger : \mathcal{X} \to \mathcal{Y}$ that pushes some latent distribution $\mathbb{P}$ on $\mathcal{X} = \mathbb{R}^D$ to be ϵ close to $\mathbb{Q} = u \circ \mathbb{S}$ in $\mathbb{W}_2^2$-sense, i.e.

$$\mathbb{W}_2^2(g^\dagger \circ \mathbb{P}, \mathbb{Q}) \leq \epsilon.$$

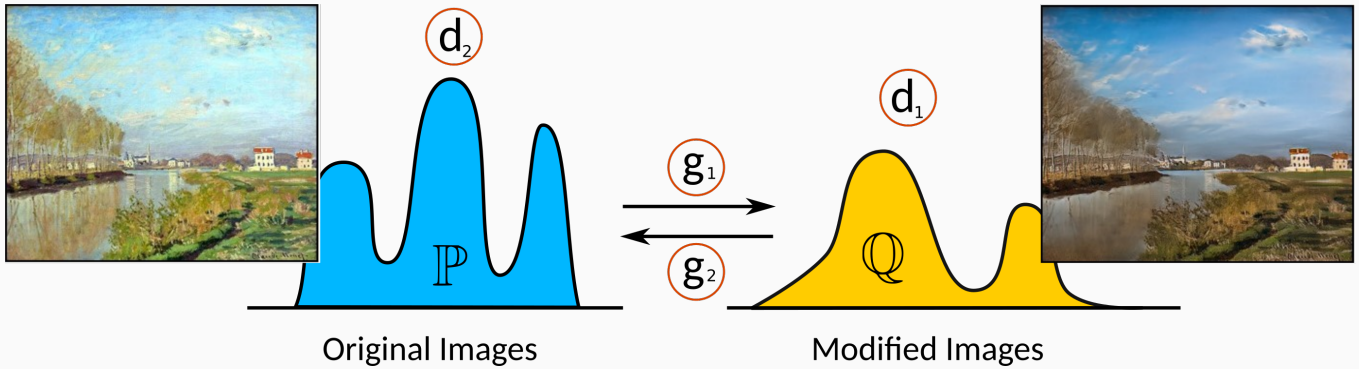
Then the following inequality holds true:

$$\mathbb{W}_2(\underbrace{v \circ g^\dagger \circ \mathbb{P}}_{\substack{\text{Generated data} \\ \text{distribuon}}}, \mathbb{S}) \leq L\sqrt{\epsilon} + \Big(\frac{1}{2}\underbrace{\mathbb{E}_{\mathbb{S}}\|s - v \circ u(s)\|_2^2}_{\substack{\text{Autoencoder's} \\ \text{reconstrution loss}}}\Big)^{\frac{1}{2}},$$

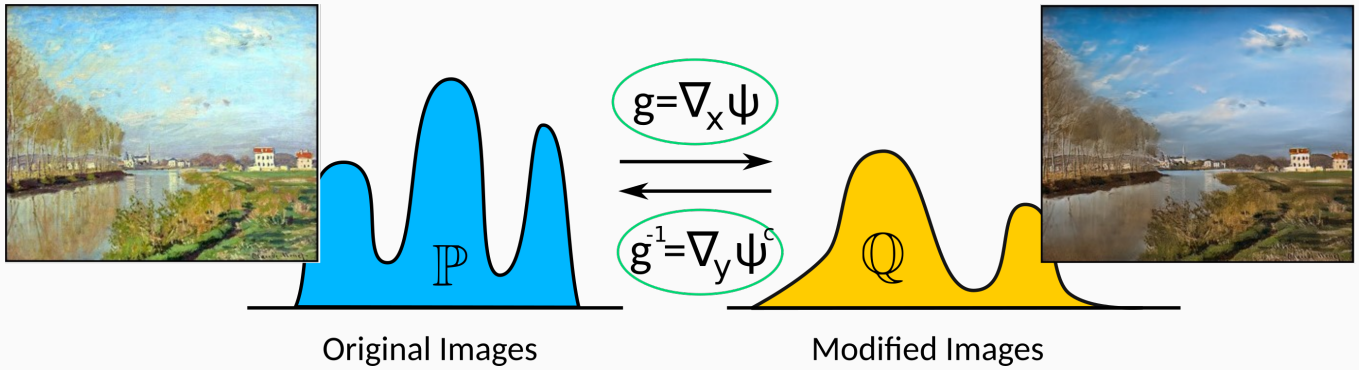
where $v \circ g^\dagger$ is the combined generative model.

Unpaired Image-to-image Style Transfer

Original Cycle GAN[10] - 4 networks



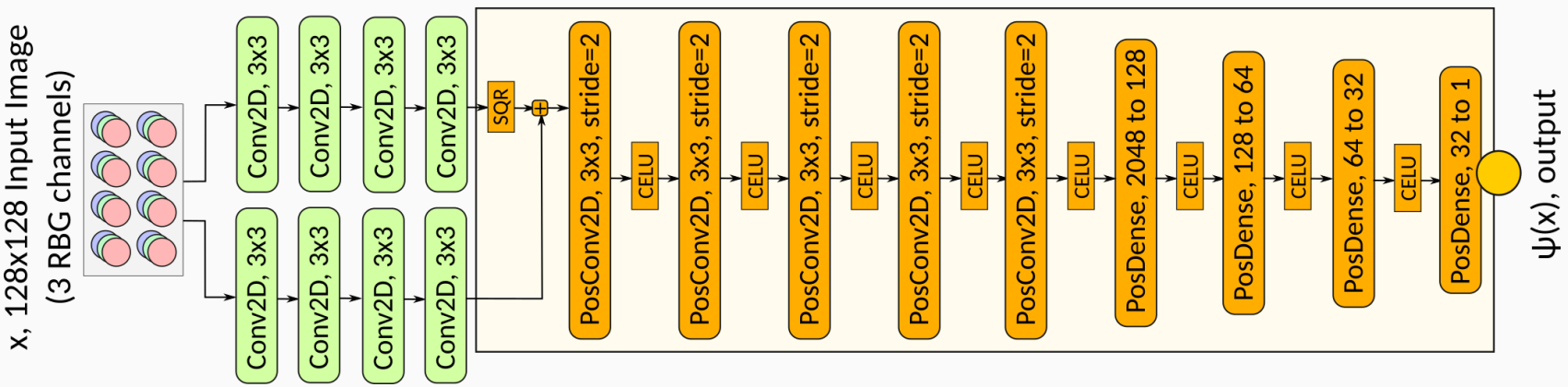
Wasserstein-2 GN - 2 networks



[10] 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.

Unpaired Image-to-image Style Transfer

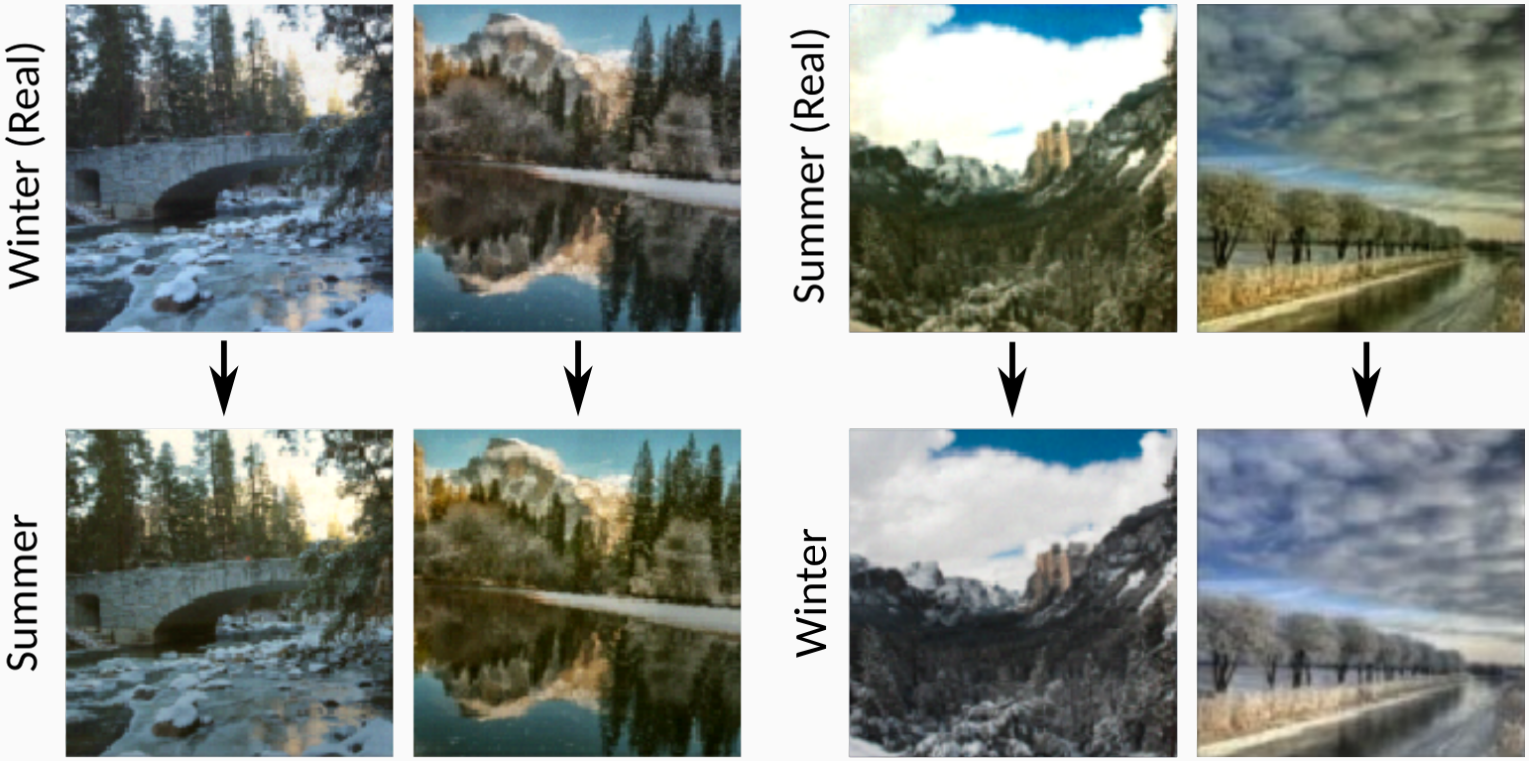
ψ_θ, ψ_ω are Convolutional Input Convex Neural Networks



Unpaired Image-to-image Style Transfer

128×128 image crops

Winter2SummerYosemite dataset



Wasserstein-2 Generative Networks - Conclusion

1. **Non-minimax** approach for fitting optimal maps by using ICNNs;
2. Superior properties of optimal maps
   - Cycle Monotonity
   - Existence
   - Uniqueness
   - Invertibility
3. Extreme ICNN sparsity: up to 90% network weights vanish.