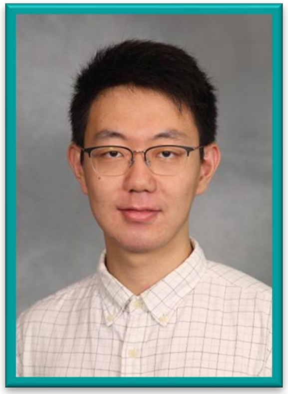
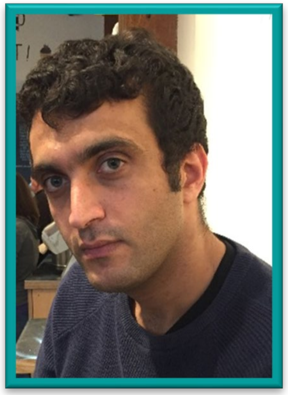
Gromov-Wasserstein Alignment: Statistics, Computation, and Geometry

Ziv Goldfeld
Cornell University



Zhengxin Zhang

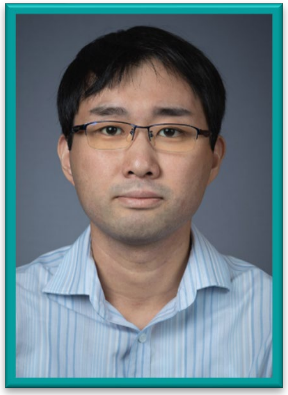
Kristjan Greenewald



Youssef Mroueh

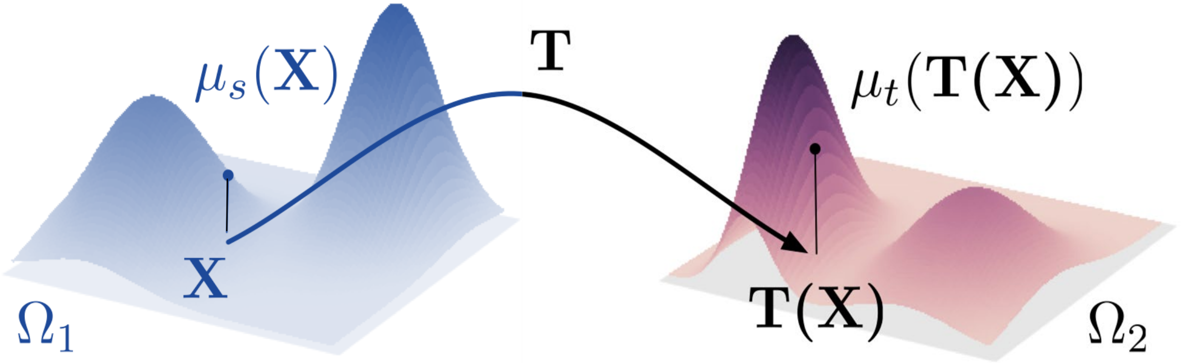
Bharath Sriprumbudur

Gabriel Rioux

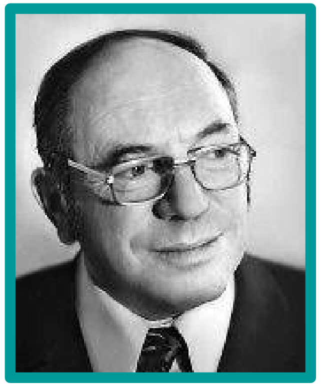


Kengo Kato

Optimal Transport



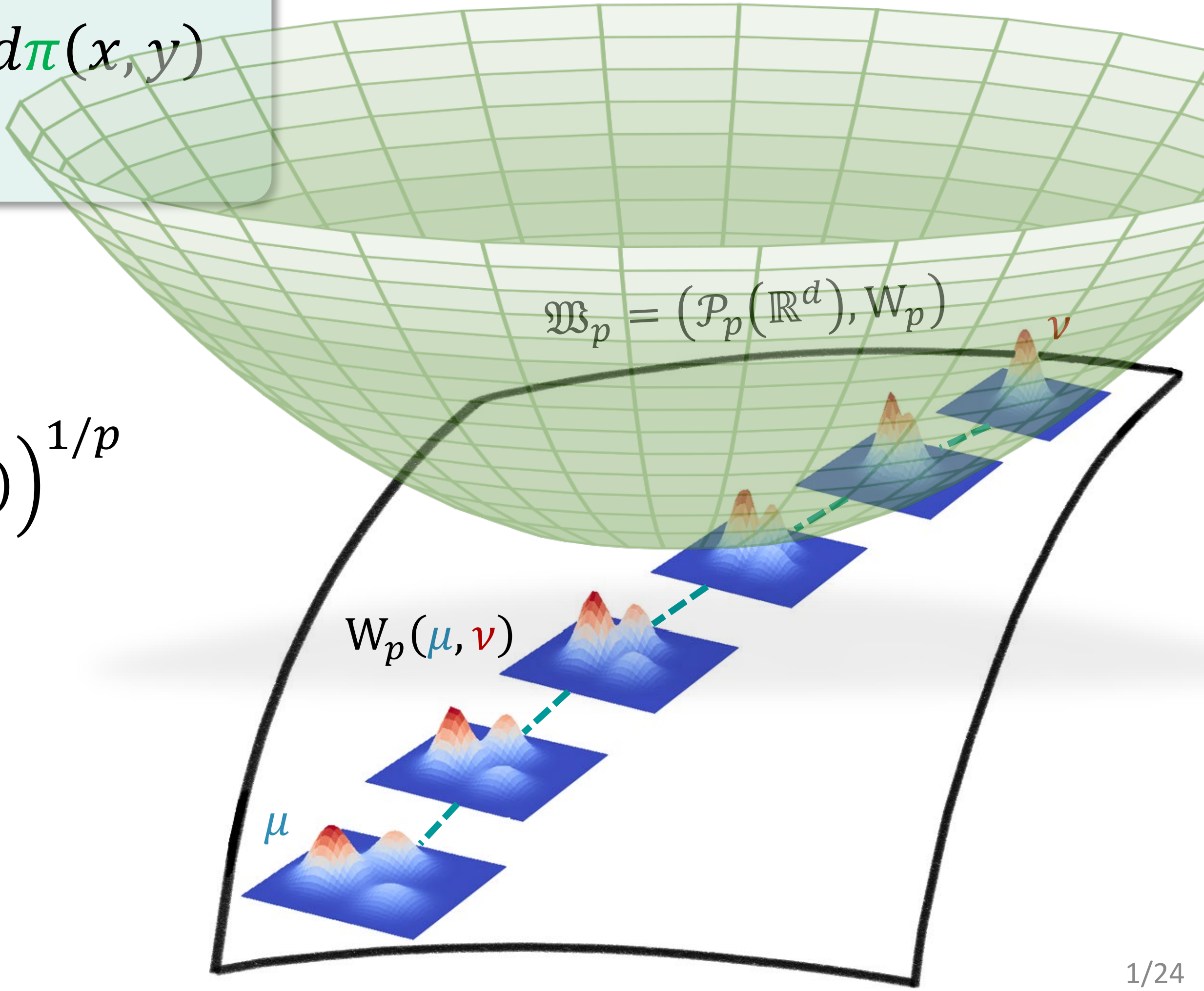
Optimal Transport



Kantorovich (1942)

Kantorovich Optimal Transport

$$\mathrm{OT}_c(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \iint c(x,y)\,d\pi(x,y)$$



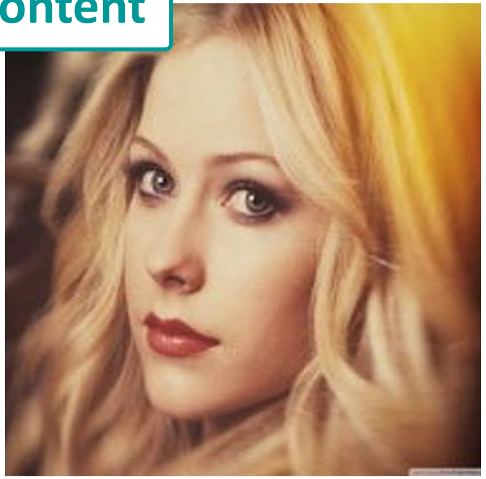
p-Wasserstein metric: $\mathrm{W}_p(\mu,\nu) := \left(\mathrm{OT}_{\|\cdot\|^p}(\mu,\nu)\right)^{1/p}$

- Geodesic curves
- Barycenters
- Gradient flows

$\vdots$

(Mixed) Style Transfer

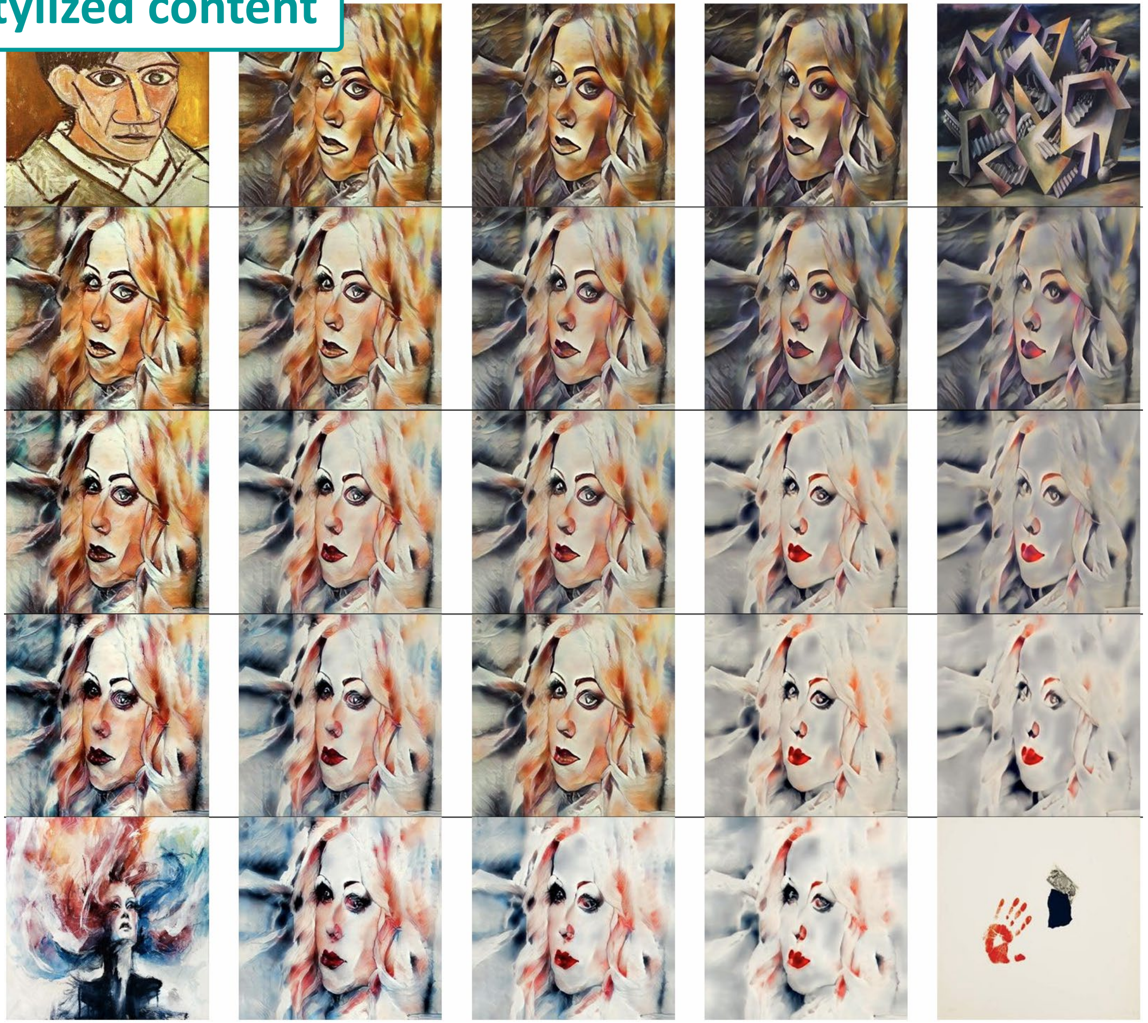
Content



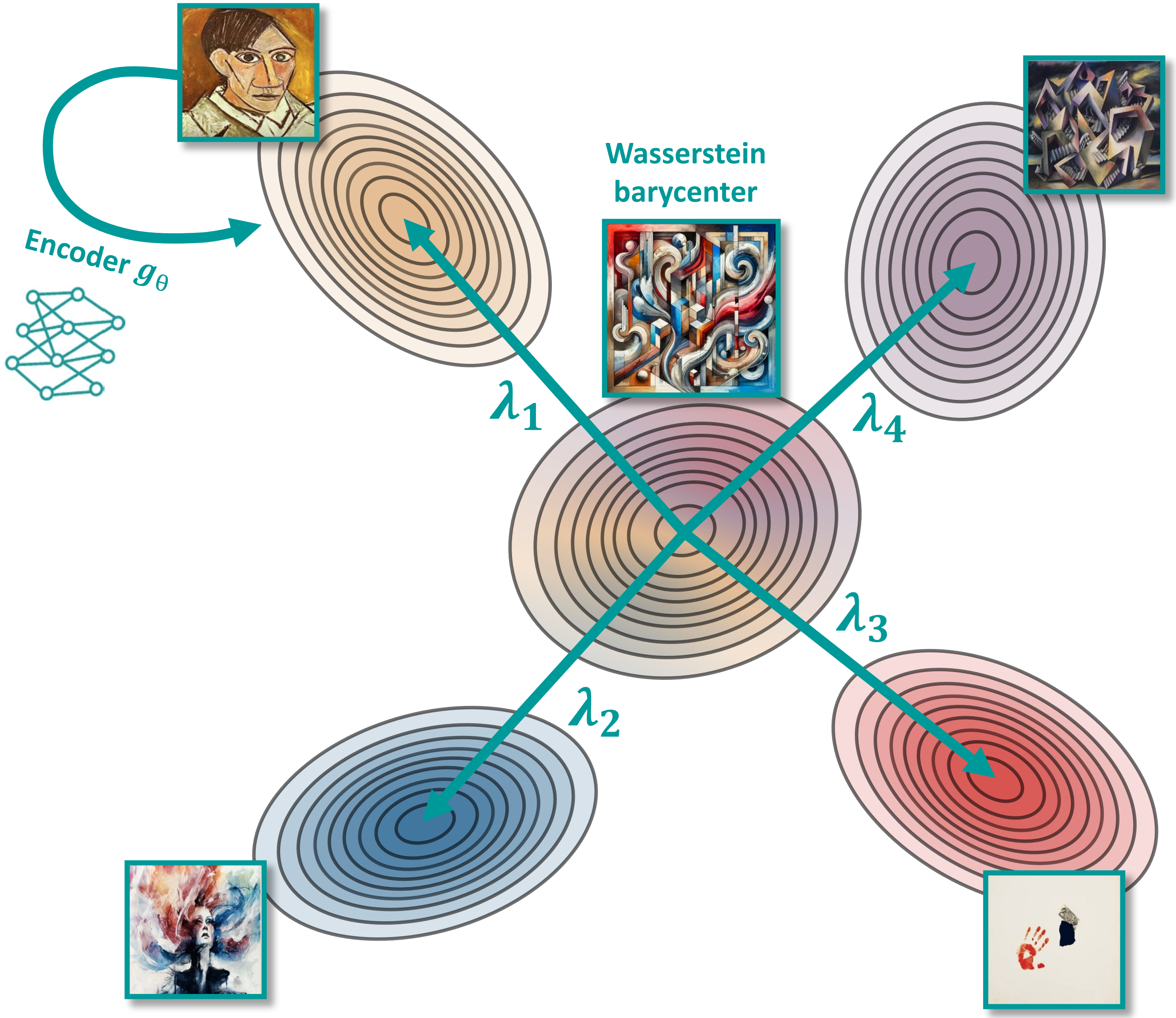
Styles

“AI”

Stylized content

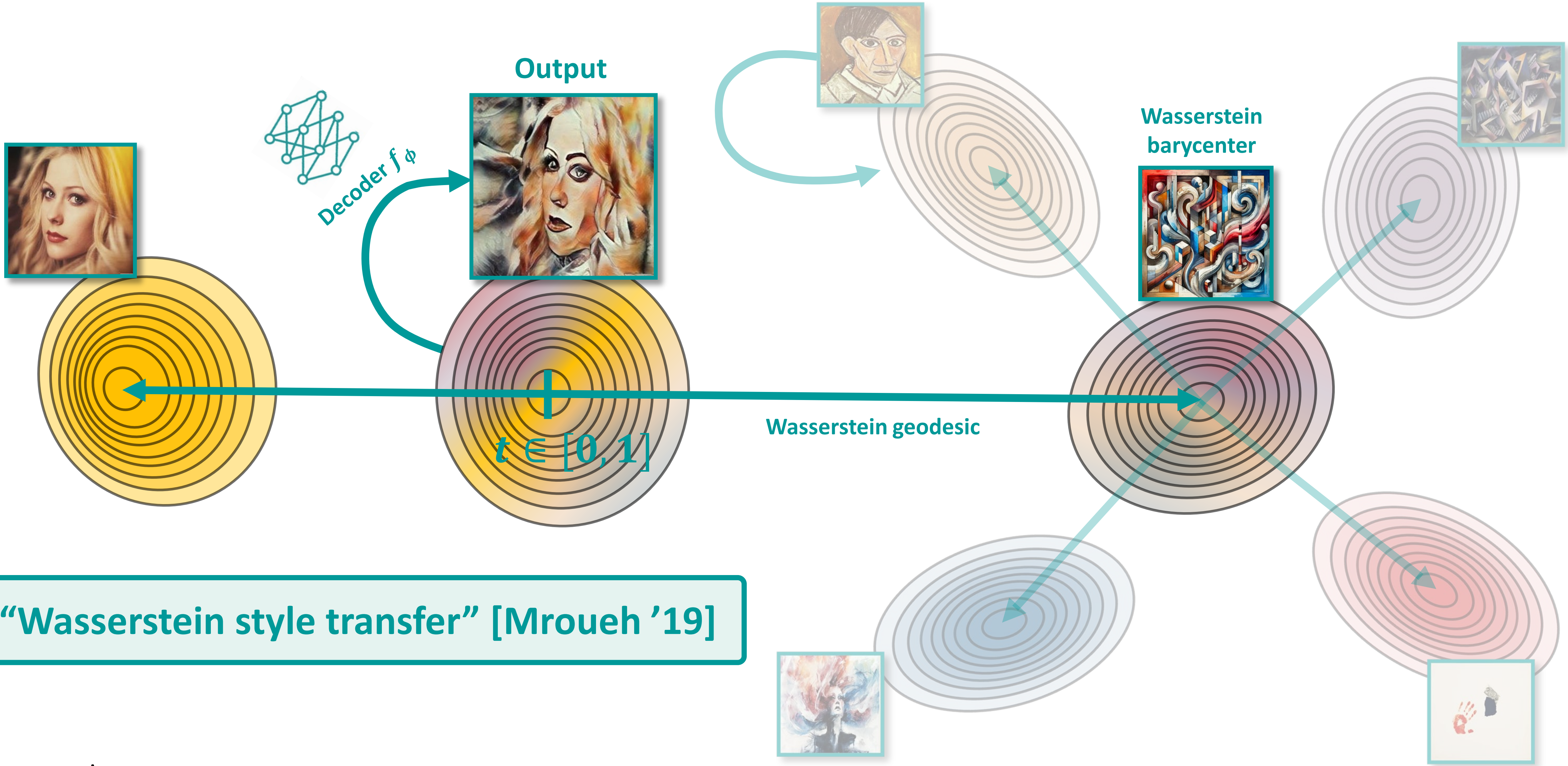


(Mixed) Style Transfer = Barycenter + Geodesic



Wasserstein space

(Mixed) Style Transfer = Barycenter + Geodesic



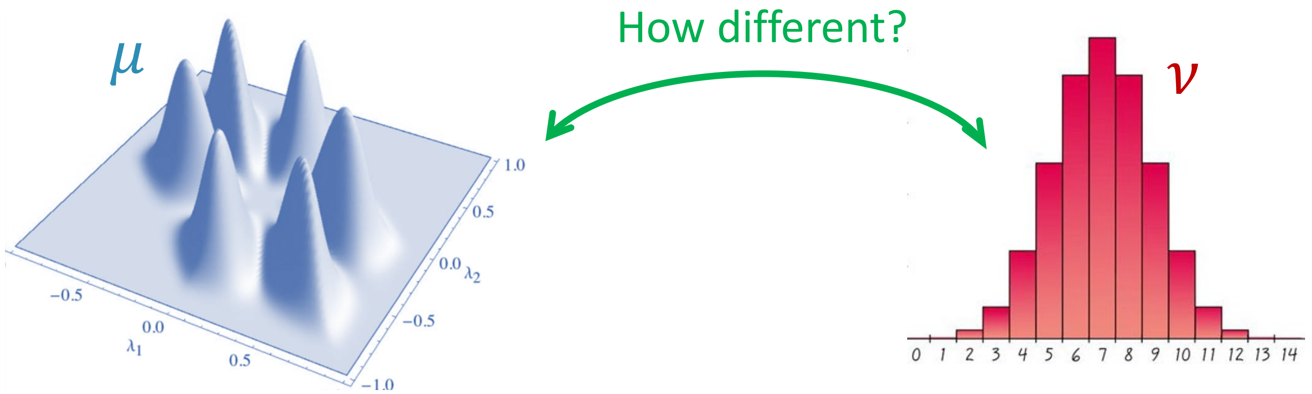
“Wasserstein style transfer” [Mroueh ’19]

Wasserstein space

Beyond OT and Wasserstein Distances

Structure Preserving Interpolation:

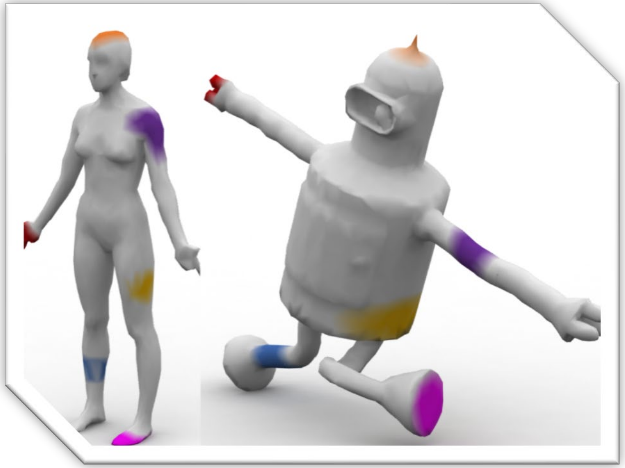
Discrepancy quantification btw incompatible spaces:



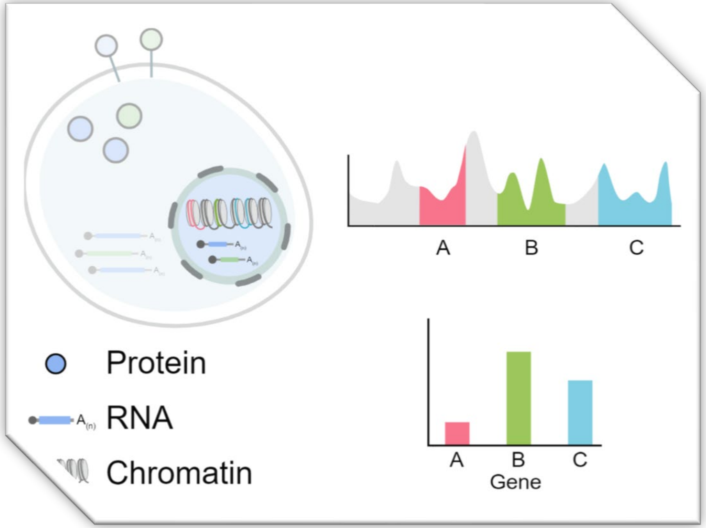
Gromov-Wasserstein Alignment

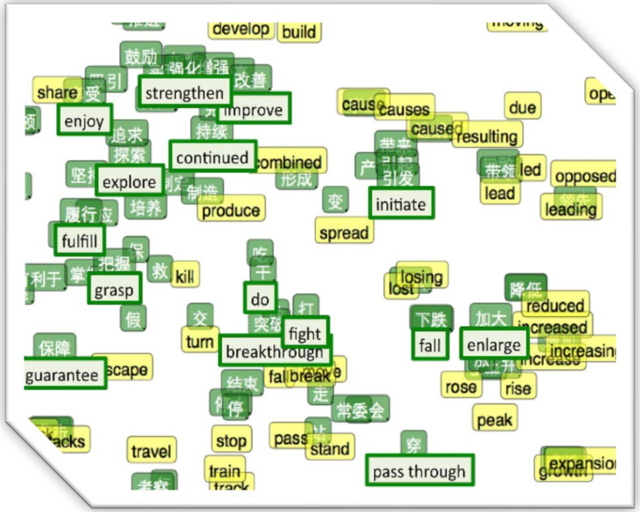
Heterogeneous & Structured Data

Dataset Matching: Various applications require matching heterogeneous **&** structured datasets



[Solomon-Peyré-Kim-Sra '16]





Goals:

1. Compare how similar/different two datasets are
2. Obtain matching/alignment

Gromov-Wasserstein Alignment

- Datasets as metric measure spaces

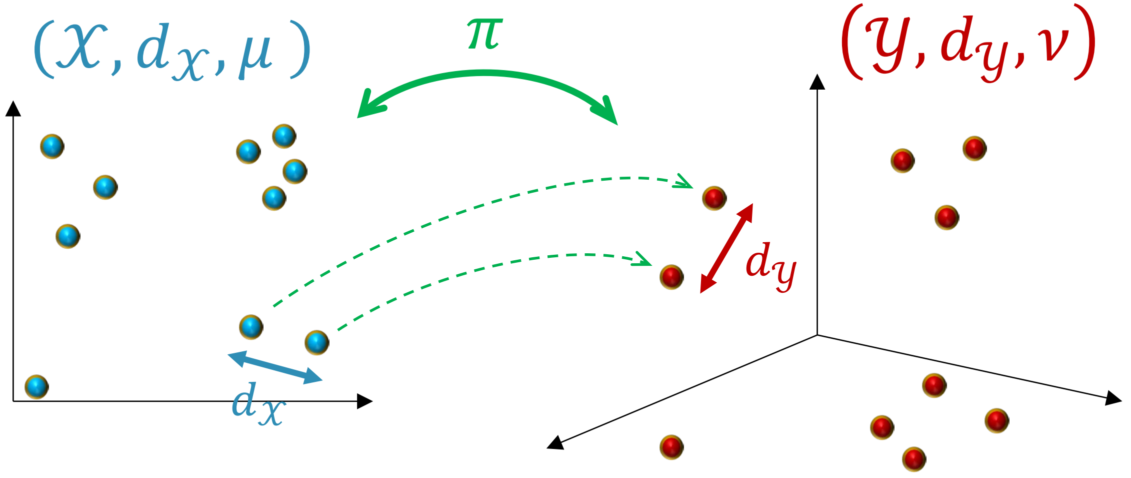
$\Longrightarrow \quad (\mathcal{X}, d_{\mathcal{X}}, \mu) \;\; \& \;\; (\mathcal{Y}, d_{\mathcal{Y}}, \nu)$

- Matching (transport map) $T: \mathcal{X} \to \mathcal{Y}$

$\Longrightarrow \quad \nu = T_{\#}\mu \;\; (X \sim \mu \Longrightarrow T(X) \sim \nu)$

- Preserve distances (minimize distance distortion)

$\Longrightarrow \quad \text{cost} = \left| d_{\mathcal{X}}(x_i, x_j)^2 - d_{\mathcal{Y}}\big(T(x_i), T(x_j)\big)^2 \right|$



Quadratic Gromov-Wasserstein Distance (Memoli '11)

$$\mathrm{D}(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \left(\mathbb{E}_{\pi \otimes \pi} \left[\left| d_{\mathcal{X}}(X, X')^2 - d_{\mathcal{Y}}(Y, Y')^2 \right|^2 \right] \right)^{1/2}$$

Gromov-Wasserstein Distance

$$\mathrm{D}_{p,q}(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \left(\mathbb{E}_{\pi\otimes\pi}\left[\left|d_{\mathcal{X}}(X,X')^q - d_{\mathcal{Y}}(Y,Y')^q\right|^p\right]\right)^{1/p}$$

Comments: L^p-Relaxation of Gromov-Hausdorff distance btw metric spaces ($p=\infty, q=1$)

- **Finiteness:** $\mathrm{D}_{p,q}(\mu,\nu) < \infty \quad \forall\, \mu,\nu$ with $\mathbb{E}_{\mu\otimes\mu}[d_{\mathcal{X}}(X,X')^{pq}] < \infty$ **&** resp. for ν
- **Identification:** $\mathrm{D}_{p,q}(\mu,\nu) = 0 \iff \exists$ isometry $T\colon \mathcal{X} \to \mathcal{Y}$ with $T_{\#}\mu = \nu$ (invariances)
- **Metric:** Metrizes space of equivalence classes of mm spaces with finite size

Duality for Quadratic GW Distance

Setting: (2,2)-GW btw $\left(\mathbb{R}^{d_x}, \|\cdot\|, \mu\right)$ and $\left(\mathbb{R}^{d_y}, \|\cdot\|, \nu\right)$ with $M_4(\mu) \coloneqq \int \|x\|^4 d\mu(x)\,, M_4(\nu) < \infty$

$$\mathrm{D}(\mu,\nu)^2 = \inf_{\pi\in\Pi(\mu,\nu)} \iint \left|\|x-x'\|^2 - \|y-y'\|^2\right|^2 d\pi\otimes\pi$$

Decomposition: Assume w.l.o.g. that μ,ν are centered (invariance to translation); then

$$\mathrm{D}(\mu,\nu)^2 = \mathrm{S}_1(\mu,\nu) + \mathrm{S}_2(\mu,\nu)$$

where $\mathrm{S}_1(\mu,\nu) = \int \|x-x'\|^4 d\mu\otimes\mu + \int \|y-y'\|^4 d\nu\otimes\nu - 4\int \|x\|^2\|y\|^2 d\mu\otimes\nu$

$$\mathrm{S}_2(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int -4\|x\|^2\|y\|^2 d\pi - 8\sum_{\substack{1\le i\le d_x\\ 1\le j\le d_y}} \left(\int x_i y_j d\pi\right)^2$$

$\Longrightarrow$ Develop variational form for $\mathrm{S}_2(\mu,\nu)$!

Duality for Quadratic GW Distance

Approach: Linearize quadratic term using auxiliary variables

$$\mathrm{S}_2(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int -4\|x\|^2\|y\|^2 d\pi - 8\sum_{\substack{1\le i\le d_x\\1\le j\le d_y}} \left(\int x_i y_j d\pi\right)^2$$

$$= \inf_{\pi\in\Pi(\mu,\nu)} \int -4\|x\|^2\|y\|^2 d\pi + 32\sum_{\substack{1\le i\le d_x\\1\le j\le d_y}} \inf_{-\frac{M_{\mu,\nu}}{2}\le a_{ij}\le\frac{M_{\mu,\nu}}{2}} \left(a_{ij}^2 - \int a_{ij}x_i y_j d\pi\right)$$

Optimality at $a_{ij}^*(\pi) = 0.5\int x_i y_j d\pi$ and define $M_{\mu,\nu} = \sqrt{M_2(\mu)M_2(\nu)}$

$\mathcal{D}_{M_{\mu,\nu}}$ = entry-wise bdd $d_x \times d_y$-sized matrices

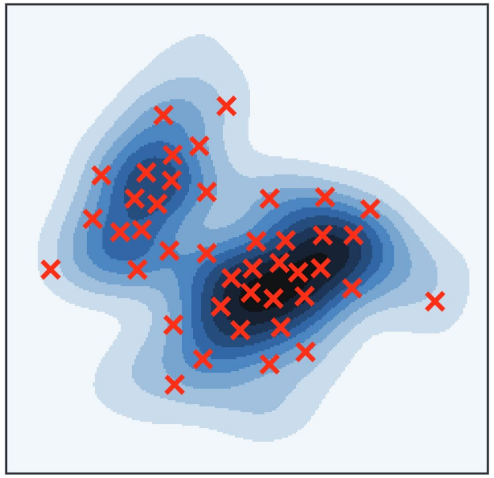
$$= \inf_{\mathbf{A}\in\mathcal{D}_{M_{\mu,\nu}}} \boxed{\inf_{\pi\in\Pi(\mu,\nu)} \int \underbrace{\left(32\|\mathbf{A}\|_{\mathrm{F}}^2 - 4\|x\|^2\|y\|^2 - 32x^{\mathrm{T}}\mathbf{A}y\right)}_{=:\, c_{\mathbf{A}}(x,y)} d\pi} = \mathrm{OT}_{\mathbf{A}}(\mu,\nu)$$

Theorem (Zhang-G.-Mroueh-Sriperumbudur '24)

$$\mathrm{S}_2(\mu,\nu) = \inf_{\mathbf{A}\in\mathcal{D}_{M_{\mu,\nu}}} \mathrm{OT}_{\mathbf{A}}(\mu,\nu)$$

Estimation from Data

Question: μ, ν are unknown; we sample $X_1, \ldots, X_n \sim \mu$ **&** $Y_1, \ldots, Y_n \sim \nu$



- **Empirical measures:** $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ and $\hat{\nu}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}$
- Approximate via plug-in $\mathrm{D}(\mu, \nu) \approx \mathrm{D}(\hat{\mu}_n, \hat{\nu}_n)$?

Asymptotic Ans: **Yes!** For μ, ν w/ finite 4–moments, $\mathrm{D}(\hat{\mu}_n, \hat{\nu}_n) \to \mathrm{D}(\mu, \nu)$ a.s. [Mémoli '11]

Non-Asymptotic Regime: What is the **rate** at which $\mathbb{E}[|\mathrm{D}(\mu, \nu) - \mathrm{D}(\hat{\mu}_n, \hat{\nu}_n)|]$ decays?

Open question: No known rates.

Sample Complexity of GW: Upper Bound

Theorem (Zhang-G.-Mroueh-Sriperumbudur '24)

Let $(\mu, \nu) \in \mathcal{P}(\mathbb{R}^{d_x}) \times \mathcal{P}(\mathbb{R}^{d_y})$ have compact support with diameter bounded by $R > 0.$ Then

$$\mathbb{E}\left[\left|\mathrm{D}(\mu,\nu)^2 - \mathrm{D}(\hat{\mu}_n,\hat{\nu}_n)^2\right|\right] \lesssim_{d_x,d_y,R} \underbrace{R^4 n^{-\frac{1}{2}}}_{\substack{\mathrm{S}_1 \text{ rate +} \\ \text{centering bias}}} + \underbrace{(1+R^4) n^{-\frac{2}{(d_x \wedge d_y)\vee 4}} (\log n)^{\mathbb{1}_{\{d_x \wedge d_y = 4\}}}}_{\mathrm{S}_2 \text{ rate}}$$

Comments:

- **Optimality:** These rates are sharp!
- **Data dimension:** Rate depends on smaller dimension (but curse of dimensionality occurs)
- **Comparison to OT:** Rate matches best known for OT

Sample Complexity of GW: Upper Bound

Theorem (Zhang-G.-Mroueh-Sriperumbudur '24)

Let $(\mu,\nu) \in \mathcal{P}(\mathbb{R}^{d_x}) \times \mathcal{P}(\mathbb{R}^{d_y})$ have compact support with diameter bounded by $R > 0$. Then

$$\mathbb{E}\left[\left|\mathrm{D}(\mu,\nu)^2 - \mathrm{D}(\hat{\mu}_n,\hat{\nu}_n)^2\right|\right] \lesssim_{d_x,d_y,R} \underbrace{R^4 n^{-\frac{1}{2}}}_{\substack{\mathrm{S}_1 \text{ rate +} \\ \text{centering bias}}} + \underbrace{(1+R^4) n^{-\frac{2}{(d_x \wedge d_y)\vee 4}} (\log n)^{\mathbb{1}_{\{d_x \wedge d_y = 4\}}}}_{\mathrm{S}_2 \text{ rate}}$$

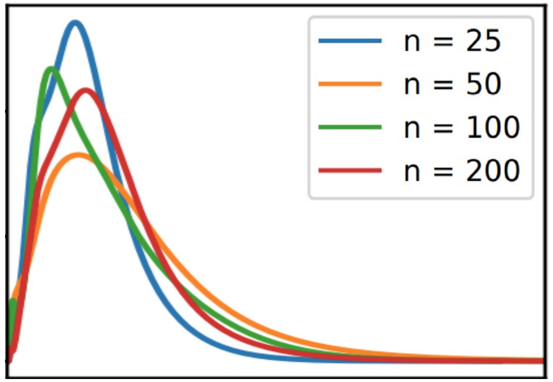
Proof outline:

1. **Decompose:** $\mathbb{E}\left[\left|\mathrm{D}^2 - \widehat{\mathrm{D}}^2\right|\right] \le \mathbb{E}\left[\left|\mathrm{S}_1 - \hat{\mathrm{S}}_1\right|\right] + \mathbb{E}\left[\left|\mathrm{S}_2 - \hat{\mathrm{S}}_2\right|\right] + n^{-1/2}$

2. **S_1 Analysis:** Estimation of moments

3. **S_2 Analysis:** $\mathbb{E}[|\mathrm{S}_2(\mu,\nu) - \mathrm{S}_2(\hat{\mu}_n,\hat{\nu}_n)|] \le \mathbb{E}\left[\sup_{\mathbf{A}\in\mathcal{D}_M} |\mathrm{OT}_{\mathbf{A}}(\mu,\nu) - \mathrm{OT}_{\mathbf{A}}(\hat{\mu}_n,\hat{\nu}_n)|\right]$

Limit Distribution Theory (teaser)



Question: Does empirical GW converge to a nondegenerate limit?

$$\boxed{r_n\big(\mathrm{D}(\hat{\mu}_n,\nu)-\mathrm{D}(\mu,\nu)\big)\underset{n\to\infty}{\longrightarrow}\chi}$$

Settings:
- Discrete-to-discrete
- Semi-discrete
- Sub-Weibull entropic GW

$$\Longrightarrow \quad \sqrt{n}(\hat{\mu}_n-\mu)\xrightarrow{d} G_\mu \ \text{ in } \ \ell^\infty(\mathcal{F})$$

Proof Recipe:

1. Weak convergence of empirical process (indexed by dual potentials)
2. Gâteaux derivative + Lipschitzness w.r.t. $\|\cdot\|_{\infty,\mathcal{F}}$ of $\mathrm{D}(\mu,\nu)$
3. Extended functional delta method $\Longrightarrow \sqrt{n}\big(\mathrm{D}(\hat{\mu}_n,\nu)-\mathrm{D}(\mu,\nu)\big)\longrightarrow \mathrm{D}'_\mu[G_\mu]$

Computation via Entropic Gromov-Wasserstein

GW is QAP: $\mathrm{D}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i},\frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}\right)^2 = \frac{1}{n^2}\min_{\sigma\in S_n}\sum_{i,j=1}^{n}\left|d_{\mathcal{X}}(x_i,x_j)^2 - d_{\mathcal{Y}}(y_{\sigma(i)},y_{\sigma(j)})^2\right|^2$

🚫 Quadratic assignment problem (non-convex) [Commander '05] $\Longrightarrow$ **NP complete**

Entropic Gromov-Wasserstein Distance (Peyré-Cuturi-Solomon '16)

$$\mathrm{D}^{\epsilon}(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{\pi\otimes\pi}\left[\left|\|X-X'\|^2 - \|Y-Y'\|^2\right|^2\right] + \epsilon \mathrm{D}_{\mathrm{KL}}(\pi\|\mu\otimes\nu)$$

Algorithms: Heuristic methods [Peyré-Cuturi-Solomon '16], [Solomon-Peyré-Kim-Sra '16]

From Stability Analysis to Convexity

$$\boxed{\mathrm{D}^{\epsilon}(\mu,\nu) = \mathrm{S}_1(\mu,\nu) + \min_{\mathbf{A}\in\mathcal{D}_M} \underbrace{\mathrm{EOT}^{\epsilon}_{\mathbf{A}}(\mu,\nu)}_{=:\Phi(\mathbf{A})}}$$

Analysis:

- Fréchet derivatives $D\Phi_{[\mathbf{A}]}$ and $D^2\Phi_{[\mathbf{A}]}$
- Bound $\lambda_{\max}\left(D^2\Phi_{[\mathbf{A}]}\right) \leq 64$ & $\lambda_{\min}\left(D^2\Phi_{[\mathbf{A}]}\right) \geq 64 - 32^2\epsilon^{-1}\sqrt{M_4(\mu)M_4(\nu)}$

Theorem (Rioux-G.-Kato '24)

1. Φ is L-smooth on $\mathcal{D}_M$ with $L \leq 64 \vee \left(32^2\epsilon^{-1}\sqrt{M_4(\mu)M_4(\nu)} - 64\right)$
2. Φ is strictly convex whenever $\epsilon > 16\sqrt{M_4(\mu)M_4(\nu)}$

Fast Gradient Method with Inexact Oracle

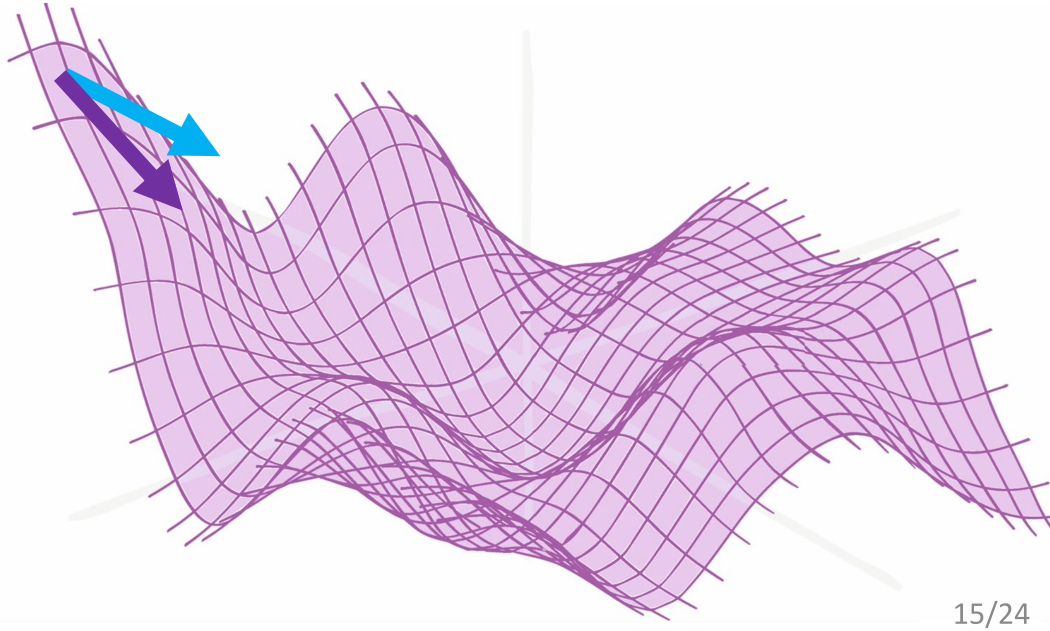
$$\boxed{\min_{\mathbf{A}\in\mathcal{D}_M} \mathrm{EOT}^{\epsilon}_{\mathbf{A}}(\mu,\nu)}$$

Gradient method: Gradient of objective at $\mathbf{A} \in \mathcal{D}_M$ depends on optimal EOT coupling $\pi^{\mathbf{A}}$

$$D\Phi_{[\mathbf{A}]} = 64\mathbf{A} - 32\sum_{i,j=1}^{n} x_i y_j^{\mathrm{T}} \pi_{i,j}^{\mathbf{A}}$$

Inexact oracle (Sinkhorn): $\tilde{\pi}^{\mathbf{A}}$ s.t. $\left\|\pi^{\mathbf{A}} - \tilde{\pi}^{\mathbf{A}}\right\|_{\infty} \leq \delta$

- Gradient approximation $\widetilde{D}\Phi_{[\mathbf{A}]}$ ($\tilde{\pi}^{\mathbf{A}}$ instead of $\pi^{\mathbf{A}}$)
- Fast gradient method



$\Longrightarrow$ Computes EGW cost and (approx.) coupling

Global Convergence Guarantees

$$\min_{\mathbf{A} \in \mathcal{D}_M} \mathrm{EOT}_{\mathbf{A}}^{\epsilon}(\mu, \nu)$$

Theorem (Rioux-G.-Kato '24; via [d'Aspremont '08])

If Φ is convex and L-smooth on $\mathcal{D}_M$ with global min $\mathbf{B}_*$, then $\mathbf{B}_k$ from Algorithm 1 satisfies

$$\Phi(\mathbf{B}_k) - \Phi(\mathbf{B}_*) \leq \frac{2L\|\mathbf{B}_*\|_{\mathrm{F}}^2}{(k+1)(k+2)} + O(M\delta)$$

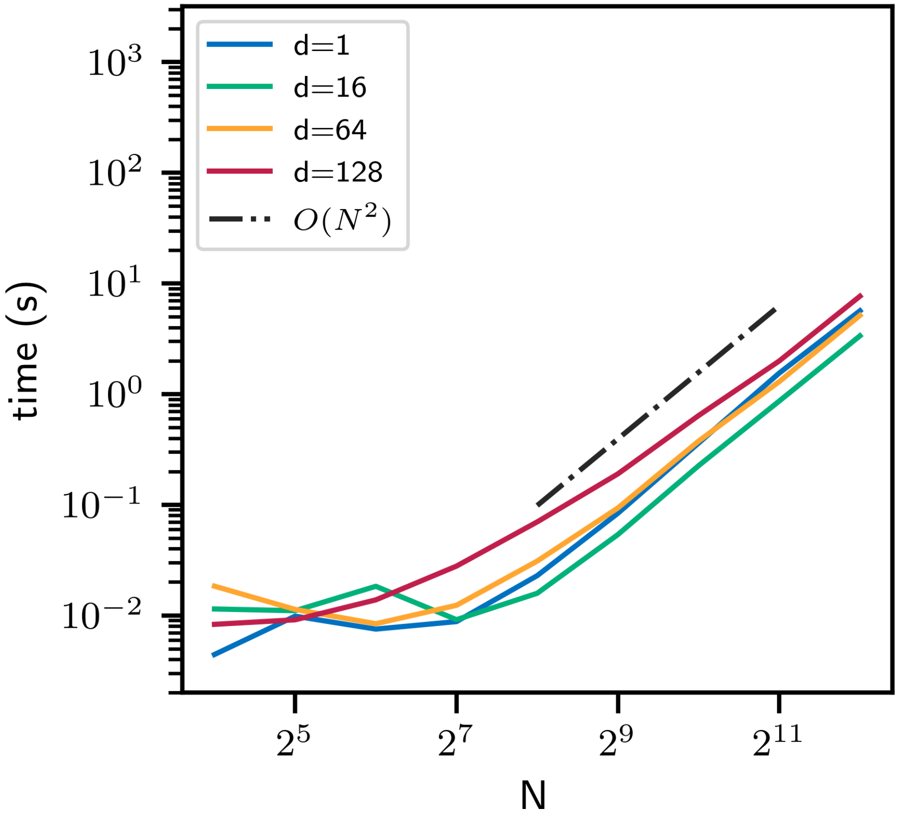
Comments:

- **Optimality:** Optimal complexity of $O\left(1/k^2\right)$ for smooth constrained opt. [Nesterov '03]
- **Non-convex regime (smooth):** Via smooth opt. with inexact oracle [Ghadimi-Lan '16]

  ↳ Adapts to convexity of Φ (improved rates if convex)

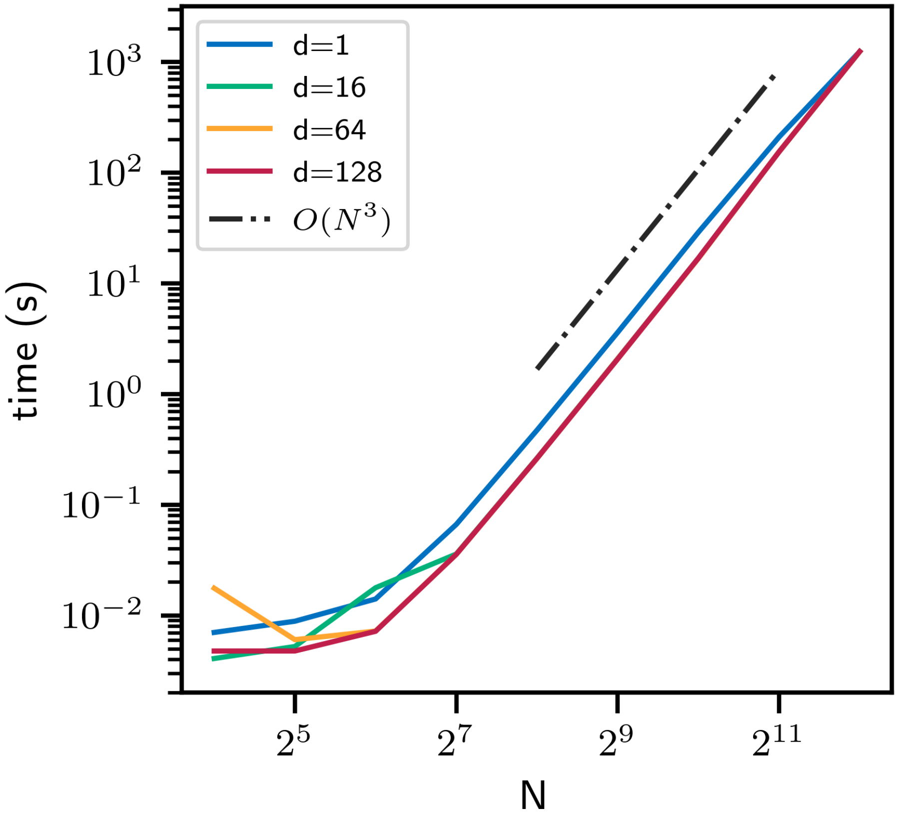
Numerical Results

Fast Gradient Method
[Rioux-G.-Kato '24]



Time = iteration × Sinkhorn
$= k \times O(N^2)$

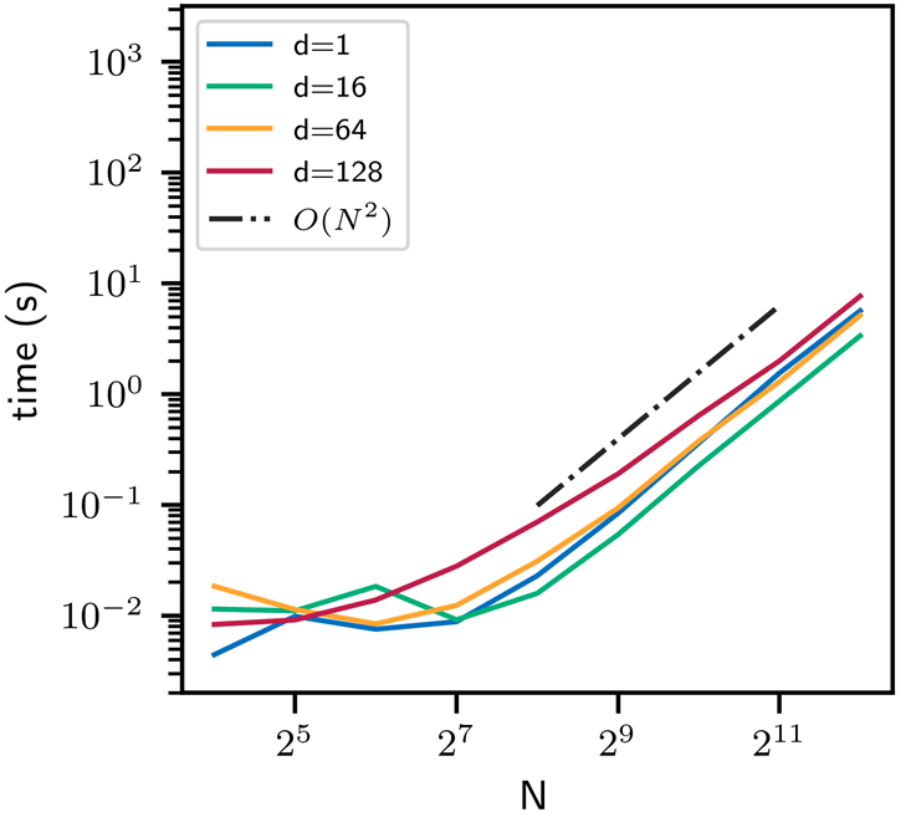
Mirror Descent
[Scetbon-Peyré-Cuturi '23]



Time = iteration × cost update
$= k \times O(N^3)$

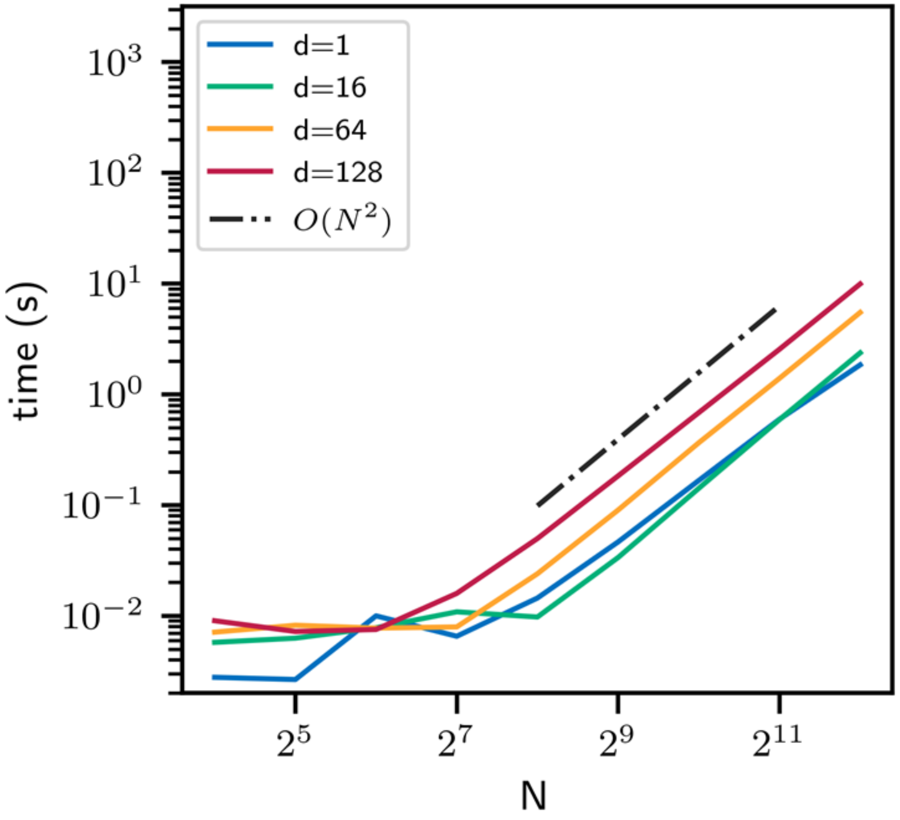
Numerical Results

Fast Gradient Method
[Rioux-G.-Kato '24]



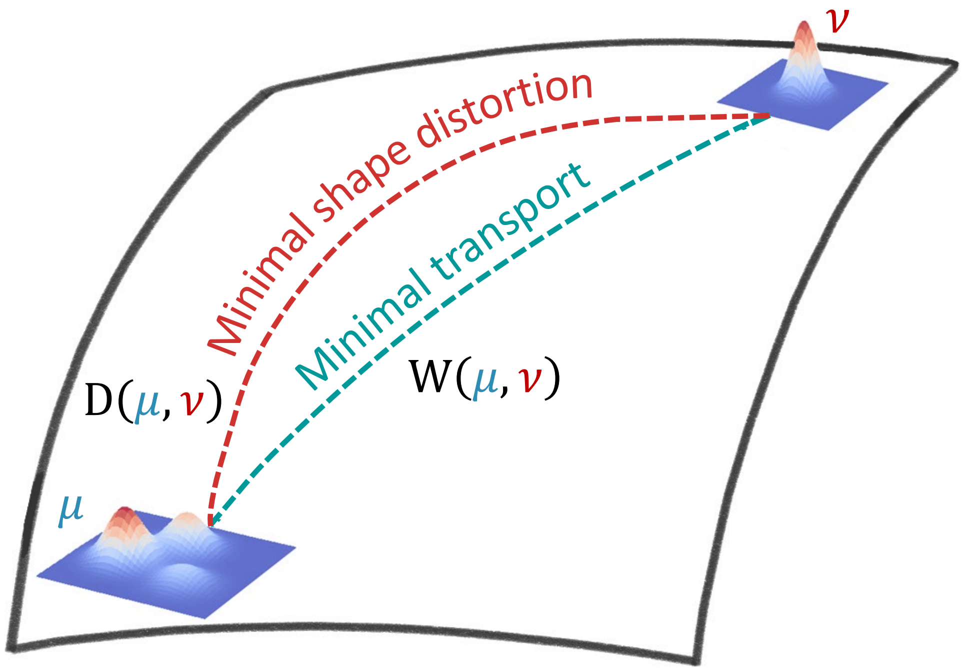
Time = iteration × Sinkhorn
$= k \times O(N^2)$

Mirror Descent
[Scetbon-Peyré-Cuturi '23]

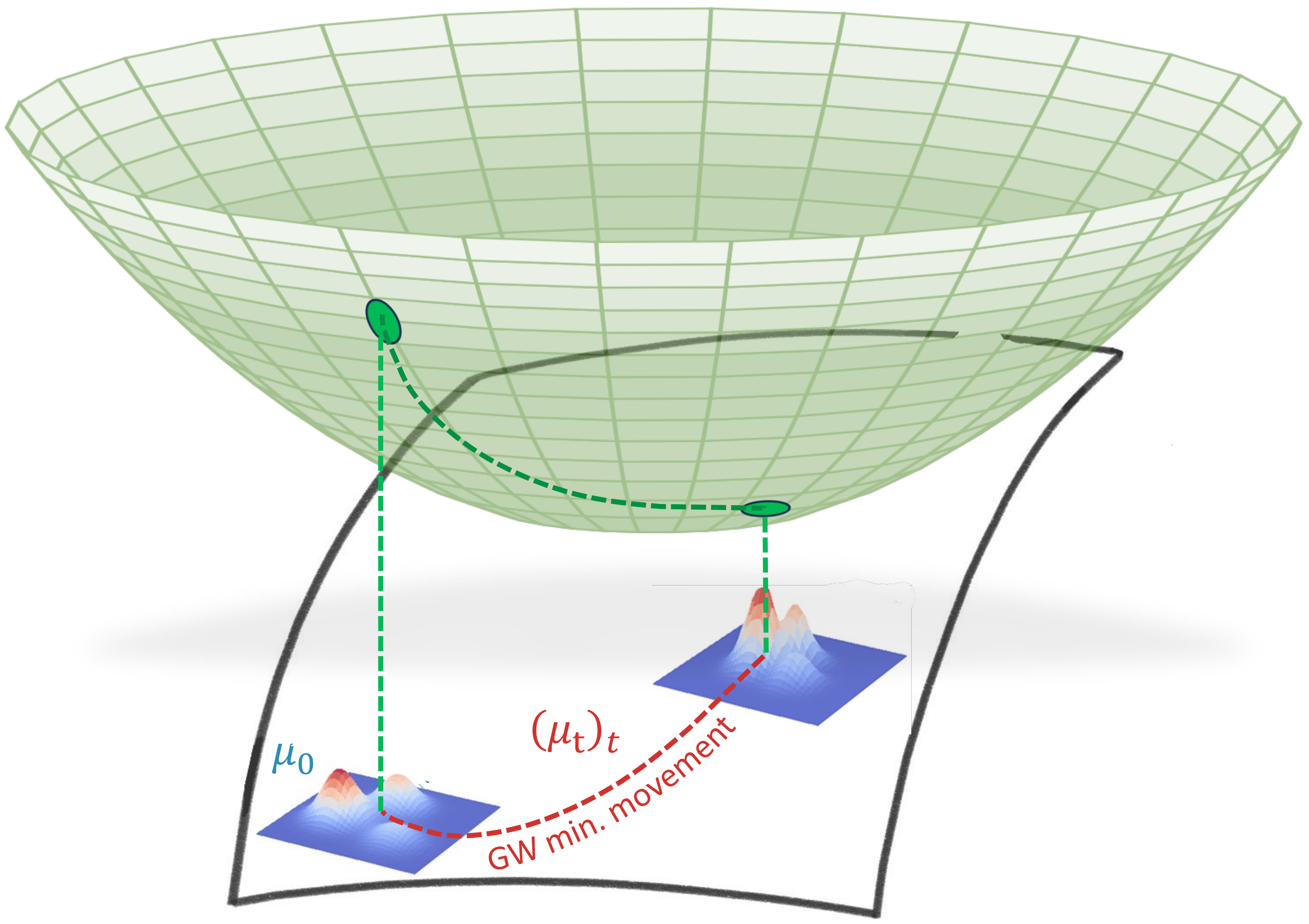


Time = iteration × cost update
$= k \times d \times O(N^2)$

Gromov-Wasserstein Geometry



Gromov-Wasserstein Geometry



Proximal Point Method for GW Gradient Flow

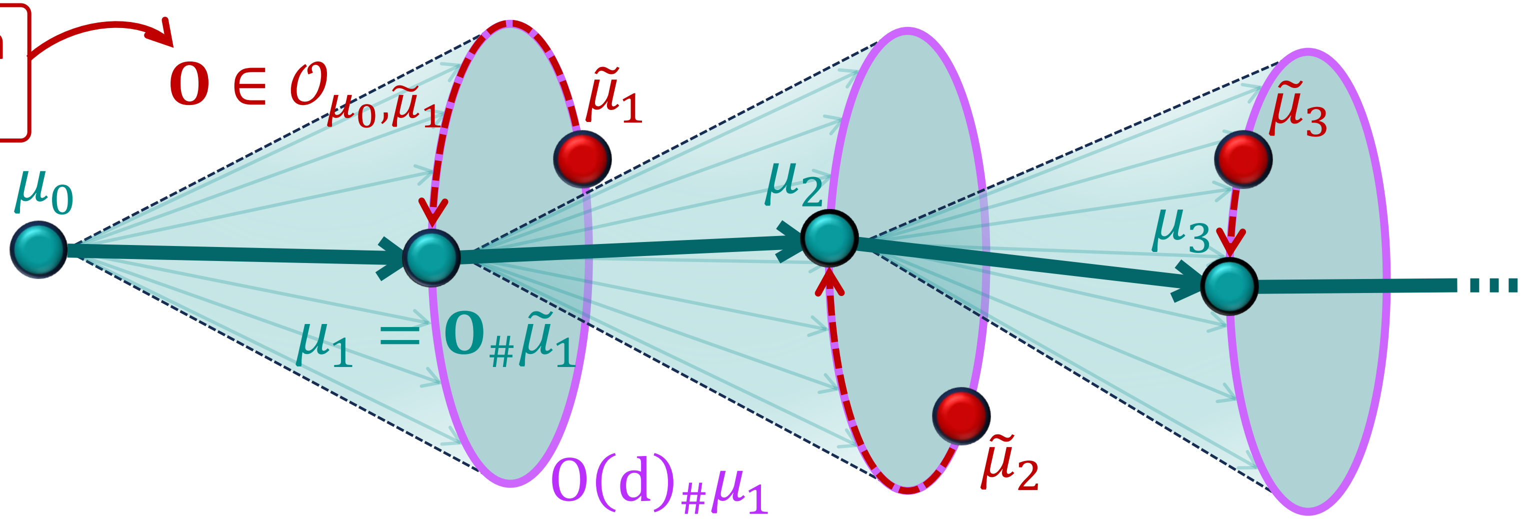
IGW JKO Scheme (Zhang-G.-Greenewald-Mroueh-Sriperumbudur '24)

$$\tilde{\mu}_{t+1} = \underset{\mu}{\operatorname{argmin}} \ \mathcal{F}(\mu) + \frac{\mathrm{IGW}(\mu,\mu_t)^2}{2\tau}, \qquad \mu_{t+1} = \left(\mathbf{O}_{\mu_t,\tilde{\mu}_{t+1}}\right)_{\#}\tilde{\mu}_{t+1}$$

Inner product GW (same space $\mathbb{R}^d$): $\mathrm{IGW}(\mu,\nu)^2 := \inf_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{\pi\otimes\pi}\left[|\langle X,X'\rangle - \langle Y,Y'\rangle|^2\right]$

IGW gradient flow: $\mathcal{F}:\mathcal{P}_2(\mathbb{R}^d)\to\mathbb{R}$ orthogonal invariance and $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$

Covariance transform (aligns steps in W_2)



Convergence of GW Gradient Flow

Theorem (Zhang-G.-Greenewald-Mroueh-Sriperumbudur '24)

The seq. $\mu_0, \mu_1, \ldots$ converges as $\tau \to 0$ to a W_2-continuous curve $(\mu_t)_t \subset \mathcal{P}_2(\mathbb{R}^d)$ with:

$$\begin{cases} \partial_t \mu_t = -\mathrm{div}(\mu_t v_t) \\ v_t = -\mathcal{L}^{-1}_{\boldsymbol{\Sigma}_t, \mu_t}[\nabla \delta \mathcal{F}(\mu_t)] \end{cases}$$

where $\boldsymbol{\Sigma}_t \coloneqq \int x x^T d\mu_t(x)$ and $\mathcal{L}_{\mathbf{A},\mu}[v](x) = \mathbf{A}v(x) + \int_{\mathbb{R}^d} v(y)\langle x, y\rangle d\mu(y)$.

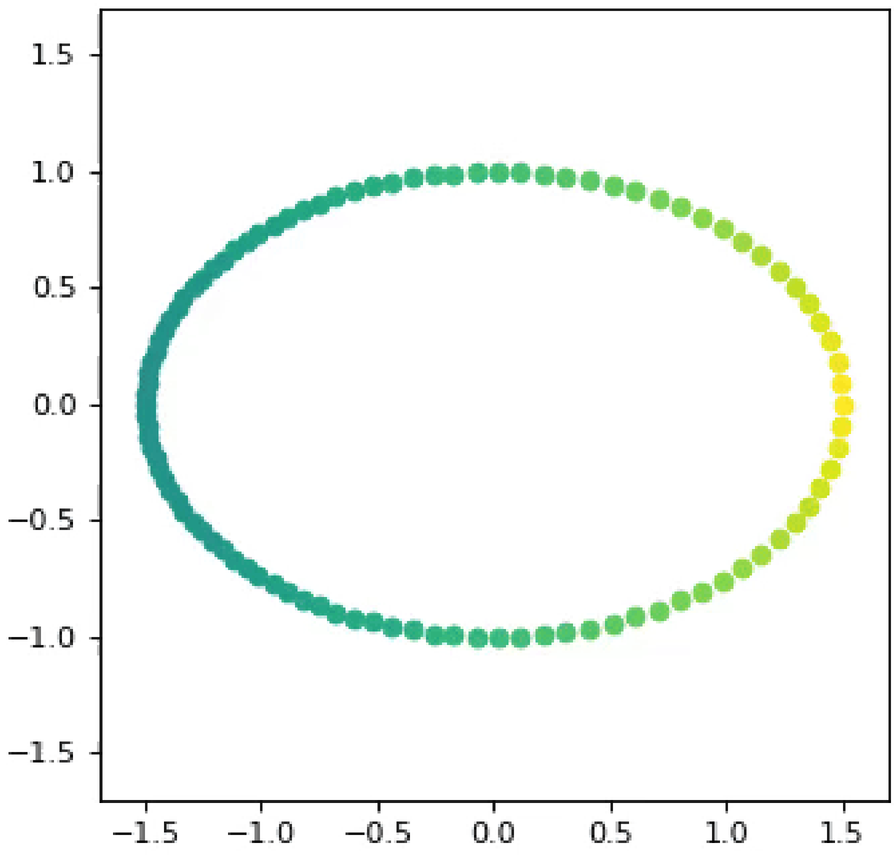
Mobility operator

Comments:

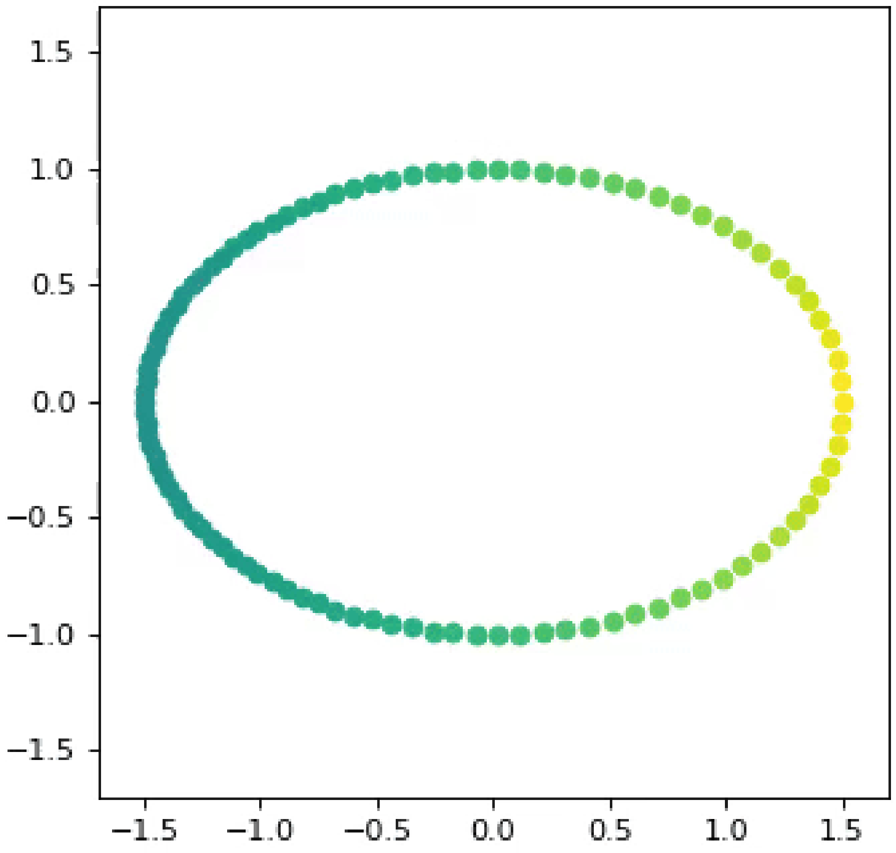
- **Action:** Decreasing functional while maintaining least distortion to global structure
- **Connection to $\mathbf{W_2}$:** $\nabla_{\mathrm{W}_2}\mathcal{F}(\mu) = \nabla \delta \mathcal{F}(\mu) \implies \nabla_{\mathrm{IGW}}\mathcal{F}(\mu) = \mathcal{L}^{-1}_{\boldsymbol{\Sigma},\mu}\left[\nabla_{\mathrm{W}_2}\mathcal{F}(\mu)\right]$
- **Riemannian structure:** $g_\mu(v, w) \coloneqq \left\langle v, \mathcal{L}_{\boldsymbol{\Sigma},\mu}[w] \right\rangle_{L^2(\mu, \mathbb{R}^d)} \implies$ **Benamou-Brenier formula**

IGW Interpolation (Ellipse)

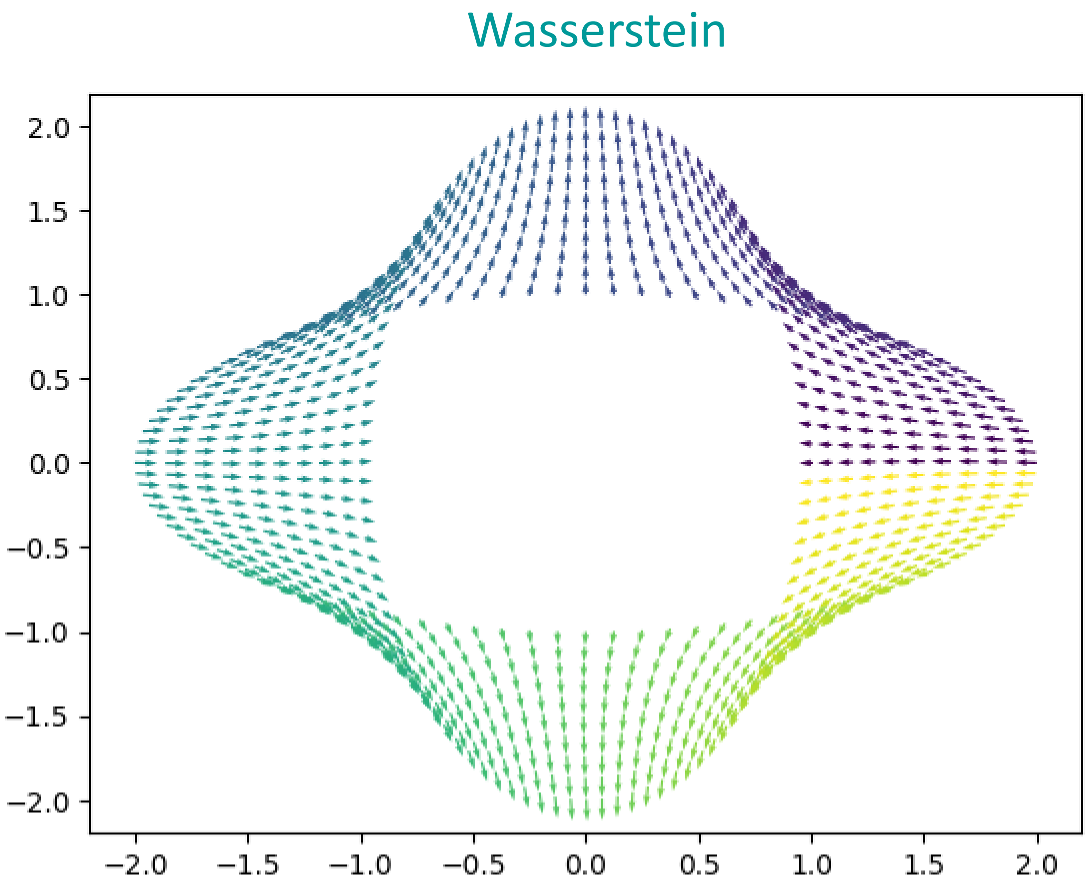
Wasserstein

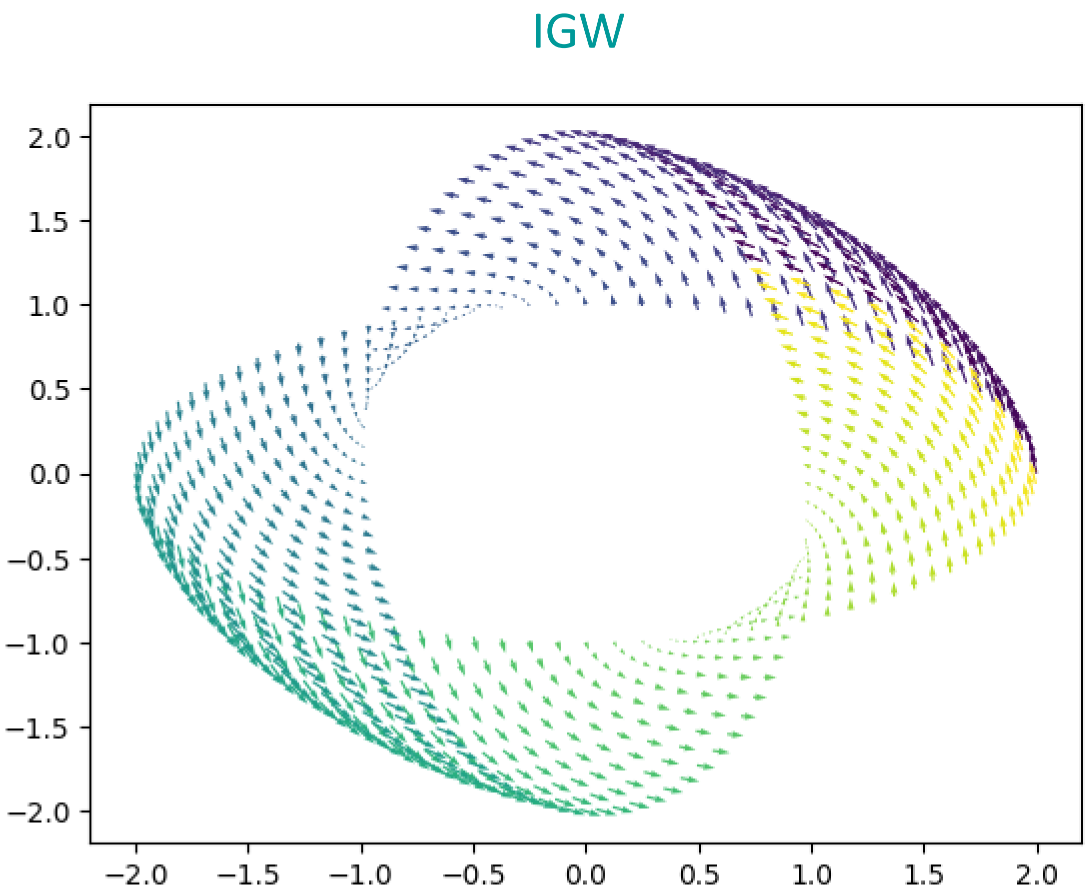


IGW



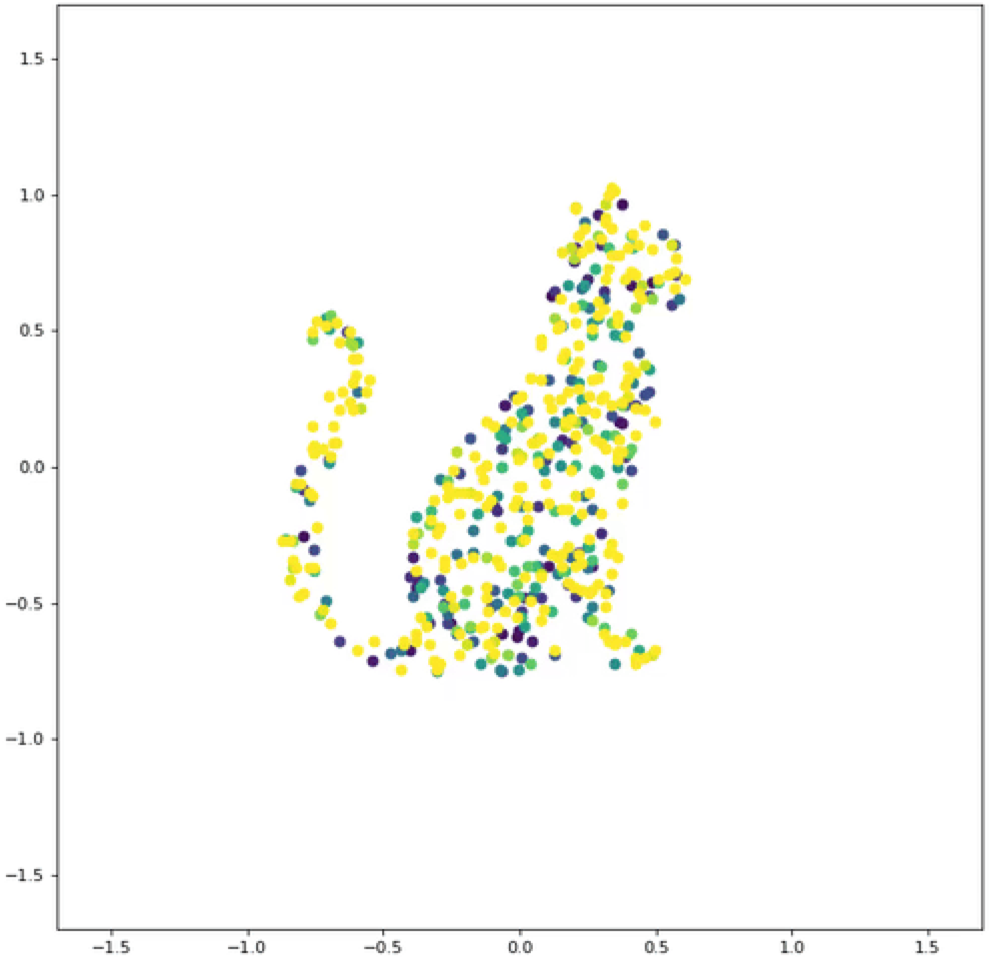
IGW Interpolation (Ellipse)



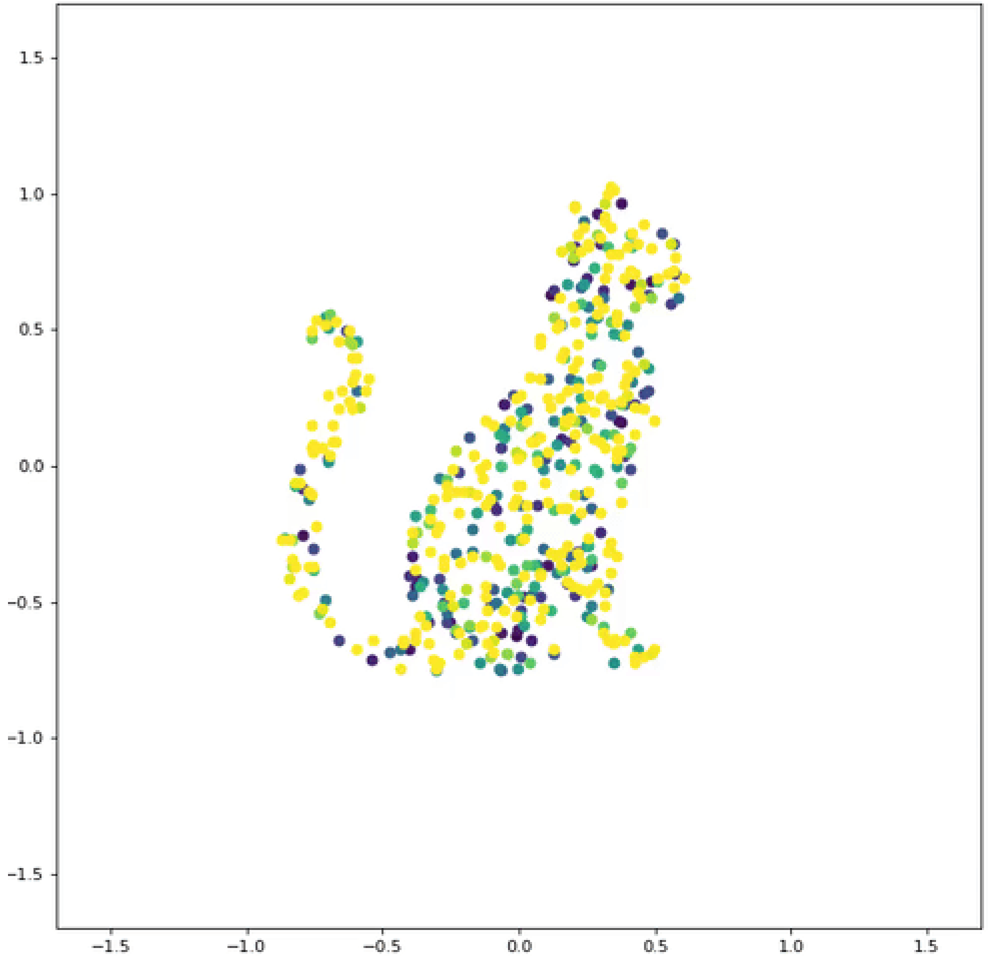


IGW Interpolation (Cat)

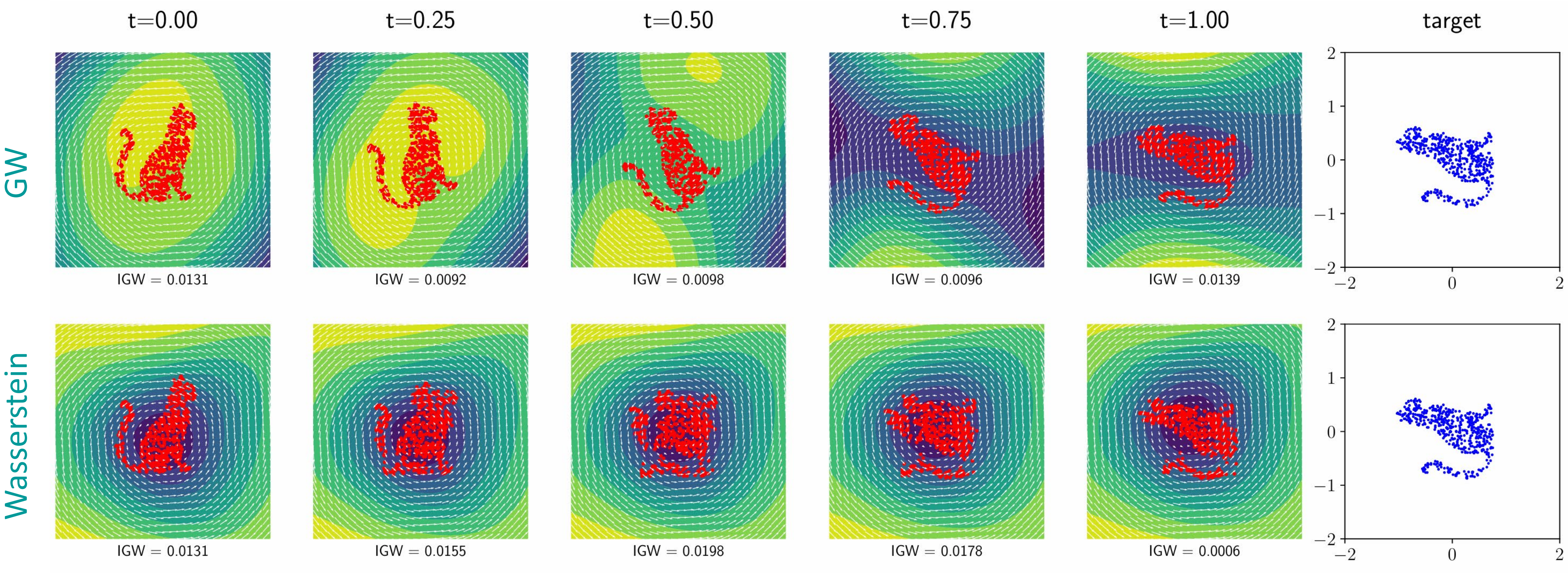
Wasserstein



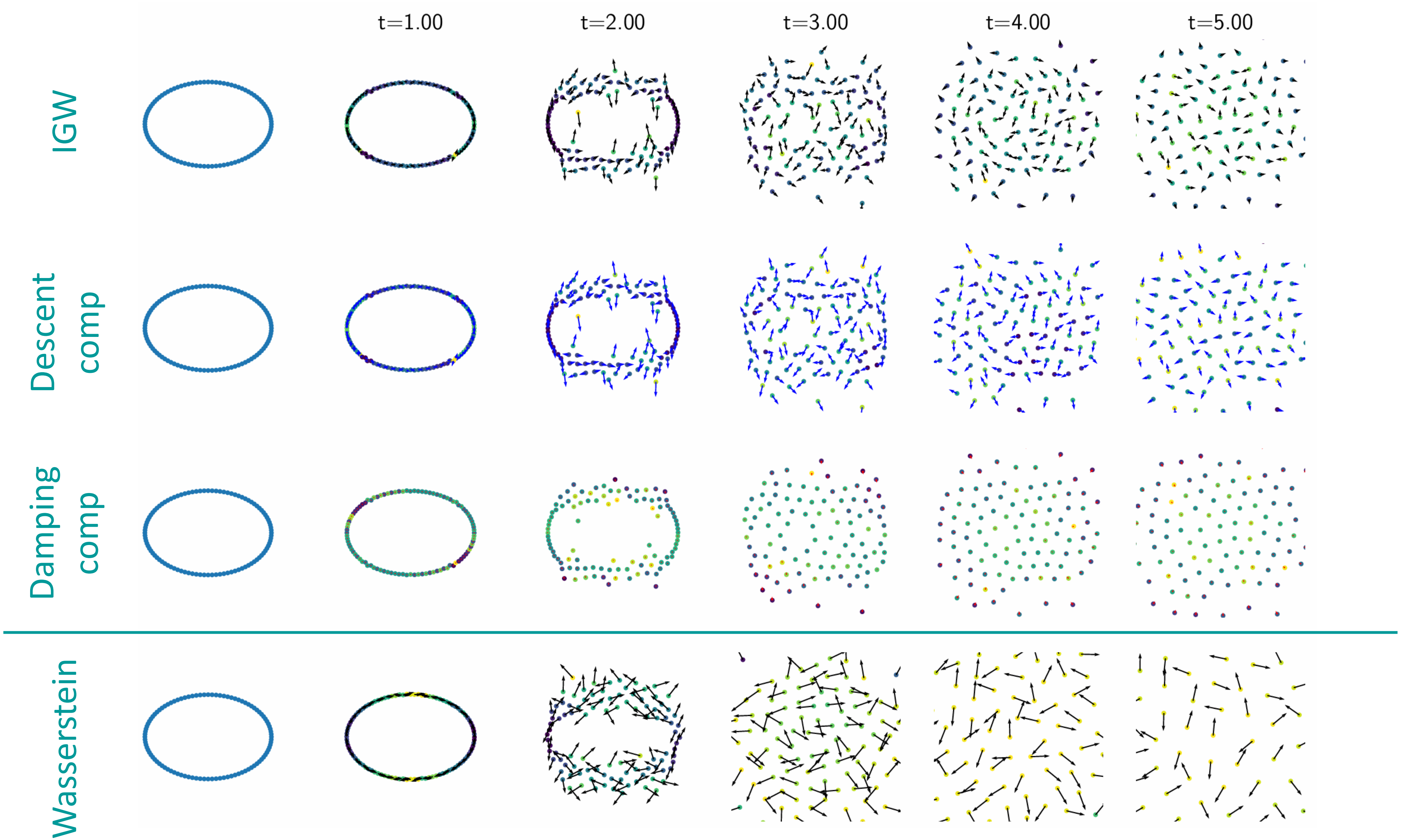
IGW



IGW Interpolation (Cat)



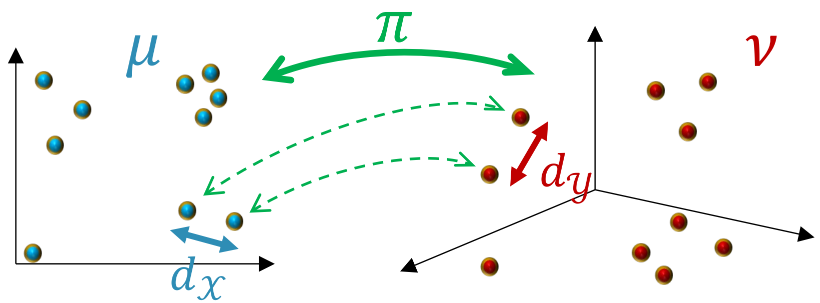
IGW Gradient Flow (Entropy)



Summary

Gromov-Wasserstein Distance: Quantifies discrepancy between mm spaces

- Alignment of hetero. datasets
- Lacking statistical, computational, geometric theory



Contributions: Duality, estimation, algorithms, and geometry

- Variational form that connects to OT
- Sharp empirical rates + limit theorems
- Algorithms w/ convergence rates for entropic GW
- GW gradient flows, Riemannian structure, Benamou-Brenier

Thank you!

[A] Zhang, Goldfeld, Mroueh, Sriperumbudur, “Gromov-Wasserstein distances: entropic regularization, duality, and sample complexity”, Annals of Statistics, ArXiv: 2212.12848

[B] Rioux, Goldfeld, Kato, “Entropic Gromov-Wasserstein distances: stability, algorithms, and distributional limits”, Journal of Machine Learning Research, ArXiv:2306.00182

[C] Zhang, Goldfeld, Greenewald, Mroueh, Sriperumbudur, “Gradient flows and Riemannian structure in the Gromov-Wasserstein geometry”, FoCM, ArXiv: 2407.11800

[D] Rioux, Goldfeld, Kato, “Limit laws for Gromov-Wasserstein alignment with applications to testing graph isomorphisms”, ArXiv:2410.18006