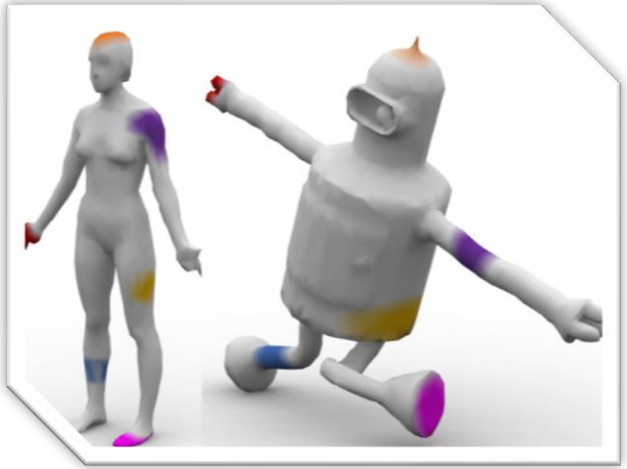
Gromov-Wasserstein Distances: Statistical & Computational Advancements via Duality Theory

Ziv Goldfeld
Cornell University

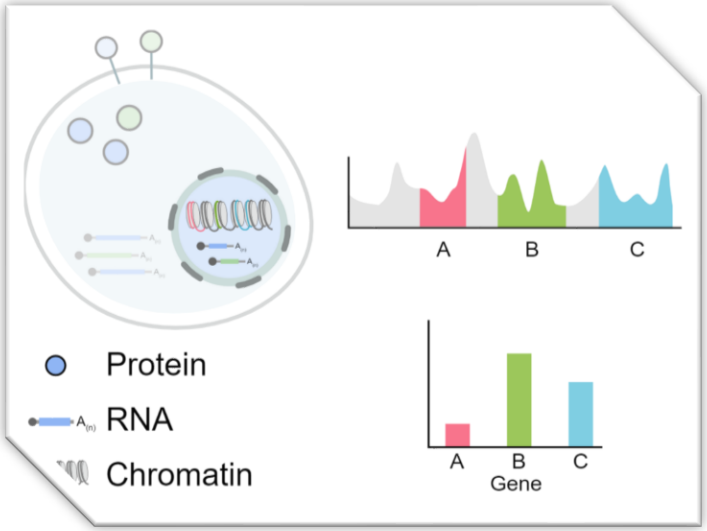
Joint work with:
Zhengxin Zhang, Youssef Mroueh, Bharath Sriperumbudur; Gabriel Rioux, Kengo Kato,

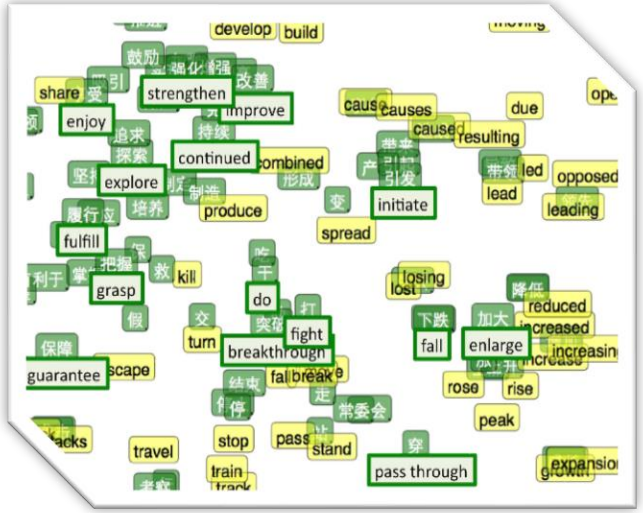
Heterogeneous & Structured Data

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 Distance

Towards the Definition:

- Datasets as metric measure spaces

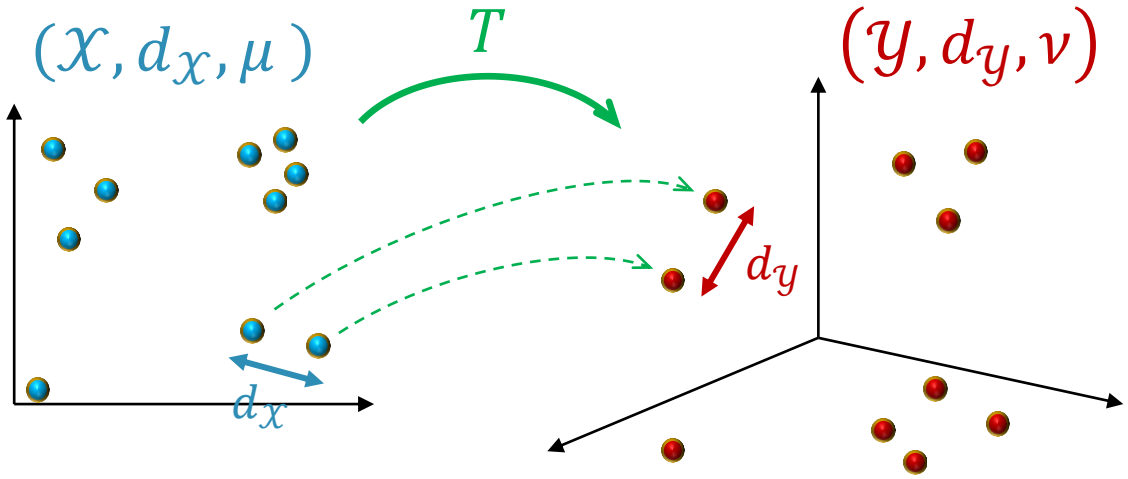
$\Longrightarrow$ $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ & $(\mathcal{Y}, d_{\mathcal{Y}}, \nu)$

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

$\Longrightarrow$ $\nu = T_{\#}\mu$ (if $X \sim \mu$ then $T(X) \sim T_{\#}\mu$)

- Preserve distances (minimize distance distortion)

$\Longrightarrow$ $\|x_i - x_j\| \approx \|T(x_i) - T(x_j)\|$



Gromov-Wasserstein Distance (almost)

The (p, q)-GW distance between mm spaces $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \nu)$ is

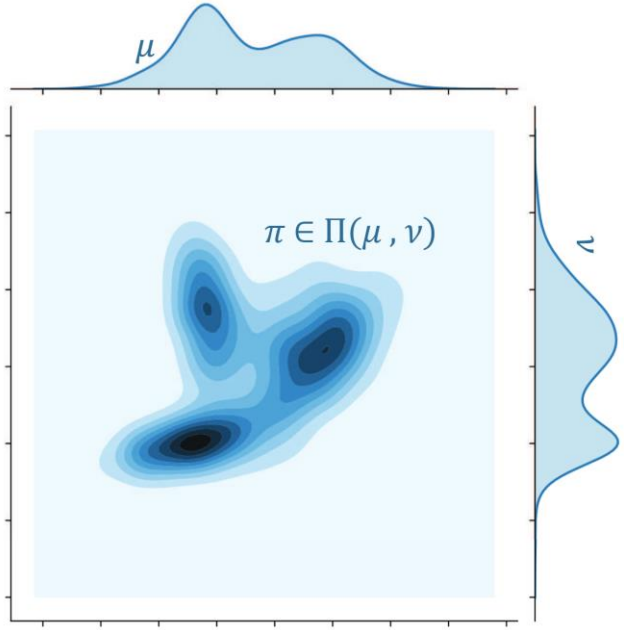
$$\mathrm{D}_{p,q}(\mu, \nu) := \inf_{\substack{T: \mathcal{X} \to \mathcal{Y} \\ T_{\#}\mu = \nu}} \left(\mathbb{E}_{(X,X') \sim \mu \otimes \mu} \left[\left| d_{\mathcal{X}}(X, X')^q - d_{\mathcal{Y}}\big(T(X), T(X')\big)^q \right|^p \right] \right)^{1/p}$$

Gromov-Wasserstein Distance

Gromov-Wasserstein Distance (Memoli '11)

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

Coupling : $\Pi(\mu,\nu) := \{\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}) : \pi(\cdot\times\mathcal{Y}) = \mu, \pi(\mathcal{X}\times\cdot) = \nu\}$



Gromov-Wasserstein Distance

Gromov-Wasserstein Distance (Memoli '11)

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

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

- **Finiteness:** $\mathrm{D}_{p,q}(\mu,\nu) < \infty \; \forall \mu,\nu$ with $\int_{\mathcal{X}\times\mathcal{X}} d_{\mathcal{X}}(x,x')^{pq} d\mu \otimes \mu(x,x') < \infty$ & resp. for ν
- **Identification:** $\mathrm{D}_{p,q}(\mu,\nu) = 0 \iff \exists$ isometry $T: \mathcal{X} \to \mathcal{Y}$ with $T_{\#}\mu = \nu$ (invariances)
- **Metric:** Metrizes space of equivalence classes of mm spaces with finite size
- **Computation:** $\mathrm{D}_{p,q}\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \frac{1}{n}\sum_{i=1}^n \delta_{y_i}\right)^p = \frac{1}{n^2} \min_{\sigma\in S_n} \sum_{i,j=1}^n \left| d_{\mathcal{X}}(x_i,x_j)^q - d_{\mathcal{Y}}(y_{\sigma(i)}, y_{\sigma(j)})^q \right|^p$

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

Entropic Gromov-Wasserstein Distance

Approach: Variants/reformulations of GW problem for computational tractability

- **Sliced GW**: Avg/max of GW btw low-dimensional projections [Vayer-Flamary-Tavenard '20]
- **Unbalanced GW:** Relax marginal constraints via f-div. penalty [Séjourné-Vialard-Peyré '23]
- **Entropic GW:** Add entropic penalty to GW cost [Peyré-Cuturi-Solomon '16]

Entropic Gromov-Wasserstein Distance

$$\mathrm{S}^{\epsilon}_{p,q}(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{\pi\otimes\pi}\left[\left|d_{\mathcal{X}}(X,X')^q - d_{\mathcal{Y}}(Y,Y')^q\right|^p\right] + \epsilon \underbrace{\mathrm{D}_{\mathrm{KL}}(\pi\|\mu\otimes\nu)}_{=\,\mathrm{I}_{\pi}\,(X;Y)}$$

⊛ Computed via mirror-descent w/ Sinkhorn iterations ($\tilde{O}\left(n^2/\epsilon^2\right)$ time, parallelizable)

↳ Convergence to stationary point (asymptotic)

Entropic Gromov-Wasserstein Theory

Open Questions:

1. Convexity regimes in ϵ?
2. Algorithms with (global) convergence rates?
3. Sample complexity for statistical estimation?

Approach: New duality theory to relate EGW to EOT

Duality for Entropic GW Distance

Setting: Quadratic cost over Euclidean spaces

- **mm-spaces:** $\left(\mathbb{R}^{d_x}, \|\cdot\|, \mu\right)$ and $\left(\mathbb{R}^{d_y}, \|\cdot\|, \nu\right)$ with $M_4(\mu) := \int \|x\|^4 d\mu(x)\,, M_4(\nu) < \infty$
- **Quadratic GW:** $\mathrm{S}_\epsilon(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \iint \left|\|x-x'\|^2 - \|y-y'\|^2\right|^2 d\pi \otimes \pi + \epsilon \mathrm{D}_{\mathrm{KL}}(\pi \| \mu \otimes \nu)$

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

$$\mathrm{S}_\epsilon(\mu,\nu) = \mathrm{S}_1(\mu,\nu) + \mathrm{S}_{2,\epsilon}(\mu,\nu)$$

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

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

$\Longrightarrow$ Derive a dual form for $\mathrm{S}_{2,\epsilon}(\mu,\nu)$!

Duality Theory for Entropic GW Distance

Approach: Linearize quadratic term using auxiliary variables

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

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

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

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

Theorem (Zhang-G.-Mroueh-Sriperumbudur '23)

Fix $\epsilon > 0$, $(\mu,\nu) \in \mathcal{P}_4(\mathbb{R}^{d_x}) \times \mathcal{P}_4(\mathbb{R}^{d_y})$, and any $M \ge \sqrt{M_2(\mu)M_2(\nu)}$, we have

$$\mathrm{S}_{2,\epsilon}(\mu,\nu) = \inf_{\mathbf{A}\in\mathcal{D}_M} 32\|\mathbf{A}\|_{\mathrm{F}}^2 + \mathrm{EOT}_{\epsilon,c_{\mathbf{A}}}(\mu,\nu)$$

Sample Complexity of Entropic GW

Setting: Data $\Longrightarrow$ $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ **&** $\hat{\nu}_n := \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ $\Longrightarrow$ $\mathrm{S}_\epsilon(\mu,\nu) \approx \mathrm{S}_\epsilon(\hat{\mu}_n,\hat{\nu}_n)$?

Theorem (Zhang-G.-Mroueh-Sriperumbudur '23)

Fix $\epsilon > 0$ and let $(\mu,\nu) \in \mathcal{P}(\mathbb{R}^{d_x}) \times \mathcal{P}(\mathbb{R}^{d_y})$ be 4-sub-Weibull with parameter. $\sigma^2 > 0$. Then

$$\mathbb{E}[|\mathrm{S}_\epsilon(\mu,\nu) - \mathrm{S}_\epsilon(\hat{\mu}_n,\hat{\nu}_n)|] \lesssim_{d_x,d_y} \underbrace{\frac{1+\sigma^4}{\sqrt{n}}}_{\substack{\mathrm{S}_1 \text{ rate +} \\ \text{centering bias}}} + \underbrace{\epsilon\left(1+\left(\frac{\sigma}{\sqrt{\epsilon}}\right)^{9\lceil (d_x \vee d_y)/2\rceil + 11}\right)\frac{1}{\sqrt{n}}}_{\mathrm{S}_{2,\epsilon} \text{ rate}}$$

Comments:

- **Optimality:** Rate is parametric and hence minimax optimal
- **Entropic OT:** Rate matches that for EOT (assuming compact support or sub-Gaussianity)
- **One-sample:** When only μ is estimated, rate is similar but with d_x instead of $d_x \vee d_y$

Sample Complexity of Entropic GW: Proof Outline

Decomposition: Split S_ϵ into $\mathrm{S}_1 + \mathrm{S}_{2,\epsilon}$ and center empirical measures

$$\mathbb{E}[|\mathrm{S}_\epsilon(\mu,\nu) - \mathrm{S}_\epsilon(\hat{\mu}_n,\hat{\nu}_n)|] \leq \mathbb{E}[|\mathrm{S}_1(\mu,\nu) - \mathrm{S}_1(\hat{\mu}_n,\hat{\nu}_n)|] + \mathbb{E}\left[\left|\mathrm{S}_{2,\epsilon}(\mu,\nu) - \mathrm{S}_{2,\epsilon}(\hat{\mu}_n,\hat{\nu}_n)\right|\right] + \frac{\sigma^2}{\sqrt{n}}$$

S_1 Analysis: Involves only estimation of moments $\Longrightarrow$ Rate is parametric $\asymp 1/\sqrt{n}$

$\mathrm{S}_{2,\epsilon}$ Analysis: Hinges on dual form + regularity analysis of optimal potentials

1. **EOT reduction:** $\left|\mathrm{S}_{2,\epsilon}(\mu,\nu) - \mathrm{S}_{2,\epsilon}(\hat{\mu}_n,\hat{\nu}_n)\right| \leq \sup_{\mathbf{A}\in\mathcal{D}_M} \left|\mathrm{EOT}_{\epsilon,c_{\mathbf{A}}}(\mu,\nu) - \mathrm{EOT}_{\epsilon,c_{\mathbf{A}}}(\hat{\mu}_n,\hat{\nu}_n)\right| \circledast$
2. **Dual potentials:** $\forall \mathbf{A} \in \mathcal{D}_M,\ (\varphi_{\boldsymbol{A}},\psi_{\boldsymbol{A}}) \in \mathcal{F}_s \times \mathcal{G}_s$ for Hölder classes of arbitrary smoothness
3. **Empirical processes:** $\mathbb{E}[\circledast] \leq \mathbb{E}\left[\sup_{\varphi\in\mathcal{F}_s} |(\mu - \hat{\mu}_n)\varphi|\right] + \mathbb{E}\left[\sup_{\psi\in\mathcal{G}_s} |(\mu - \hat{\mu}_n)\psi|\right] \lesssim 1/\sqrt{n}$

Bound Dudley entropy integral of $\mathcal{F}_s$ and $\mathcal{G}_s$ (Hölder) with $s = \left\lceil \frac{d_x}{2} \right\rceil + 1$

From Stability Analysis to Convexity

$$\boxed{\mathrm{S}_\epsilon(\mu,\nu) = \mathrm{S}_1(\mu,\nu) + \min_{\mathbf{A}\in\mathcal{D}_M}\Big\{\underbrace{32\|\mathbf{A}\|_{\mathrm{F}}^2 + \mathrm{EOT}_{\epsilon,c_{\mathbf{A}}}(\mu,\nu)}_{=:\,\Phi(\mathbf{A})}\Big\}}$$

Analysis:

- Fréchet derivatives $D\Phi_{[\mathbf{A}]}$ and $D^2\Phi_{[\mathbf{A}]}$
- Bound $\lambda_{max}\big(D^2\Phi_{[\mathbf{A}]}\big) \le 64$ **&** $\lambda_{min}\big(D^2\Phi_{[\mathbf{A}]}\big) \ge 32^2\epsilon^{-1}\sqrt{M_4(\mu)M_4(\nu)} - 64$

Theorem (Rioux-G.-Kato '23)

1. For $M \ge \sqrt{M_2(\mu)M_2(\nu)}$, all minimizers of Φ are in $\mathcal{D}_M$
2. Φ is strictly convex whenever $\epsilon > 16\sqrt{M_4(\mu)M_4(\nu)}$
3. Φ is L-smooth on $\mathcal{D}_M$ with $L \le 64 \vee \left(32^2\epsilon^{-1}\sqrt{M_4(\mu)M_4(\nu)} - 64\right)$

First-Order Inexact Oracle Methods

$$\min_{\mathbf{A}\in\mathcal{D}_M} 32\|\mathbf{A}\|_{\mathrm{F}}^2 + \mathrm{EOT}_{\epsilon,c_{\mathbf{A}}}(\mu,\nu)$$

First-order methods: 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): $\widetilde{\Pi}^{\mathbf{A}}$ s.t. $\left\|\Pi^{\mathbf{A}} - \widetilde{\Pi}^{\mathbf{A}}\right\|_\infty \leq \delta$

- Gradient approximation $\widetilde{D}\Phi_{[\mathbf{A}]}$ ($\widetilde{\Pi}^{\mathbf{A}}$ instead of $\Pi^{\mathbf{A}}$)
- First-order method under convexity [d'Aspremont '08]

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

Algorithm 1 Fast gradient method with inexact oracle

Fix $L = 64$ and let $\alpha_k = \frac{k+1}{2}$, and $\tau_k = \frac{2}{k+3}$

$k \leftarrow 0$
$\boldsymbol{A}_0 \leftarrow \mathbf{0}$
$\boldsymbol{G}_0 \leftarrow \widetilde{D}\Phi_{[\boldsymbol{A}_0]}$
$\boldsymbol{W}_0 \leftarrow \alpha_0 \boldsymbol{G}_0$
while stopping condition is not met **do**
  $\boldsymbol{B}_k \leftarrow \frac{M}{2}\mathrm{sign}(\boldsymbol{A}_k - L^{-1}\boldsymbol{G}_k)\min\left(\frac{2}{M}\left|\boldsymbol{A}_k - L^{-1}\boldsymbol{G}_k\right|, 1\right)$
  $\boldsymbol{C}_k \leftarrow \frac{M}{2}\mathrm{sign}(-L^{-1}\boldsymbol{W}_k)\min\left(\frac{2}{M}\left|L^{-1}\boldsymbol{W}_k\right|, 1\right)$
  $\boldsymbol{A}_{k+1} \leftarrow \tau_k \boldsymbol{C}_k + (1-\tau_k)\boldsymbol{B}_k$
  $\boldsymbol{G}_{k+1} \leftarrow \widetilde{D}\Phi_{[\boldsymbol{A}_{k+1}]}$
  $\boldsymbol{W}_{k+1} \leftarrow \boldsymbol{W}_k + \alpha_{k+1}\boldsymbol{G}_{k+1}$
  $k \leftarrow k+1$
return $\boldsymbol{B}_k$

Global Convergence Guarantees (Convex)

$$\min_{\mathbf{A}\in\mathcal{D}_M} 32\|\mathbf{A}\|_{\mathrm{F}}^2 + \mathrm{EOT}_{\epsilon,c_{\mathbf{A}}}(\mu,\nu)$$

Theorem (Rioux-G.-Kato '23)

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

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

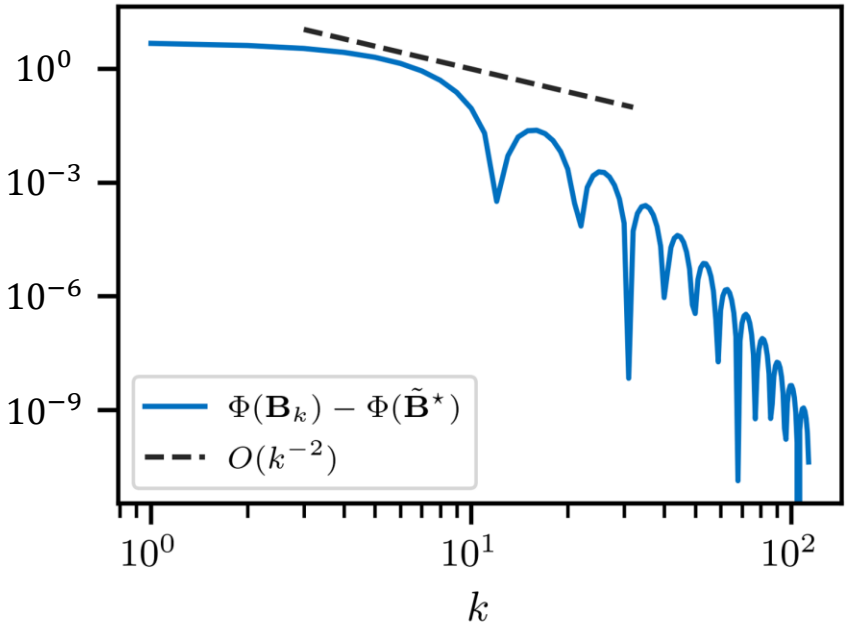
Comments:

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

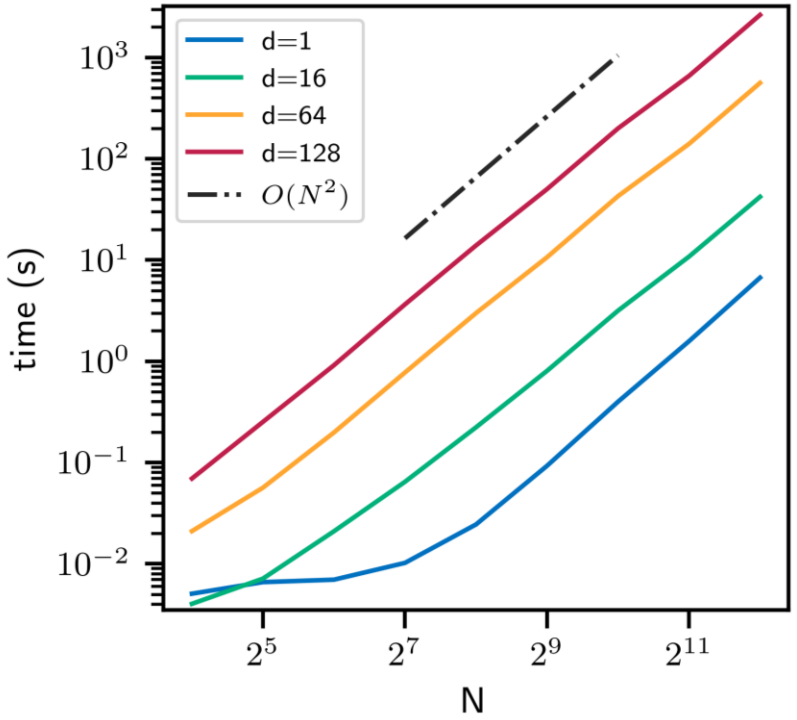
  ↳ Adapts to convexity of Φ (yields improved rates if convex)

Numerical Results

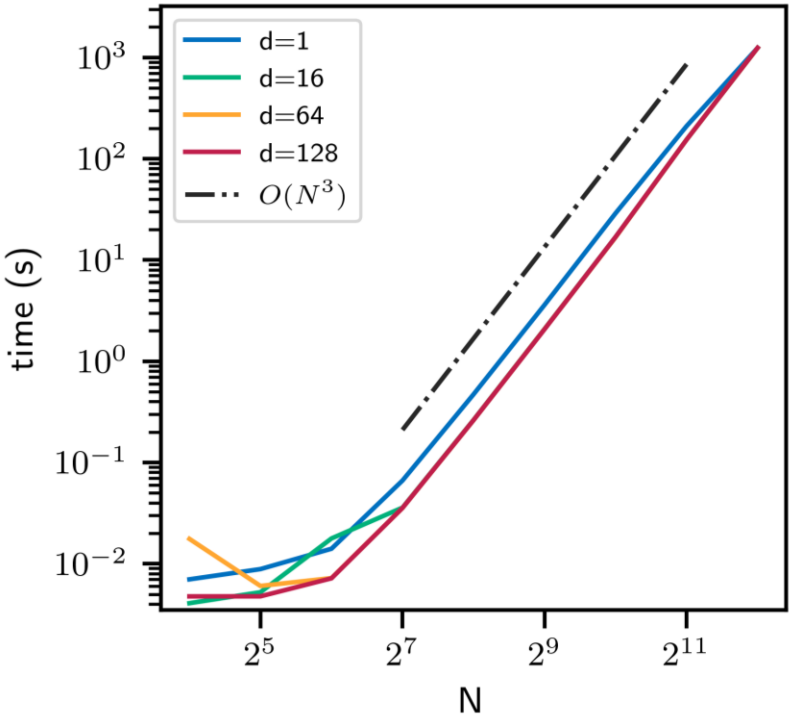
Convergence Rate



Fast Gradient Method
[Rioux-G.-Kato '26]



Mirror Descent
[Peyré-Cuturi-Solomon '16]



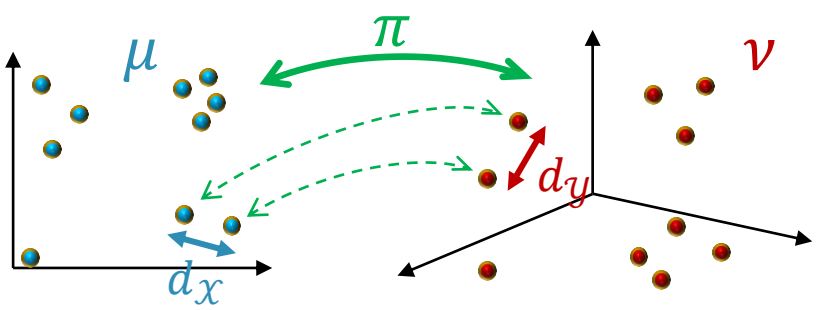
Solves in matrix space:
$d_x \times d_y$ size

Solves in coupling space:
$N_x \times N_y$ size

Summary

Gromov-Wasserstein Distance: Quantifies discrepancy between mm spaces

- Applications in ML and beyond for heterogeneous data
- Foundational statistical **&** computational questions open



Contributions: Duality, empirical rates, and algorithms

- Dual form that connects to EOT
- First sample complexity result for EGW (quadratic cost over Euclidean spaces)
- First algorithms with convergence rates (global optimality under convexity)
- Duality and empirical rates also derived for non-entropic GW

[A] Zhang, Goldfeld, Mroueh, Sriperumbudur, "Gromov-Wasserstein distances: entropic regularization, duality, and sample complexity", ArXiv: 2212.12848

[B] Rioux, Goldfeld, Kato, "Entropic Gromov-Wasserstein distances: stability, algorithms, and distributional limits", in prep.

Thank you!