Gromov-Wasserstein Distances: Entropic Regularization, Duality, and Sample Complexity

Ziv Goldfeld
Cornell University

Information Theory and Applications Workshop, February 13th, 2023
Heterogeneous & Structured Data

Dataset Matching: Various applications require matching heterogeneous & structured datasets

Goals:
1. Compare how similar/different two datasets are
2. Obtain matching/alignment that respects individual structure
The Gromov-Wasserstein Distance

Object matching: Correspondence between geometric objects

• Represent objects as metric measure spaces
  \((X, d_X, \mu) \quad \& \quad (Y, d_Y, \nu)\)

• Find matching (transport map) \(T: X \rightarrow Y\)
  \(\nu = T_{\#}\mu\) (if \(X \sim \mu\) then \(T(X) \sim T_{\#}\mu\))

• Preserve distances (minimize distortion)
  \[\|x_i - x_j\| \approx \|T(x_i) - T(x_j)\|\]

Gromov-Wasserstein Distance (Memoli '11)

The \((p, q)\)-GW distance btw. mm spaces \((X, d_X, \mu)\) and \((Y, d_Y, \nu)\) is

\[
D_{p,q}(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \left( \int_{X \times Y} \int_{X \times Y} |d_X(x, x')^q - d_Y(y, y')^q|^p \, d\pi \otimes \pi(x, y, x', y') \right)^{1/p}
\]
The Gromov-Wasserstein Distance

**Gromov-Wasserstein Distance (Memoli ‘11)**

\[ D_{p,q}(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \left( \int_{X \times Y} \int_{X \times Y} |d_X(x, x')^q - d_Y(y, y')^q|^p \, d\pi \otimes \pi(x, y, x', y') \right)^{1/p} \]

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

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

\(\bullet\) Quadratic assignment problem (non-convex) [Commander ‘05] \(\iff\) **NP complete**
Entropic GW vs. Computational Hardness

**Approach:** Explore variants of the 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**

$$S_{p,q}^\varepsilon (\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \iint |d_X(x,x')^q - d_Y(y,y')^q|^p \, d\pi \otimes \pi(x,y,x',y') + \varepsilon D_{KL}(\pi \| \mu \otimes \nu)$$

- Computed via mirror-descent w/ Sinkhorn iterations [Solomon et al ‘16]
- Sinkhorn algorithm time complexity is $\tilde{O}(n^2/\varepsilon^2)$ (highly parallelizable) [Lin et al ‘22]
A Statistical Question

**Question:** $\mu, \nu$ 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 $\hat{\nu}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{Y_i}$

  Can we approximate $D_{p,q}(\mu, \nu) \approx D_{p,q}(\hat{\mu}_n, \hat{\nu}_n)$?

  $\Rightarrow$ $S_{p,q}^\varepsilon(\mu, \nu) \approx S_{p,q}^\varepsilon(\hat{\mu}_n, \hat{\nu}_n)$?

**Asymptotic Ans:** Yes! For $\mu, \nu$ w/ finite $pq$–size, $D_{p,q}(\hat{\mu}_n, \hat{\nu}_n) \to D_{p,q}(\mu, \nu)$ a.s. [Mémoli ’11]

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

- **Open question:** No available results for either $D_{p,q}$ or $S_{p,q}^\varepsilon$

  **Statistical implications:** Principled sample-size selection + further stat. advancements

  **Computational implications:** Time complexity depends on sample size
Duality Theory for (Entropic) GW Distance

Setting: $(2,2)$-cost over Euclidean spaces

- **mm-spaces:** $(\mathbb{R}^d_x, \|\cdot\|, \mu)$ and $(\mathbb{R}^d_y, \|\cdot\|, \nu)$ with $M_4(\mu) := \int \|x\|^4 d\mu(x), M_4(\nu) < \infty$

- **Quadratic GW:**
  \[ S_\epsilon(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int \|x - x'\|^2 - \|y - y'\|^2 \right|^2 d\pi \otimes \pi + \epsilon D_{KL}(\pi\|\mu \otimes \nu) \]

Decomposition: Assume wlog that $\mu, \nu$ are centered (invariance to translation); then

\[ S_\epsilon(\mu, \nu) = S_1(\mu, \nu) + S_{2,\epsilon}(\mu, \nu) \]

where

\[ S_1(\mu, \nu) = \int \|x - x'\|^4 d\mu \otimes \mu + \int \|y - y'\|^4 d\nu \otimes \nu - 4 \int \|x\|^2 \|y\|^2 d\mu \otimes \nu \]

\[ S_{2,\epsilon}(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int -4\|x\|^2 \|y\|^2 d\pi - 8 \sum_{1 \leq i \leq d_x} \left( \int x_i y_j d\pi \right)^2 + \epsilon D_{KL}(\pi\|\mu \otimes \nu) \]

Derive a dual form for $S_{2,\epsilon}(\mu, \nu)$ by linearizing quadratic term!
Duality Theory for (Entropic) GW Distance

**Approach:** Linearize quadratic term using auxiliary variables

\[
S_{2,\epsilon}(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int -4\|x\|^2\|y\|^2 d\pi - 8 \sum_{1 \leq i \leq d_x} (\int x_i y_j d\pi)^2 + \epsilon D_{KL}(\pi \| \mu \otimes \nu)
\]

Optimality at \(a_{i,j}(\pi) = 0.5 \int x_i y_j d\pi\)

and define

\(M_{\mu, \nu} = \sqrt{M_2(\mu)M_2(\nu)}\)

\[
\begin{align*}
S_{2,\epsilon}(\mu, \nu) &= \inf_{\pi \in \Pi(\mu, \nu)} \int -4\|x\|^2\|y\|^2 d\pi - 8 \sum_{1 \leq i \leq d_x} \frac{\inf_{1 \leq j \leq d_y} \frac{M_{\mu, \nu}}{2} \leq a_{ij} \leq \frac{M_{\mu, \nu}}{2}}{(a_{i,j}^2 - \int a_{ij} x_i y_j d\pi)} + \epsilon D_{KL}(\pi \| \mu \otimes \nu) \\
&= \inf_{A \in D_{M_{\mu, \nu}}} 32\|A\|_F^2 + \inf_{\pi \in \Pi(\mu, \nu)} \int \left(-4\|x\|^2\|y\|^2 - 32^T Ay\right) d\pi + \epsilon D_{KL}(\pi \| \mu \otimes \nu) \\
&= c_A(x, y) = EOT_{\epsilon, c_A}(\mu, \nu)
\end{align*}
\]

**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 \(M \geq \sqrt{M_2(\mu)M_2(\nu)},\) we have

\[
S_{2,\epsilon}(\mu, \nu) = \inf_{A \in D_M} 32\|A\|_F^2 + \sup_{(\varphi, \psi) \in L^1(\mu) \times L^1(\nu)} \int \varphi d\mu + \int \psi d\nu - \epsilon \int \frac{\varphi(x) + \psi(y) - c_A(x, y)}{\epsilon} d\mu \otimes \nu + \epsilon
\]
Sample Complexity of Entropic GW

**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 param. $\sigma^2 > 0$. Then

$$
\mathbb{E}[|S_\epsilon(\mu, \nu) - S_\epsilon(\hat{\mu}_n, \hat{\nu}_n)|] \lesssim d_x, d_y \frac{1 + \sigma^4}{\sqrt{n}} + \epsilon \left( 1 + \left( \frac{\sigma}{\sqrt{\epsilon}} \right)^{9[(d_x \vee d_y)/2]+11} \right) \frac{1}{\sqrt{n}}
$$

**Comments:**

- **Optimality:** Rate is parametric and hence minimax optimal
- **Entropic OT:** Rate is matches that for EOT (assuming compact support or sub-Gaussianity)
- **Constants:** May not be optimal but matches best known dependence on $\epsilon, \sigma$ for EOT
- **One-sample:** When only $\mu$ 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 $S_\epsilon$ into $S_1 + S_{2,\epsilon}$ and center empirical measures

\[
\mathbb{E}[|S_\epsilon(\mu, \nu) - S_\epsilon(\hat{\mu}_n, \hat{\nu}_n)|] \leq \mathbb{E}[|S_1(\mu, \nu) - S_1(\hat{\mu}_n, \hat{\nu}_n)|] + \mathbb{E}[|S_{2,\epsilon}(\mu, \nu) - S_{2,\epsilon}(\hat{\mu}_n, \hat{\nu}_n)|] + \frac{\sigma^2}{\sqrt{n}}
\]

$S_1$ Analysis: Involves only estimation of moments

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

1. EOT reduction: $|S_{2,\epsilon}(\mu, \nu) - S_{2,\epsilon}(\hat{\mu}_n, \hat{\nu}_n)| \leq \sup_{A \in \mathcal{D}_M} \text{EOT}_{\epsilon, c_A}(\mu, \nu) - \text{EOT}_{\epsilon, c_A}(\hat{\mu}_n, \hat{\nu}_n)$

2. Potentials: For each $A \in \mathcal{D}_M$:

\[
|\varphi(x)| \leq C_{d_x, d_y}(1 + \bar{\sigma}^5)(1 + \|x\|^4)
\]

\[
|D^\alpha \varphi(x)| \leq C_{\alpha, d_x, d_y}(1 + \bar{\sigma}^{4.5|\alpha|})(1 + \|x\|^3|\alpha|), \forall \alpha \in \mathbb{N}_0^{d_x}
\]

$\forall A \in \mathcal{D}_M$ $(\varphi^*, \psi^*) \in \mathcal{F}_S \times \mathcal{G}_S$ for Hölder classes of arbitrary smoothness

3. Reduction to emp. Process: $\mathbb{E}[\star] \leq \mathbb{E}\left[\sup_{\varphi \in \mathcal{F}_S} |(\mu - \hat{\mu}_n)\varphi|\right] + \mathbb{E}\left[\sup_{\psi \in \mathcal{G}_S} |(\mu - \hat{\mu}_n)\psi|\right]$  

$\longrightarrow$ Partition $\mathbb{R}^{d_x}$ into compact shells & bound ent. integral of $\mathcal{F}_S$ (Hölder) with $s = \left\lfloor \frac{d_x}{2} \right\rfloor + 1$
Proof: Same argument as before but apply **standard OT duality** in the last step

\[ S_{2, \varepsilon}(\mu, \nu) = \ldots = \inf_{A \in D_M_{\mu, \nu}} 32\|A\|_F^2 + \left\{ \inf_{\pi \in \Pi(\mu, \nu)} \int \left( -4\|x\|^2\|y\|^2 - 32x^TAy \right) d\pi \right\} = \text{OT}_{c_A}(\mu, \nu) \]
Standard GW: Duality & Sample Complexity

**Theorem (Zhang-G.-Mroueh-Sriperumbudur ‘23)**

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

$$
\mathbb{E}[|D(\mu, \nu)^2 - D(\hat{\mu}_n, \hat{\nu}_n)^2|] \leq d_x, d_y, R \frac{R^4}{\sqrt{n}} + (1 + R^4)n^{-\frac{2}{d_x v d_y \sqrt{4}}} (\log n)^{\mathbb{1}\{d_x v d_y = 4\}}
$$

**Proof:** Similar argument using Lipschitness & concavity of optimal potentials (via cost concavity)

- **Low dimension:** Potential class is Donsker for $d_x \lor d_y \leq 3$ [Hundrieser et al ‘22]

**Comments:**

- **OT:** Rate is matches that for empirical OT with compact support [Manole-Niles Weed ‘22]
- **Unbdd. support:** [Manole-Niles Weed ‘22] have argument for OT under strong assumptions
- **Non-squared GW:** If $D(\mu, \nu) > 0$ then the same rates hold for empirical $D$ itself
- **One-sample:** When only $\mu$ is estimated, rate is similar but with $d_x$ instead of $d_x \lor d_y$
**Summary**

**Gromov-Wasserstein Distance:** Quantifies discrepancy between metric spaces
- Applications in ML and beyond for heterogeneous data
- Foundational statistical & computational questions open

**Contributions:** Duality and first steps towards statistical theory
- Dual form using auxiliary matrix-valued variable
- First sample complexity results for GW and EGW (quadratic cost over Euclidean spaces)
- Additional results: stability or GW cost and coupling in regular parameter

**Directions:** New optimization algorithms, limit distribution theory, GW gradient flow, etc.


**Thank you!**