The Unbalanced Gromov Wasserstein Distance: Conic Formulation and Relaxation

Thibault Séjourné, François-Xavier Vialard, Gabriel Peyré

2021.3.26

Outlines

Background

- Comparing data distributions on different metric spaces is a basic problem in machine learning.
- The most popular distance between such metric measure spaces is the Gromov-Wasserstein (GW) distance.
- Limitations
  - The GW distance is limited to the comparison of metric measure spaces endowed with a **probability** distribution. This strong limitation is problematic for many applications in ML.
  - Imposing an exact conservation of mass across spaces is not robust to outliers and often leads to irregular matching.
- To behave better wrt mass variation and outliers → unbalanced OT.

Metric Measure Spaces

A mm-space is denoted as $\mathcal{X} = (X, d, \mu)$ where X is a complete separable set endowed with a distance d and a positive Borel measure $\mu \in \mathcal{M}_+(X)$.

Csiszár Divergences (φ-Divergences)

$$\mathbf{D}_\varphi(\mu \mid \nu) \stackrel{\text{def.}}{=} \int_X \varphi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\nu + \varphi'_\infty \int_X \mathrm{d}\mu^\perp \tag{1}$$

- $\varphi : \mathbb{R}_+ \to [0, +\infty]$, a convex, lower semi-continuous, positive function, $\varphi(1) = 0$.
- $\mu = \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\nu + \mu^\perp$, the Lebesgue decomposition of μ with respect to ν
- $\varphi'_\infty = \lim_{r\to\infty} \varphi(r)/r \in \mathbb{R} \cup \{+\infty\}$, recession constant.

D_φ is convex, positive, 1-homogeneous and weak lower-semicontinuous.
Particular instances of φ -divergences:

- Kullback-Leibler (KL) for $\varphi(r) = r\log(r) - r + 1$ ($\varphi'_\infty = \infty$);
- Total Variation (TV) for $\varphi(r) = |r - 1|$.

Balanced and Unbalanced Optimal Transport

Balanced optimal transport: μ, ν are probability distributions, i.e. $\mu(X) = 1$, $\nu(Y) = 1$.
Unbalanced optimal transport: arbitrary positive measures $(\mu, \nu) \in \mathcal{M}_+(X)^2$.

Unbalanced Wasserstein Distances

$$\mathrm{UW}(\mu,\nu)^q \overset{\text{def.}}{=} \inf_{\pi\in\mathcal{M}(X\times X)} \int \lambda(d(x,y))\mathrm{d}\pi(x,y) + \mathrm{D}_\varphi(\pi_1 \mid \mu) + \mathrm{D}_\varphi(\pi_2 \mid \nu) \tag{2}$$

- (π_1, π_2) are the two marginals of the joint distribution π.
- Often take $\rho\mathrm{D}_\varphi$ instead of D_φ (1.e. take $\psi = \rho\varphi$) to adjust the strength of the marginals penalization:
  - Balanced OT: the convex indicator $\varphi = \iota_{\{1\}}$ or the limit $\rho \to +\infty$(enforces $\pi_1 = \mu$ and $\pi_2 = \nu$).
  - Unbalanced OT: $0 < \rho < +\infty$, operates a trade-off between transportation and creation of mass.

Contributions

The two main contributions of this paper are the definition of two formulations relaxing the GW distance:

- Unbalanced Gromov-Wasserstein(UGW) divergence: can be computed efficiently on GPUs.
- Conic Gromov-Wasserstein distance(CGW): a distance between mm-spaces endowed with positive measures up to isometries.

Quadratic φ -Divergences

$$\mathrm{D}_\varphi^\otimes(\rho \mid \nu) \overset{\text{def.}}{=} \mathrm{D}_\varphi(\rho \otimes \rho \mid \nu \otimes \nu) \tag{3}$$

- $\rho \otimes \rho \in \mathcal{M}_+\left(X^2\right)$ is the tensor product measure
- $\mathrm{d}(\rho \otimes \rho)(x, y) = \mathrm{d}\rho(x)\mathrm{d}\rho(y)$
- $D_\varphi^\otimes$ is not a convex function in general.

Unbalanced GW Divergence

$$\mathrm{UGW}(\mathcal{X},\mathcal{Y}) = \inf_{\pi\in\mathcal{M}^+(X\times Y)} \mathcal{L}(\pi)$$

where

$$\mathcal{L}(\pi) \overset{\text{def.}}{=} \int_{X^2\times Y^2} \lambda\left(\left|d_X(x,x') - d_Y(y,y')\right|\right) \mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y') + \mathrm{D}_\varphi^\otimes(\pi_1 \mid \mu) + \mathrm{D}_\varphi^\otimes(\pi_2 \mid \nu)$$

Why quadratic divergences?
Using quadratic divergences results in UGW being 2 -homogeneous: for $\theta \geq 0$, writing $(\mathcal{X}_\theta,\mathcal{Y}_\theta)$ equiped with $(\theta\mu,\theta\nu)$, one has $\theta^{-2}\,\mathrm{UGW}(\mathcal{X}_\theta,\mathcal{Y}_\theta) = \mathrm{UGW}(\mathcal{X},\mathcal{Y})$.
Using tensorized divergences ensure that the behavior does not depends on θ.

Existence and Definiteness of UGW

Proposition 1 (Existence of minimizers)
We assume that (X, Y) are compact and that either $(i)\varphi$ superlinear, i.e $\varphi'_\infty = \infty$, or (ii) λ has compact sublevel sets in $\mathbb{R}_+$ and $2\varphi'_\infty + \inf \lambda > 0$. Then there exists $\pi \in \mathcal{M}_+(X \times Y)$ such that $\mathrm{UGW}(\mathcal{X}, \mathcal{Y}) = \mathcal{L}(\pi)$

Proposition 2 (Definiteness of UGW)
Assume that $\varphi^{-1}(\{0\}) = \{1\}$ and $\lambda^{-1}(\{0\}) = \{0\}$. Then UGW $(\mathcal{X}, \mathcal{Y}) \geq 0$ and is 0 if and only if $\mathcal{X} \sim \mathcal{Y}$.

Reformulation of UGW

Lemma 1
Defining $L_c(a,b) \overset{\text{def.}}{=} c + a\varphi(1/a) + b\varphi(1/b)$, and writing $\left(f \overset{\text{def.}}{=} \frac{d\mu}{d\pi_1}, g \overset{\text{def.}}{=} \frac{d\nu}{d\pi_2}\right)$ the Lebesgue densities of (μ, ν) w.r.t. (π_1, π_2) such that $\mu = f\pi_1 + \mu^\perp$ and $\nu = g\pi_2 + \nu^\perp$, one has

$$\mathcal{L}(\pi) = \int_{X^2 \times Y^2} L_{\lambda(|d_X - d_Y|)}(f \otimes f, g \otimes g) \mathrm{d}\pi \mathrm{d}\pi + \varphi(0)\left(\left|(\mu \otimes \mu)^\perp\right| + \left|(\nu \otimes \nu)^\perp\right|\right)$$

Background on Cone Sets and Distances

The conic formulation lifts a point $x \in X$ to a couple $(x, r) \in X \times \mathbb{R}^+$ where r encodes some (power of a) mass. Then we seek optimal transport plans defined over $\mathfrak{C}[X] \stackrel{\text{def.}}{=} X \times \mathbb{R}_+/(X \times \{0\})$.
In the sequel, points of $X \times \mathbb{R}_+$ are noted (x, r), while $[x, r]$ are quotiented points of $\mathfrak{C}[X]$.

Background on Cone Sets and Distances

Consider coordinates of the form $([u,a],[v,b]) = ([d_X(x,x'), rr'], [d_Y(y,y'), ss']) \in \mathfrak{C}[\mathbb{R}_+] \times \mathfrak{C}[\mathbb{R}_+]$. Thus conic discrepancies $\mathcal{D}$ on $\mathfrak{C}[\mathbb{R}_+]$ are defined for $(p,q) \geq 0$ as

$$\mathcal{D}([u,a],[v,b])^q \overset{\text{def.}}{=} H_{\lambda(|u-v|)}(a^p, b^p) \quad \text{where} \quad H_c(a^p, b^p) \overset{\text{def.}}{=} \inf_{\theta \geq 0} \theta L_c\left(\frac{a^p}{\theta}, \frac{b^p}{\theta}\right) \tag{4}$$

Proposition 3.
Assume $\lambda^{-1}(\{0\}) = \{0\}, \varphi^{-1}(\{0\}) = \{1\}$ and φ is coercive. Then $\mathcal{D}$ is definite on $\mathfrak{C}[\mathbb{R}^+]$, i.e. $\mathcal{D}([u,a],[v,b]) = 0$ if and only if $(a = b = 0)$ or $(a = b$ and $u = v)$.

Conic GW Distance

$$\mathrm{CGW}(\mathcal{X},\mathcal{Y}) \overset{\text{def.}}{=} \inf_{\alpha\in\mathcal{U}_p(\mu,\nu)} \mathcal{H}(\alpha)$$

where

$$\mathcal{H}(\alpha) \overset{\text{def.}}{=} \int \mathcal{D}\left(\left[d_X(x,x'), rr'\right], \left[d_Y(y,y'), ss'\right]\right)^q \, \mathrm{d}\alpha([x,r],[y,s])\mathrm{d}\alpha\left([x',r'],[y',s']\right)$$

$$\mathcal{U}_p(\mu,\nu) \overset{\text{def.}}{=} \left\{\alpha \in \mathcal{M}_+(\mathfrak{C}[X]\times\mathfrak{C}[Y]) : \int_{\mathbb{R}_+} r^p \, \mathrm{d}\alpha_1(\cdot,r) = \mu, \int_{\mathbb{R}_+} s^p \, \mathrm{d}\alpha_2(\cdot,s) = \nu\right\}$$

Properties of Conic GW Distance

Theorem 1
(i) The divergence CGW is symmetric, positive and definite up to isometries.
(ii) If $\mathcal{D}$ is a distance on $\mathfrak{C}[\mathbb{R}_+]$, then $\mathrm{CGW}^{1/q}$ is a distance on the set of mm -spaces up to isometries.
(iii) For any $\left(\mathrm{D}_\varphi, \lambda, p, q\right)$ with associated cost $\mathcal{D}$ on the cone, one has $\mathrm{UGW} \geq \mathrm{CGW}$.

Proposition 4
For a fixed γ, the optimal $\pi \in \arg\min \mathcal{F}(\pi, \gamma) + \varepsilon \mathrm{KL}\left(\pi \otimes \gamma \mid (\mu \otimes \nu)^{\otimes 2}\right)$ is the solution of
$\min_\pi \int c_\gamma^\varepsilon(x, y) \mathrm{d}\pi(x, y) + \rho m(\gamma) \mathrm{KL}\left(\pi_1 \mid \mu\right) + \rho m(\gamma) \mathrm{KL}\left(\pi_2 \mid \nu\right) + \varepsilon m(\gamma) \mathrm{KL}(\pi \mid \mu \otimes \nu)$,
where $m(\gamma) \stackrel{\text{def.}}{=} \gamma(X \times Y)$ is the mass of γ, and where we define the cost associated to γ as $c_\gamma^\varepsilon(x, y) \stackrel{\text{def.}}{=}$
$\int \lambda\left(\left|d_X(x, \cdot) - d_Y(y, \cdot)\right|\right) \mathrm{d}\gamma + \rho \int \log\left(\frac{\mathrm{d}\gamma_1}{\mathrm{d}\mu}\right) \mathrm{d}\gamma_1 + \rho \int \log\left(\frac{\mathrm{d}\gamma_2}{\mathrm{d}\nu}\right) \mathrm{d}\gamma_2 + \varepsilon \int \log\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\mu \mathrm{d}\nu}\right) \mathrm{d}\gamma$

Numerical Computation of UGW

We introduce a lower bound obtained by introducing two transportation plans. To further accelerate the method and enable GPU-friendly iterations, we consider an entropic regularization. It reads, for any $\epsilon \geq 0$,

$$\begin{aligned} r\mathrm{UGW}_{\varepsilon}(\mathcal{X},\mathcal{Y}) &\stackrel{\text{def.}}{=} \inf_{\pi} \mathcal{L}(\pi) + \varepsilon \mathrm{KL}^{\otimes}(\pi \mid \mu \otimes \nu) \\ &\geq \inf_{\pi,\gamma} \mathcal{F}(\pi,\gamma) + \varepsilon \mathrm{KL}\left(\pi \otimes \gamma \mid (\mu \otimes \nu)^{\otimes 2}\right), \\ \text{and } \mathcal{F}(\pi,\gamma) &\stackrel{\text{def.}}{=} \int_{X^2 \times Y^2} \lambda\left(|d_X - d_Y|\right) \mathrm{d}\pi \otimes \gamma \\ &+ \mathrm{D}_{\varphi}\left(\pi_1 \otimes \gamma_1 \mid \mu \otimes \mu\right) + \mathrm{D}_{\varphi}\left(\pi_2 \otimes \gamma_2 \mid \nu \otimes \nu\right) \end{aligned} \tag{5}$$

where (γ_1, γ_2) denote the marginals of the plan γ.

Algorithm

Algorithm 1 – UGW($\mathcal{X}, \mathcal{Y}, \rho, \varepsilon$)

Input: mm-spaces $(\mathcal{X}, \mathcal{Y})$, relaxation ρ, regularization ε
Output: approximation (π, γ) minimizing $\boxed{6}$

1: Initialize $\pi = \gamma = \mu \otimes \nu / \sqrt{m(\mu)m(\nu)}$, $g = 0$.
2: **while** (π, γ) has not converged **do**
3: Update $\pi \leftarrow \gamma$, then $c \leftarrow c_\pi^\varepsilon$, $\tilde{\rho} \leftarrow m(\pi)\rho$, $\tilde{\varepsilon} \leftarrow m(\pi)\varepsilon$
4: **while** (f, g) has not converged **do**
5: $\forall x,\ f(x) \leftarrow -\frac{\tilde{\varepsilon}\tilde{\rho}}{\tilde{\varepsilon}+\tilde{\rho}} \log\left(\int e^{(g(y)-c(x,y))/\tilde{\varepsilon}} \mathrm{d}\nu(y) \right)$
6: $\forall y,\ g(y) \leftarrow -\frac{\tilde{\varepsilon}\tilde{\rho}}{\tilde{\varepsilon}+\tilde{\rho}} \log\left(\int e^{(f(x)-c(x,y))/\tilde{\varepsilon}} \mathrm{d}\mu(x) \right)$
7: Update $\gamma(x, y) \leftarrow \exp\left[(f(x) + g(y) - c(x, y))/\tilde{\varepsilon} \right] \mu(x)\nu(y)$
8: Rescale $\gamma \leftarrow \sqrt{m(\pi)/m(\gamma)}\gamma$
9: Return (π, γ).

- Computing the cost c_γ^ε for spaces X and Y of n points has in general a cost $O\left(n^4\right)$ in time and memory.
- Each iteration of Sinkhorn thus has a cost n^2.

Robustness to Imbalanced Classes

- $X = Y = \mathbb{R}^2$
- $\mathcal{E}, \mathcal{C}$ and $\mathcal{S}$: uniform distributions on an ellipse, a disk and a square.
- Figure 1 contrasts the transportation plan obtained by GW and UGW for a fixed $\mu = 0.5\mathcal{E} + 0.5\mathcal{C}$ and ν obtained using two different mixtures of $\mathcal{E}$ and $\mathcal{S}$.

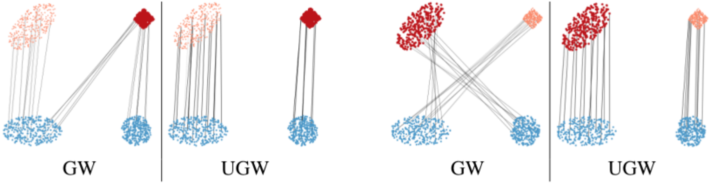

Figure 1: GW vs. UGW transportation plan, using $\nu = 0.3\mathcal{E} + 0.7\mathcal{S}$ on the left, and $\nu = 0.7\mathcal{E} + 0.3\mathcal{S}$ on the right.

Robustness to Outlier

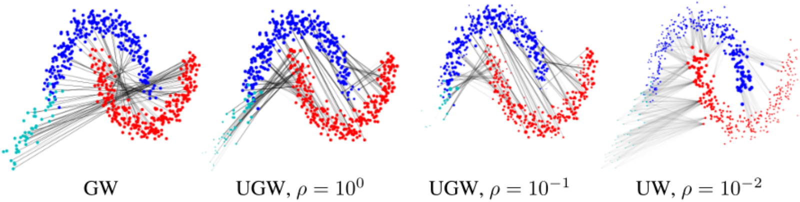

Figure 2: GW and UGW applied to two moons with outliers.

Decreasing the value of ρ (thus allowing for more mass creation/destruction in place of transportation) is able to reduce and even remove the influence of the outliers.

Graph Matching and Comparison with Partial-GW

The colors $c(x)$ are defined on the "source" graph X and are mapped by an optimal plan π on $y \in Y$ to a color $\frac{1}{\pi_1(u)} \int_X c(x) \mathrm{d}\pi(x, y)$.

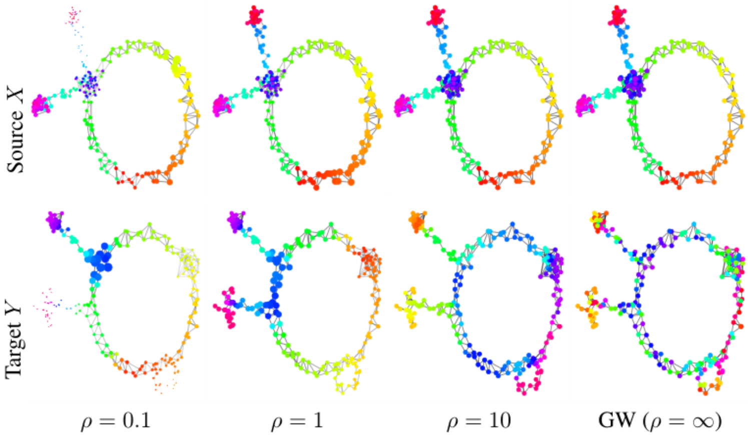

Figure 3: Comparison of UGW and GW for graph matching.

Graph Matching and Comparison with Partial-GW

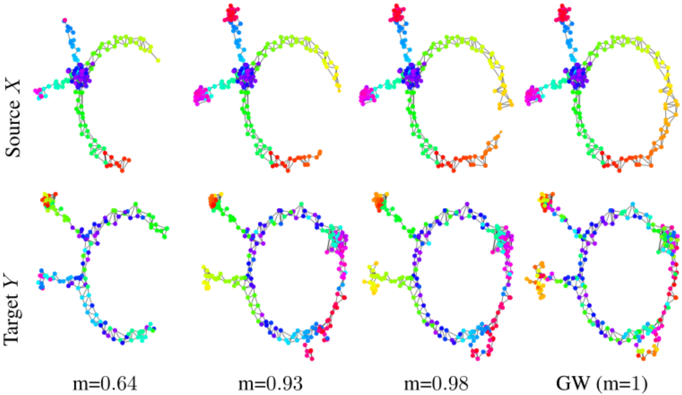

Figure 4: Comparison of Partial-GW for graph matching. Here m is the budget of transported mass.

PGW is equivalent to solving GW on sub-graphs, so the color distribution of GW and PGW are the same.