The unbalanced Gromov-Wasserstein distance: Conic Formulation and Relaxation

Thibault Séjourné

Mokameeting – 14th October, 2020

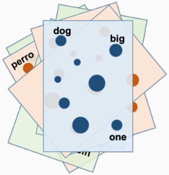
Joint work with Francois-Xavier Vialard and Gabriel Peyré

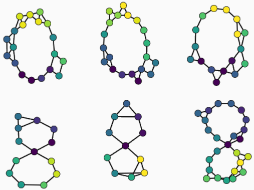
Introduction

Motivations

Some of machine learning challenges are

- matching point clouds of $\mathbb{R}^d$ up to isometries[1]
- graph matching[2]





[1] Alaux, J., Grave, E., Cuturi, M., & Joulin, A. (2018). Unsupervised hyperalignment for multilingual word embeddings.

[2] Vayer, T., Chapel, L., Flamary, R., Tavenard, R., & Courty, N. (2018). Fused Gromov-Wasserstein distance for structured objects.

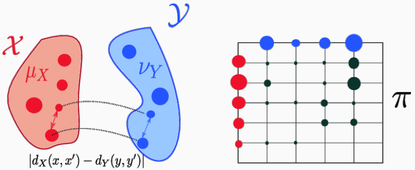
Metric measure spaces and Gromov-Wasserstein distance

mm-space: $\mathcal{X} = (X, d_X, \mu)$ with (X, d_X) complete separable, μ positive measure

Definition - GW distance[3]

Take $\mathcal{X} = (X, d_X, \mu)$ and $\mathcal{Y} = (Y, d_Y, \nu)$ equipped with **probabilities**. Define a penalty $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ (e.g. $\lambda(t) = t^q$). One defines $GW(\mathcal{X}, \mathcal{Y}) = \inf_{\{\pi_1=\mu,\ \pi_2=\nu\}} \mathcal{G}(\pi)$ where

$$\mathcal{G}(\pi) \stackrel{\text{def.}}{=} \int \lambda(|d_X(x,x') - d_Y(y,y')|)\mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y')$$

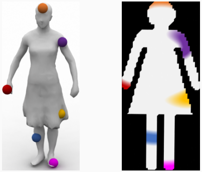


[3] Mémoli, F. (2011). Gromov–Wasserstein distances and the metric approach to object matching.

Specificities of GW

Two key differences with OT

- GW is non-convex (quadratic assignment program)
- $(\mathcal{X}, \mathcal{Y})$ can differ radically in nature.[4]



Isometric mm-spaces

Def: $\mathcal{X} \sim \mathcal{Y} \Leftrightarrow \exists \psi : X \to Y$ bijective isometry s.t.

$$d_X(x, x') = d_Y(\psi(x), \psi(x')) \quad \text{and} \quad \nu = \psi_\sharp \mu$$

Prop: With $\lambda(t) = t^q$, $GW^{\frac{1}{q}}$ distance and definite iff $\mathcal{X} \sim \mathcal{Y}$

[4] Solomon, J., Peyré, G., Kim, V. G., & Sra, S. (2016). Entropic metric alignment for correspondence problems.

Motivations of GW

- GW makes sense in a variety of problems (NLP, generative learning[5], domain adaptation[6], etc...)
- Though GW suffers limitations similar to OT
    - restricted to probabilities
    - sensitive to outliers
- Some attempts to extend to positive measures
    - Generalize Sturm's[7] Lq transportation distance[8] (no fast algo)
    - Partial-GW[9] (no properties)

---

[5] Bunne, C., Alvarez-Melis, D., Krause, A., & Jegelka, S. (2019). Learning generative models across incomparable spaces.

[6] Redko, I., Vayer, T., Flamary, R., & Courty, N. (2020). CO-Optimal Transport.

[7] Sturm, K. T. (2006). On the geometry of metric measure spaces.

[8] De Ponti, N., & Mondino, A. (2020). Entropy-Transport distances between unbalanced metric measure spaces.

[9] Chapel, L., Alaya, M. Z., & Gasso, G. (2020). Partial Gromov-Wasserstein with Applications on Positive-Unlabeled Learning.

Unbalanced OT - Static and conic formulations

Csiszàr divergences[10]

Define $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ l.s.c., convex, $\varphi(1) = 0$, $\varphi'^{\infty} = \lim_{x\to\infty} \frac{\varphi(x)}{x}$.
Decompose $\forall(\alpha, \beta)$, $\alpha = \frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\beta + \alpha^{\top}$ (discrete: $\alpha = \sum_i \alpha_i \delta_{x_i}$)

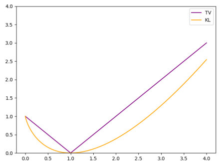
Definition - φ-divergence

$$\mathrm{D}_{\varphi}(\alpha, \beta) \stackrel{\text{def.}}{=} \int_{\mathcal{X}} \varphi\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}(x)\right)\mathrm{d}\beta(x) + \varphi'^{\infty} \int_{\mathcal{X}} \mathrm{d}\alpha^{\top}(x),$$

$$\mathrm{D}_{\varphi}(\alpha, \beta) = \sum_{\beta_i \neq 0} \varphi\left(\frac{\alpha_i}{\beta_i}\right)\beta_i + \varphi'^{\infty} \sum_{\beta_i = 0} \alpha_i.$$

Examples:

- $\mathrm{KL}(\alpha, \beta) = \sum_i (\log(\frac{\alpha_i}{\beta_i})\alpha_i - \alpha_i + \beta_i)$:
  $\varphi(x) = x \log x - x + 1$,
- $\mathrm{TV}(\alpha, \beta) = \sum_i |\alpha_i - \beta_i|$: $\varphi(x) = |x - 1|$.



[10] Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observation.

Unbalanced optimal transport - relaxed formulation

Idea: Soften the hard constraint $\pi_1 = \mu \to \mathrm{D}_\varphi(\pi_1|\mu)$.

Definition - Unbalanced OT[11]

For any φ-divergence D_φ and any measures (μ, ν) one defines $\mathrm{UOT}(\mu, \nu) = \inf_{\pi \geq 0} \mathcal{L}_1(\pi)$ where

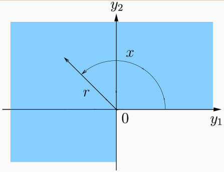
$$\mathcal{L}_1(\pi) \stackrel{\text{def.}}{=} \int_{\mathcal{X}} \lambda(d) \mathrm{d}\pi + \mathrm{D}_\varphi(\pi_1, \mu) + \mathrm{D}_\varphi(\pi_2, \nu).$$

- **Two dynamics:** transportation vs creation/destruction.
- **Balanced OT** is retrieved with $\mathrm{D}_\varphi = \iota_{(=)}$.
- **Choice of D_φ:** prior on the mass variation dynamics.

[11] Liero, M., Mielke, A., & Savaré, G. (2018). Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures.

Unbalanced optimal transport - conic formulation

Idea: Add a mass variable and lift on a cone $\mathfrak{C} = (X \times \mathbb{R}_+)/(X \times \{0\})$ endowed with a distance $\mathcal{D}_{\mathfrak{C}}([x,r],[y,s])$ (e.g. $X \subset \mathbb{S}^1$)



Definition

Define $\mathrm{COT}(\mu,\nu) \overset{\text{def.}}{=} \inf_{\gamma\in\mathcal{U}_p(\mu,\nu)} \mathcal{H}_1(\gamma)$ where

$$\mathcal{H}_1(\gamma) \overset{\text{def.}}{=} \int \mathcal{D}_{\mathfrak{C}}([x,r],[y,s])^q \mathrm{d}\gamma([x,r],[y,s]),$$

$$\mathcal{U}_p(\mu,\nu) \overset{\text{def.}}{=} \{\gamma \geq 0,\ \int r^p \mathrm{d}\gamma_1(\cdot, r) = \mu, \int s^p \mathrm{d}\gamma_2(\cdot, s) = \nu\}$$

Key point: $\mathrm{COT} = \inf_{\gamma\in\mathcal{U}_p(\mu,\nu)} \int \mathcal{D}_{\mathfrak{C}} \mathrm{d}\gamma$ (Wasserstein over $\mathfrak{C}$)

$\Rightarrow$ **If $\mathcal{D}_{\mathfrak{C}}$ is a distance, then COT is a distance**

Equivalence between both formulations

Proposition

If one defines $\mathcal{D}_{\mathfrak{C}}$ as

$$\mathcal{D}_{\mathfrak{C}}([x,r],[y,s])^q = H_{\lambda(d(x,y))}(r^p, s^p),$$

$$H_c(r,s) \overset{\text{def.}}{=} \inf_{\theta \geq 0} \theta L_c(r/\theta, s/\theta),$$

$$L_c(r,s) \overset{\text{def.}}{=} c + r\varphi(1/r) + s\varphi(1/s),$$

Then $\mathbf{UOT} = \mathbf{COT}$.

- In general $\mathcal{D}_{\mathfrak{C}}$ is definite but not necessarily a distance.
- Note that H_c depends on φ, $\mathcal{D}_{\mathfrak{C}}$ on (φ, λ, p, q)

When is $\mathcal{D}_{\mathfrak{C}}$ a distance over $\mathfrak{C}$?

Explicit conic settings inducing distances

In those settings $\mathcal{D}_{\mathfrak{C}}$ is a distance, thus so is the conic formulation.

Gaussian Hellinger

$$\mathrm{D}_\varphi = \mathrm{KL}, \quad \lambda(t) = t^2 \quad \text{and} \quad q = p = 2,$$
$$\mathcal{D}_{\mathfrak{C}}([x,r],[y,s])^2 = r^2 + s^2 - 2rse^{-d(x,y)/2}.$$

Hellinger-Kantorovich / Wasserstein-Fisher-Rao

$$\mathrm{D}_\varphi = \mathrm{KL}, \quad \lambda(t) = -\log\cos^2(t \wedge \tfrac{\pi}{2}) \quad \text{and} \quad q = p = 2,$$
$$\mathcal{D}_{\mathfrak{C}}([x,r],[y,s])^2 = r^2 + s^2 - 2rs\cos(\tfrac{\pi}{2} \wedge d(x,y)).$$

Partial optimal transport

$$\mathrm{D}_\varphi = \mathrm{TV}, \quad \lambda(t) = t^q \quad \text{and} \quad q \geq 1 \quad \text{and} \quad p = 1,$$
$$\mathcal{D}_{\mathfrak{C}}([x,r],[y,s])^q = r + s - (r \wedge s)(2 - d(x,y)^q)_+.$$

Unbalanced Gromov-Wasserstein

Unbalanced Gromov-Wasserstein

Definition

One defines $UGW(\mathcal{X},\mathcal{Y}) = \inf_{\pi\geq 0}\mathcal{L}_2(\pi)$ where

$$\mathcal{L}_2(\pi) = \int \lambda(|d_X - d_Y|)\mathrm{d}\pi\mathrm{d}\pi + \mathrm{D}_\varphi(\pi_1\otimes\pi_1, \mu\otimes\mu) + \mathrm{D}_\varphi(\pi_2\otimes\pi_2, \nu\otimes\nu).$$

To be compared with

$$\mathcal{G}(\pi) = \int \lambda(|d_X(x,x') - d_Y(y,y')|)\mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y'),$$

$$\mathcal{L}_1(\pi) = \int_{\mathcal{X}} \lambda(d)\mathrm{d}\pi + \mathrm{D}_\varphi(\pi_1,\mu) + \mathrm{D}_\varphi(\pi_2,\nu).$$

Unbalanced Gromov-Wasserstein

Definition

One defines $UGW(\mathcal{X}, \mathcal{Y}) = \inf_{\pi \geq 0} \mathcal{L}_2(\pi)$ where

$$\mathcal{L}_2(\pi) = \int \lambda(|d_X - d_Y|)\mathrm{d}\pi\mathrm{d}\pi + \mathrm{D}_\varphi(\pi_1 \otimes \pi_1, \mu \otimes \mu) + \mathrm{D}_\varphi(\pi_2 \otimes \pi_2, \nu \otimes \nu).$$

To be compared with

$$\mathcal{G}(\pi) = \int \lambda(|d_X(x, x') - d_Y(y, y')|)\mathrm{d}\pi(x, y)\mathrm{d}\pi(x', y'),$$

$$\mathcal{L}_1(\pi) = \int_{\mathcal{X}} \lambda(d)\mathrm{d}\pi + \mathrm{D}_\varphi(\pi_1, \mu) + \mathrm{D}_\varphi(\pi_2, \nu).$$

Unbalanced Gromov-Wasserstein

Definition

One defines $UGW(\mathcal{X}, \mathcal{Y}) = \inf_{\pi \geq 0} \mathcal{L}_2(\pi)$ where

$$\mathcal{L}_2(\pi) = \int \lambda(|d_X - d_Y|) \mathrm{d}\pi \mathrm{d}\pi + \mathrm{D}_\varphi(\pi_1 \otimes \pi_1, \mu \otimes \mu) + \mathrm{D}_\varphi(\pi_2 \otimes \pi_2, \nu \otimes \nu).$$

To be compared with

$$\mathcal{G}(\pi) = \int \lambda(|d_X(x, x') - d_Y(y, y')|) \mathrm{d}\pi(x, y) \mathrm{d}\pi(x', y'),$$

$$\mathcal{L}_1(\pi) = \int_{\mathcal{X}} \lambda(d) \mathrm{d}\pi + \mathrm{D}_\varphi(\pi_1, \mu) + \mathrm{D}_\varphi(\pi_2, \nu).$$

Unbalanced Gromov-Wasserstein

Definition

One defines $UGW(\mathcal{X},\mathcal{Y}) = \inf_{\pi\geq 0} \mathcal{L}_2(\pi)$ where

$$\mathcal{L}_2(\pi) = \int \lambda(|d_X - d_Y|)\mathrm{d}\pi\mathrm{d}\pi + \mathrm{D}_\varphi(\pi_1 \otimes \pi_1, \mu \otimes \mu) + \mathrm{D}_\varphi(\pi_2 \otimes \pi_2, \nu \otimes \nu).$$

To be compared with

$$\mathcal{G}(\pi) = \int \lambda(|d_X(x,x') - d_Y(y,y')|)\mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y'),$$

$$\mathcal{L}_1(\pi) = \int_{\mathcal{X}} \lambda(d)\mathrm{d}\pi + \mathrm{D}_\varphi(\pi_1,\mu) + \mathrm{D}_\varphi(\pi_2,\nu).$$

Conic Gromov-Wasserstein

Definition

One defines $CGW(\mathcal{X}, \mathcal{Y}) = \inf_{\gamma \in \mathcal{U}_p(\mu,\nu)} \mathcal{H}_2(\gamma)$ where

$$\mathcal{H}_2(\gamma) \overset{\text{def.}}{=} \int \mathcal{D}_{\mathfrak{C}}([d_X(x,x'), rr'], [d_Y(y,y'), ss'])^q \mathrm{d}\gamma([x,r],[y,s]) \, \mathrm{d}\gamma([x',r'],[y',s']),$$

$$\mathcal{U}_p(\mu,\nu) \overset{\text{def.}}{=} \{\gamma \geq 0, \int r^p \mathrm{d}\gamma_1(\cdot, r) = \mu, \int s^p \mathrm{d}\gamma_2(\cdot, s) = \nu\}.$$

To be compared with

$$\mathcal{G}(\pi) = \int \lambda(|d_X(x,x') - d_Y(y,y')|) \mathrm{d}\pi(x,y) \mathrm{d}\pi(x',y'),$$

$$\mathcal{H}_1(\gamma) = \int \mathcal{D}_{\mathfrak{C}}([x,r],[y,s])^q \mathrm{d}\gamma([x,r],[y,s]).$$

Conic Gromov-Wasserstein

Definition

One defines $CGW(\mathcal{X}, \mathcal{Y}) = \inf_{\gamma \in \mathcal{U}_p(\mu,\nu)} \mathcal{H}_2(\gamma)$ where

$$\mathcal{H}_2(\gamma) \stackrel{\text{def.}}{=} \int \mathcal{D}_{\mathfrak{c}}([d_X(x,x'), rr'], [d_Y(y,y'), ss'])^q \mathrm{d}\gamma([x,r],[y,s]) \mathrm{d}\gamma([x',r'],[y',s']),$$

$$\mathcal{U}_p(\mu,\nu) \stackrel{\text{def.}}{=} \{\gamma \geq 0, \int r^p \mathrm{d}\gamma_1(\cdot, r) = \mu, \int s^p \mathrm{d}\gamma_2(\cdot, s) = \nu\}.$$

To be compared with

$$\mathcal{G}(\pi) = \int \lambda(|d_X(x,x') - d_Y(y,y')|) \mathrm{d}\pi(x,y) \mathrm{d}\pi(x',y'),$$

$$\mathcal{H}_1(\gamma) = \int \mathcal{D}_{\mathfrak{c}}([x,r],[y,s])^q \mathrm{d}\gamma([x,r],[y,s]).$$

Conic Gromov-Wasserstein

Definition

One defines $CGW(\mathcal{X}, \mathcal{Y}) = \inf_{\gamma \in \mathcal{U}_p(\mu,\nu)} \mathcal{H}_2(\gamma)$ where

$$\mathcal{H}_2(\gamma) \stackrel{\text{def.}}{=} \int \mathcal{D}_{\mathfrak{C}}([d_X(x, x'), rr'], [d_Y(y, y'), ss'])^q \mathrm{d}\gamma([x, r], [y, s]) \\ \mathrm{d}\gamma([x', r'], [y', s']),$$

$$\mathcal{U}_p(\mu, \nu) \stackrel{\text{def.}}{=} \{\gamma \geq 0, \int r^p \mathrm{d}\gamma_1(\cdot, r) = \mu, \int s^p \mathrm{d}\gamma_2(\cdot, s) = \nu\}.$$

To be compared with

$$\mathcal{G}(\pi) = \int \lambda(|d_X(x, x') - d_Y(y, y')|) \mathrm{d}\pi(x, y) \mathrm{d}\pi(x', y'),$$

$$\mathcal{H}_1(\gamma) = \int \mathcal{D}_{\mathfrak{C}}([x, r], [y, s])^q \mathrm{d}\gamma([x, r], [y, s]).$$

Definition

One defines $CGW(\mathcal{X},\mathcal{Y}) = \inf_{\gamma\in\mathcal{U}_p(\mu,\nu)} \mathcal{H}_2(\gamma)$ where

$$\mathcal{H}_2(\gamma) \overset{\text{def.}}{=} \int \mathcal{D}_{\mathfrak{C}}([d_X(x,x'), rr'], [d_Y(y,y'), ss'])^q \mathrm{d}\gamma([x,r],[y,s]) \, \mathrm{d}\gamma([x',r'],[y',s']),$$

$$\mathcal{U}_p(\mu,\nu) \overset{\text{def.}}{=} \{\gamma \geq 0, \int r^p \mathrm{d}\gamma_1(\cdot, r) = \mu, \int s^p \mathrm{d}\gamma_2(\cdot, s) = \nu\}.$$

To be compared with

$$\mathcal{G}(\pi) = \int \lambda(|d_X(x,x') - d_Y(y,y')|) \mathrm{d}\pi(x,y) \mathrm{d}\pi(x',y'),$$

$$\mathcal{H}_1(\gamma) = \int \mathcal{D}_{\mathfrak{C}}([x,r],[y,s])^q \mathrm{d}\gamma([x,r],[y,s]).$$

Results

Theorem

1. If $\lambda^{-1}(\{0\}) = \{0\}$ and $\varphi^{-1}(\{0\}) = \{1\}$, then UGW and CGW are definite up to isometries.
2. If $\mathcal{D}_{\mathfrak{C}}$ is a cone distance then $\text{CGW}^{1/q}$ is a distance.
3. For any (φ, λ, p, q) and associated $\mathcal{D}_{\mathfrak{C}}$, UGW $\geq$ CGW.

Same results as UOT, except UGW $\geq$ CGW while UOT $=$ COT (non-convexity).

Implementation and numerics

Implementing UGW

Idea: Entropic regularization + alternate minimization

$$\begin{aligned}\mathrm{UGW}_\varepsilon(\mathcal{X},\mathcal{Y}) &\overset{\text{def.}}{=} \inf_{\pi\geq 0} \mathcal{L}_2(\pi) + \varepsilon\mathrm{KL}(\pi\otimes\pi,(\mu\otimes\nu)^{\otimes 2})\\ &\geq \inf_{\pi,\gamma\geq 0} \mathcal{F}(\pi,\gamma) + \varepsilon\mathrm{KL}(\pi\otimes\gamma,(\mu\otimes\nu)^{\otimes 2}),\end{aligned}$$

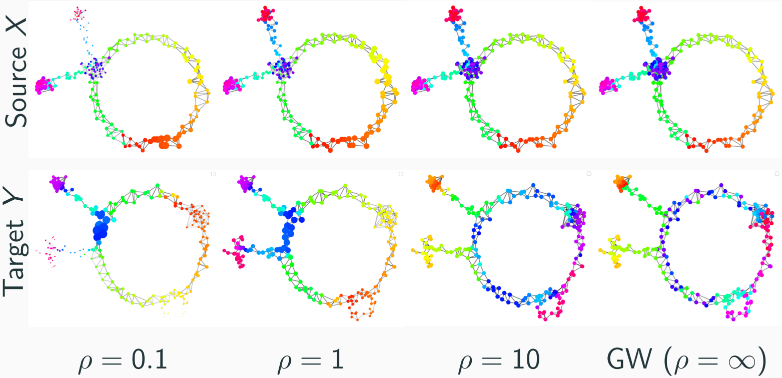
$$\begin{aligned}\text{where } \mathcal{F}(\pi,\gamma) \overset{\text{def.}}{=} &\int \lambda(|d_X - d_Y|)\mathrm{d}\pi\mathrm{d}\gamma + \mathrm{D}_\varphi(\pi_1\otimes\gamma_1,\mu\otimes\mu)\\ &+ \mathrm{D}_\varphi(\pi_2\otimes\gamma_2,\nu\otimes\nu)\end{aligned}$$

- Alternate descent = sequence of UOT_ε problems (Sinkhorn).
- Numerically we observe at optimality $\pi^* = \gamma^*$.

$\Rightarrow$ Computes a local minimizer of UGW_ε.

Numerical experiment

Setup: $\lambda(t) = t^2$, $\mathrm{D}_\varphi = \rho\mathrm{KL}$, $(\mathcal{X}, \mathcal{Y})$ = graphs with geodesic distance, (μ, ν) uniform probabilities



Take home message: UGW encodes partial and/or geometrically consistent matchings.

Conclusion

- Blending of Unbalanced OT with Gromov-Wasserstein distances
- The relaxed formulation UGW is appealing for computations
- The conic formulation CGW is a metric between mm-spaces up to isometry

Thank you !

Implementation

Implementing UGW

Idea: Entropic regularization + alternate minimization

$$\begin{aligned}\mathrm{UGW}_\varepsilon(\mathcal{X},\mathcal{Y}) &\stackrel{\text{def.}}{=} \inf_{\pi\geq 0} \mathcal{L}_2(\pi) + \varepsilon \mathrm{KL}(\pi\otimes\pi, (\mu\otimes\nu)^{\otimes 2}) \\ &\geq \inf_{\pi,\gamma\geq 0} \mathcal{F}(\pi,\gamma) + \varepsilon \mathrm{KL}(\pi\otimes\gamma, (\mu\otimes\nu)^{\otimes 2}),\end{aligned}$$

$$\begin{aligned}\text{where } \mathcal{F}(\pi,\gamma) \stackrel{\text{def.}}{=} \int \lambda(|d_X - d_Y|) \mathrm{d}\pi \mathrm{d}\gamma + \mathrm{D}_\varphi(\pi_1\otimes\gamma_1, \mu\otimes\mu) \\ + \mathrm{D}_\varphi(\pi_2\otimes\gamma_2, \nu\otimes\nu)\end{aligned}$$

- $\forall s > 0$, (π,γ) optimal $\Rightarrow$ $(s\pi, \frac{1}{s}\gamma)$ optimal

$\Rightarrow$ impose $m(\pi) = m(\gamma)$

- Numerically we observe at optimality $\pi^* = \gamma^*$.

Alternate UGW = sequence of Sinkhorn updates

- Focus on $\lambda(t) = t^2$ for improved time and memory complexity
- Focus on $\mathrm{D}_\varphi = \mathrm{KL}$ which verifies

$$\mathrm{KL}(\mu \otimes \nu, \alpha \otimes \beta) = m(\nu)\mathrm{KL}(\mu, \alpha) + m(\mu)\mathrm{KL}(\nu, \beta) \\ +(m(\mu) - m(\alpha))(m(\nu) - m(\beta)).$$

$\Rightarrow$ Given γ, minimizing w.r.t. π amounts to solve a regularized UOT problem.

Reformulation of the alternate minimization

We focus on $\mathrm{D}_\varphi = \rho\mathrm{KL}$.

Proposition - alternate descent $\leftrightarrow$ solve UOT

For a fixed γ, $\pi \in \arg\min_\pi \mathcal{F}(\pi,\gamma) + \varepsilon\mathrm{KL}(\pi\otimes\gamma|(\mu\otimes\nu)^{\otimes 2})$ is the solution of

$$\min_\pi \int c_\gamma^\varepsilon(x,y)\mathrm{d}\pi(x,y) + \rho m(\gamma)\mathrm{KL}(\pi_1|\mu) + \rho m(\gamma)\mathrm{KL}(\pi_2|\nu) + \varepsilon m(\gamma)\mathrm{KL}(\pi|\mu\otimes\nu), \qquad \text{where}$$

$$c_\gamma^\varepsilon(x,y) \stackrel{\text{def.}}{=} \int \lambda(|d_X(x,\cdot) - d_Y(y,\cdot)|)\mathrm{d}\gamma + \rho\int \log(\frac{\mathrm{d}\gamma_1}{\mathrm{d}\mu})\mathrm{d}\gamma_1 + \rho\int\log(\frac{\mathrm{d}\gamma_2}{\mathrm{d}\nu})\mathrm{d}\gamma_2 + \varepsilon\int\log(\frac{\mathrm{d}\gamma}{\mathrm{d}\mu\mathrm{d}\nu})\mathrm{d}\gamma.$$

Algorithm

Algorithm 1 – UGW($\mathcal{X}$, $\mathcal{Y}$, ρ, ε)

Input: mm-spaces $(\mathcal{X}, \mathcal{Y})$, relaxation ρ, regularization ε
Output: approximation (π, γ) minimizing $\mathcal{F} + \varepsilon \mathrm{KL}^{\otimes}$

Initialize (π, γ) and (f, g)
while (π, γ) has not converged **do**
  Update $\pi \leftarrow \gamma$ and compute the cost $c \leftarrow c_\pi^\varepsilon$
  Update parameters $(\tilde{\rho}, \tilde{\varepsilon}) \leftarrow (m(\pi)\rho, m(\pi)\varepsilon)$
  Compute (f, g) that solves UOT(μ, ν, c_π^ε, $\tilde{\rho}$, $\tilde{\varepsilon}$)
  Update $\gamma(x, y) \leftarrow \exp\left[(f(x) + g(y) - c(x, y))/\tilde{\varepsilon}\right]\mu(x)\nu(y)$
  Rescale $\gamma \leftarrow \sqrt{m(\pi)/m(\gamma)}\gamma$
Return (π, γ).

Detailed algorithm

Algorithm 2 – UGW($\mathcal{X}$, $\mathcal{Y}$, ρ, ε)

Input: mm-spaces $(\mathcal{X}, \mathcal{Y})$, relaxation ρ, regularization ε
Output: approximation (π, γ) minimizing $\mathcal{F} + \varepsilon \mathrm{KL}^{\otimes}$

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