Some Extensions of Optimal Transport

Binh Nguyen – Telecom Paris

M2DS Research Seminar Course

(With credit of some illustrations from G.Peyré and R.Flamary)

Outline

Reminder on Optimal Transport

Some Extensions of Optimal Transport

Optimal Transport across different spaces

Outline

Previously...

Monge optimal transport (1781)

$$\mathrm{MOT}(\alpha, \beta) \overset{\text{def.}}{=} \min_{T : T_{\#}\alpha = \beta} \int d(x, T(x))\alpha(dx)$$

But:

- Not guarantee there exists a solution T
- Not guarantee uniqueness of the solution T
- Not symmetric: $\mathrm{MOT}(\alpha, \beta) \neq \mathrm{MOT}(\beta, \alpha)$

Previously...

Kantorovic optimal transport (1942)

$$\mathrm{OT}(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\pi : \pi_1 = \alpha, \pi_2 = \beta} \int \int C(x, y) d\pi(x, y)$$

But:

- Guarantee there exists a solution π (with some assumptions on C)
- Solution still not unique
- Symmetric
- Not differentiable

Previously...

Kantorovic optimal transport – discrete formulation
(discrete measure $\rightarrow$ discrete measure)

$$\alpha = \sum_{i=1}^{n} \alpha_i \delta_{x_i} \quad \beta = \sum_{j=1}^{m} \beta_j \delta_{y_j}$$

$$\mathrm{OT}(\alpha, \beta) \overset{\text{def.}}{=} \min_{\mathrm{P}:\mathrm{P}\mathbb{1}=\alpha, \mathrm{P}^\top\mathbb{1}=\beta} \sum_{i,j} C_{ij} P_{ij} \quad \text{with} \quad C_{ij} = d(x_i, y_j)$$

- Easiest to understand
- C and P now are just two matrices in $\mathbb{R}^{n\times m}$
- Solved with linear programming techniques, *e.g.* simplex algo.
- But: $\mathcal{O}(n^3 \log(n)) \rightarrow$ costly to solve when n large

Previously...

Entropic (regularized) optimal transport – discrete formulation
(discrete measure $\rightarrow$ discrete measure)

$$\alpha = \sum_{i=1}^{n} \alpha_i \delta_{x_i} \quad \beta = \sum_{j=1}^{m} \beta_j \delta_{y_j}$$

$$\mathrm{OT}_\varepsilon(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\mathrm{P}:\mathrm{P}\mathbb{1}=\alpha, \mathrm{P}^\top \mathbb{1}=\beta} \sum_{i,j} C_{ij} P_{ij} + \varepsilon E(\mathrm{P})$$

with $E(P) = \sum_{ij} P_{ij} log(P_{ij})$

- Can be solved using Sinkhorn algorithm: matrix product update only with element-wise operations

---

Cuturi, Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances. NeuRIPS 2013

Previously...

Entropic (regularized) optimal transport

$$\mathrm{OT}_\varepsilon(\alpha,\beta) \stackrel{\text{def.}}{=} \min_{\mathrm{P}:\mathrm{P}\mathbb{1}=\alpha,\mathrm{P}^\top\mathbb{1}=\beta} \sum_{i,j} C_{ij}P_{ij} + \varepsilon E(\mathrm{P})$$

with $E(P) = \sum_{ij} P_{ij} \log\left(\frac{P_{ij}}{\alpha_i\beta_j}\right)$

- Initialize: $\mathrm{K} = e^{-C/\varepsilon}, v = \mathbb{1}$
- Update till convergence:
  - $u = \frac{\alpha}{Kv}$
  - $v = \frac{\beta}{K^\top u}$
- $P_{ij} = u_i K_{ij} v_j$
- Element-wise operations: $\mathcal{O}(n^2)$; can be done in parallel with GPU

Cuturi, Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances. NeuRIPS 2013

Previously...

Entropic (regularized) optimal transport – general form

$$\mathrm{OT}_\varepsilon(\alpha, \beta) \overset{\text{def.}}{=} \min_{\pi : \pi_1 = \alpha, \pi_2 = \beta} \int \int C(x, y) d\pi(x, y) + \varepsilon E(\pi)$$

with $E(P) = \sum_{ij} P_{ij} \log \left(\frac{P_{ij}}{\alpha_i \beta_j} \right)$

- Solution always exists and unique
- Differentiable
- But: not a distance

Cuturi, Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances. NeuRIPS 2013

Outline

Partial Optimal Transport

Motivation: standard OT requires

- $\sum_i \alpha_i = \sum_j \beta_j$ (and usually equals 1).
- All of the mass from α needs to be transfer to β

$\longrightarrow$ Partial OT focuses on transporting a fraction of mass $0 \leq m \leq \min(\sum_i \alpha_i, \sum_j \beta_j)$

Partial Optimal Transport

A relaxation of constraint of the Kantorovic OT problem

$$\text{Partial OT}(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\mathrm{P}:\mathrm{P}\mathbb{1} \leq \alpha, \mathrm{P}^\top \mathbb{1} \leq \beta} \sum_{i,j} C_{ij} P_{ij} \quad \text{with} \quad \mathbb{1}^\top P \mathbb{1} = m$$

- Equality constraints are relaxed, now only need total transported mass to be equal to $m > 0$
- Allow distributions with different total mass when $m \leq \min(\mathbb{1}^\top \alpha, \mathbb{1}^\top \beta)$
- But: cannot be solved using linear programming/Sinkhorn because constraints are now different

Figalli. The optimal partial transport problem. Archive for Rational Mechanics and Analysis, 2010

Partial Optimal Transport

Solution: Adding dummy variables to make Partial OT become standard OT

$$\widetilde{\text{Partial OT}}(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\tilde{\mathrm{P}}:\tilde{\mathrm{P}}\mathbb{1}=\tilde{\alpha}, \tilde{\mathrm{P}}^\top \mathbb{1}=\tilde{\beta}} \sum_{i,j} \tilde{C}_{ij} \tilde{P}_{ij} \quad \text{, with}$$

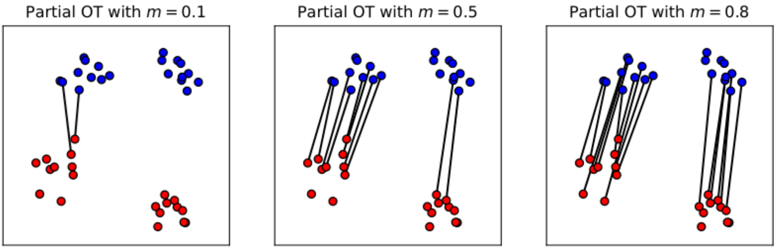
$$\tilde{P} = \begin{bmatrix} P & b \\ a^\top & 0 \end{bmatrix}, \tilde{C} = \begin{bmatrix} C & \xi \mathbb{1}_n \\ \xi \mathbb{1}_n^\top & 2\xi + c_{max} \end{bmatrix}, \tilde{\alpha} = [\tilde{\alpha}, \beta^\top \mathbb{1} - m], \tilde{\beta} = [\beta, \tilde{\alpha}^\top \mathbb{1} - m]$$

- This means: solving the augmented problem $\widetilde{\text{Partial OT}}$ to find P.

Chapel et al. Partial Optimal Transport with Applications on Positive-Unlabeled Learning. NeuRIPS 2020.

Partial Optimal Transport

Assuming initial mass $\sum_i \alpha_i = \sum_j \beta_j = 1.0$



$\longrightarrow$ With small m only a small fraction of the mass get transported, and vice versa

Unbalanced Optimal Transport

Another type of relaxation for the contraint: adding divergence as regularisation and removing mass contraint completely
$\longrightarrow$ Unbalanced OT

$$\mathrm{UOT}(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\mathrm{P}} \sum_{i,j} C_{ij} P_{ij} + \tau \mathrm{KL}(P||\alpha) + \tau \mathrm{KL}(P||\beta)$$

with $\mathrm{KL}(p||q)$ the Kullback-Leibler divergence

- $\tau \rightarrow +\infty$: standard OT
- $\tau \rightarrow 0$: some thing call the Hellinger distance:

$$H^2(\alpha, \beta) \stackrel{\text{def.}}{=} \frac{1}{2} \|\sqrt{\alpha} - \sqrt{\beta}\|_2^2$$

[Liereo, Mielke, Savaré 2015], [Chizat, Schmitzer, Peyré, Vialard 2015]

Unbalanced Optimal Transport

Another type of relaxation for the contraint: adding divergence as regularisation and removing mass contraint completely
$\longrightarrow$ Unbalanced OT

$$\mathrm{UOT}(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\mathrm{P}} \sum_{i,j} C_{ij} P_{ij} + \tau \mathrm{KL}(P||\alpha) + \tau \mathrm{KL}(P||\beta)$$

with $\mathrm{KL}(p||q)$ the Kullback-Leibler divergence

- $\tau \to +\infty$: standard OT
- $\tau \to 0$: some thing call the Hellinger distance:

$$H^2(\alpha, \beta) \stackrel{\text{def.}}{=} \frac{1}{2} \|\sqrt{\alpha} - \sqrt{\beta}\|_2^2$$

But: how to solve this problem given now it looks more complicated?

[Liereo, Mielke, Savaré 2015], [Chizat, Schmitzer, Peyré, Vialard 2015]

Unbalanced Optimal Transport

Entropic regularization to the rescue:

$$\mathrm{UOT}_{\varepsilon}(\alpha, \beta) \overset{\text{def.}}{=} \min_{\mathrm{P}} \sum_{i,j} C_{ij} P_{ij} + \tau \mathrm{KL}(P||\alpha) + \tau \mathrm{KL}(P||\beta) + \varepsilon \mathrm{KL}(P||\alpha \otimes \beta)$$

where $\alpha \otimes \beta \overset{\text{def.}}{=} \alpha\beta^{\top}$ is the measure product.

- $\mathrm{UOT}_{\varepsilon}$ objective is convex and differentiable
- Sinkhorn's algorithm update
  - $u = \left(\frac{\alpha}{Kv}\right)^{1+\varepsilon/\tau} \qquad v = \left(\frac{\beta}{K^{\top} u}\right)^{1+\varepsilon/\tau}$
- $P_{ij} = u_i K_{ij} v_j$
- Note: formula is simplified, only for KL-divergence; but can be any type of divergence belongs to the so-called f-divergence

Chizat, Schmitze, Peyré, Vialard. Scaling algorithms for unbalanced optimal transport problems. Mathematics of Computation 2018.

Outline

Motivation for Gromov-Wasserstein distance

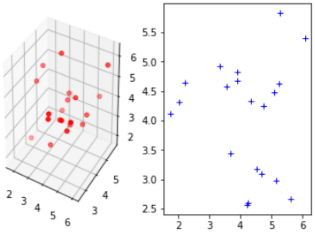
Reminder: the OT problem we define above is technicall called Wasserstein distance

$$\alpha = \sum_{i=1}^{n} \alpha_i \delta_{x_i} \quad \beta = \sum_{j=1}^{m} \beta_j \delta_{y_j}$$

$$W_p^p(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\mathrm{P}:\mathrm{P}\mathbb{1}=\alpha, \mathrm{P}^\top \mathbb{1}=\beta} \sum_{i,j} C_{ij} P_{ij} \quad \text{with} \quad C_{ij} = d(x_i, y_j)^p$$

However...

Motivation for Gromov-Wasserstein distance



- Objective: matching points between $\mathcal{X}$ a 3D surface and $\mathcal{Y}$ 2D surface
- How to measure distance between the 3D and 2D space? ($d(x_i, y_j)$ does not exist)

  $\longrightarrow$ Need to define different kind of distance

Gromov-Wasserstein distance

$$(D, \alpha) \quad \alpha = \sum_{i=1}^{n} \alpha_i \delta_{x_i} \quad D_{i,i'} = d(x_i, x_{i'})$$

$$(\bar{D}, \beta) \quad \beta = \sum_{j=1}^{m} \beta_j \delta_{y_j} \quad \bar{D}_{j,j'} = d(x_i, x_{j'})$$

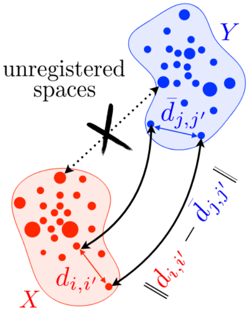
$\longrightarrow$ Gromov Wasserstein distance

$$\mathrm{GW}_p^p(D, \alpha, \bar{D}, \beta) \overset{\text{def.}}{=} \mathcal{E}_{D,\bar{D}}^p = \min_{\mathrm{P}:\mathrm{P}\mathbb{1}=\alpha, \mathrm{P}^\top \mathbb{1}=\beta} \sum_{i,i',j,j'} |D_{i,i'} - \bar{D}_{j,j'}|^p P_{i,j} P_{i',j'}$$

- GW-2 defines a distance (up to isometries – skip definition) [Memoli 2011]
- Search for transport plans that preserve the pairwise relationships between samples

Memoli (2011); Sturm (2012)

Gromov-Wasserstein distance



General formulation:

$$\mathrm{GW}_2^2(d_X, \alpha, d_Y, \beta) \stackrel{\text{def.}}{=} \min_{\pi:\pi_1=\alpha,\pi_2=\beta} \int_{X^2 \times Y^2} |d_X(x, x') - d_y(y, y')|^2 \, d\pi(x, y) \, d\pi(x', y')$$

Solving GW problem

- Non-convex
- NP-hard to solve (means: very long time to find solutions, if they exist)

Solving GW problem

- Non-convex
- NP-hard to solve (means: very long time to find solutions, if they exist)

  $\longrightarrow$ Solution 1: Entropic-regularized Gromov-Wasserstein

Entropic Gromov-Wasserstein

$$\mathrm{GW}_p^p(D, \alpha, \bar{D}, \beta) \stackrel{\text{def.}}{=} \min_{\mathrm{P}:\mathrm{P}\mathbb{1}=\alpha, \mathrm{P}^\top \mathbb{1}=\beta} \sum_{i,i',j,j'} |D_{i,i'} - \bar{D}_{j,j'}|^p P_{i,j} P_{i',j'} - \varepsilon \sum_{i,j} P_{i,j} \log\left(\frac{P_{i,j}}{\alpha_i \beta_j}\right)$$

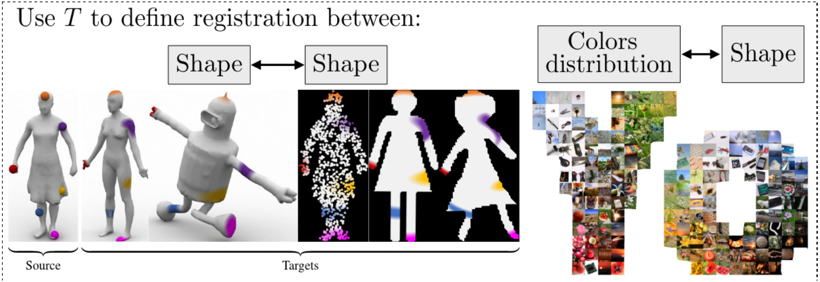
$\longrightarrow$ Sinkhorn's algorithm update

- Initialize $P = \alpha \otimes \beta$
- Repeat until convergence:
  - $\widetilde{P} = -DP\bar{D}$
  - $P = \mathtt{sinkhorn}(\alpha, \beta, \widetilde{P})$

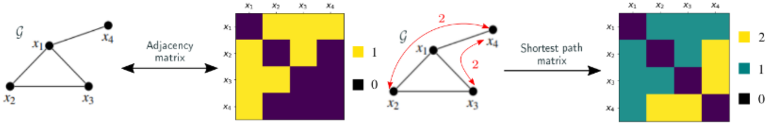
Note: technically the algorithm we solve above is projected mirror descent [Benamou et al. 2015]

Peyré, Cuturi, Solomon. Gromov-wasserstein averaging of kernel and distance matrices. ICML 2016

Application: shape analysis



Application: graph learning



- Caveat: AFAIK current OT works deal only undirected graphs $\mathcal{G} \stackrel{\text{def.}}{=} (V, E)$ with n nodes
- $V \stackrel{\text{def.}}{=} \{x_i\}_{i \in [n]}$ set of nodes (vertices)
- $E \stackrel{\text{def.}}{=} \{(x_i, x_j)\}_{x_i, x_j \in V}$
- Possible distance matrices: Adjacency matrix, graph Laplacian, geodesic (shortest path distance)

Fused Gromov-Wasserstein Distance

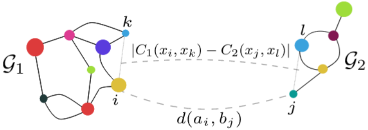


Figure 2. FGW loss E_q for a coupling π depends on both a similarity between each feature of each node of each graph $(d(a_i, b_j))_{i,j}$ and between all intra-graph structure similarities $(|C_1(x_i, x_k) - C_2(x_j, x_l)|)_{i,j,k,l}$.

- Each node in V now represents a feature ($x \in \mathbb{R}^d$)
- Fused GW: interpolating between Wasserstein and Gromov-Wasserstein distance

---

Vayer et al. Optimal Transport for structured data with application on graphs. ICML 2019

Fused Gromov-Wasserstein Distance

For $\alpha \in [0, 1]$:

$$\mathrm{FGW}^p_{p,\alpha}(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\mathrm{P}:\mathrm{P}\mathbb{1}=\alpha, \mathrm{P}^\top\mathbb{1}=\beta} \sum_{i,j,i',j'} \{(1-\alpha) d(a_i, b_j)^p + \alpha |d_X(x_i, y_k) - d_Y(x_j, y_l)|^p P_{i,j} P_{k,l}\}$$

Interpolating between Wasserstein and Gromov-Wasserstein distance:

- $\lim_{\alpha \to 0} \mathrm{FGW}_{p,\alpha}(\alpha, \beta) = \mathrm{W}_p(\alpha, \beta)^p$
- $\lim_{\alpha \to 1} \mathrm{FGW}_{p,\alpha}(\alpha, \beta) = \mathrm{GW}_p(\alpha, \beta)^p$
- Define a metric for $p = 1$ and semi-metric for $p > 1$

Vayer et al. Optimal Transport for structured data with application on graphs. ICML 2019

Fused Gromov-Wasserstein Distance

Algorithm 1 Conditional Gradient (CG) for FGW

$\pi^{(0)} \leftarrow \mu_X \mu_Y^\top$
for $i = 1, \ldots,$ **do**
  $G \leftarrow$ Gradient from Eq. (7) *w.r.t.* $\pi^{(i-1)}$
  $\tilde{\pi}^{(i)} \leftarrow$ Solve OT with ground loss G
  $\tau^{(i)} \leftarrow$ Line-search for loss (1) with $\tau \in (0, 1)$ using Alg. 2
  $\pi^{(i)} \leftarrow (1 - \tau^{(i)})\pi^{(i-1)} + \tau^{(i)}\tilde{\pi}^{(i)}$
end for

When $p = 2$:

- Gradient of FGW can be factorized, similar to [Peyré et al 2016]
- Finding optimal plan with Conditional Gradient (Frank-Wolfe) method

Vayer et al. Optimal Transport for structured data with application on graphs. ICML 2019

Unbalanced Gromov-Wasserstein Distance

Similar to standard OT: relaxing the mass contraint, but for GW distance

$$
\begin{aligned}
\mathrm{UGW}_p(\alpha, \beta)^p &\stackrel{\text{def.}}{=} \min_P \mathcal{L}(P) \\
&= \min_P \sum_{i,i',j,j'} |D_{i,i'} - \bar{D}_{j,j'}|^p P_{i,j} P_{i',j'} + \\
&\qquad \tau \mathrm{KL}(P_1 \otimes P_1 || \alpha \otimes \alpha) + \tau \mathrm{KL}(P_2 \otimes P_2 || \beta \otimes \beta)
\end{aligned}
$$

- Note that now the divergence term are between tensor product measure $\rightarrow$ quadratic divergence
- Solutions exist on compact space (and a additional technical condition)
- However: NP-hardness to find the minimizer, not proper distance

Séjourné et al. 2021. The Unbalanced Gromov Wasserstein Distance: Conic Formulation and Relaxation. ICML 2021

Unbalanced Gromov-Wasserstein Distance

Idea: ease up computation by entropic regularization

$$\mathrm{UGW}_{p,\varepsilon}(\alpha, \beta)^p \stackrel{\text{def.}}{=} \min_P \mathcal{L}(P) + \varepsilon \mathrm{KL}(P \otimes P || \alpha \otimes \beta)$$

But: computation is heavy, no Sinkhorn-update scheme available
$\longrightarrow$ For special case $p = 2$, lower bound with a different term that can be efficiently approximate with Sinkhorn-algorithm

$$\mathrm{UGW}_{2,\varepsilon}(\alpha, \beta) \geq \inf_{P,G} \mathcal{F}(P, G) + \varepsilon \mathrm{KL}(P \otimes G || \alpha \otimes \beta)^2$$

where

$$\mathcal{F}(P, G) \stackrel{\text{def.}}{=} \sum_{i,j,k,l} |d_X(x_i, y_k) - d_Y(x_j, y_l)|^2 P_{i,j} G_{k,l} +$$
$$\mathrm{KL}(P_1 \otimes G_1 || \alpha \otimes \alpha) + \mathrm{KL}(P_2 \otimes G_2 || \beta \otimes \beta)$$

Séjourné et al. 2021. The Unbalanced Gromov Wasserstein Distance: Conic Formulation and Relaxation. ICML 2021

Unbalanced Gromov-Wasserstein Distance

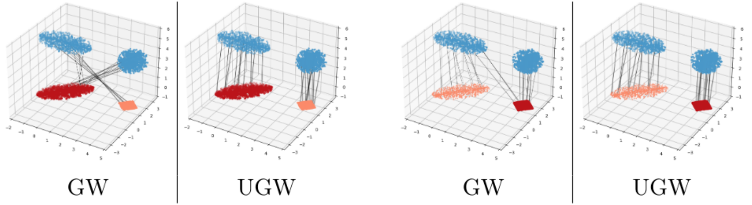


Figure 3: GW vs. UGW transportation plan, using $\nu = 0.3\mathcal{E}_2 + 0.7\mathcal{C}$ on the left, and $\nu = 0.7\mathcal{E}_2 + 0.3\mathcal{C}$ on the right. The 2D mm-spaces is lifted into $\mathbb{R}^3$ by padding the third coordinate to zero.