Optimal Transport in ML

Rémi Gilleron

Inria Lille & CRIStAL & Univ. Lille

Feb. 2018

Main Source – also some figures

Computational Optimal transport, G. Peyré and M. Cuturi

Compare documents using word embeddings

Given: word embeddings in $\mathbb{R}^d$, a similarity on $\mathbb{R}^d$
Problem: compare documents (phrases, sentences)
Idea: a document as an histogram of frequencies of words in V

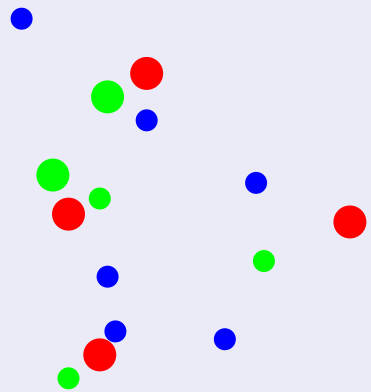


Figure: Three documents in $\mathbb{R}^2$. Each circle corresponds to a word vector. The size of the circle is proportional to the frequency.

Measure the transportation plan between histograms over the vocabulary using the similarity between word vectors

Unsupervised domain adaptation

Unsupervised domain adaptation

Source: labeled data; target: unlabeled data; problem: classify data in the target domain.

Base idea: transport data from the source domain into the target domain, then learn a classifier. The transportation plan between the two clouds use a ground distance and empirical distributions of the clouds

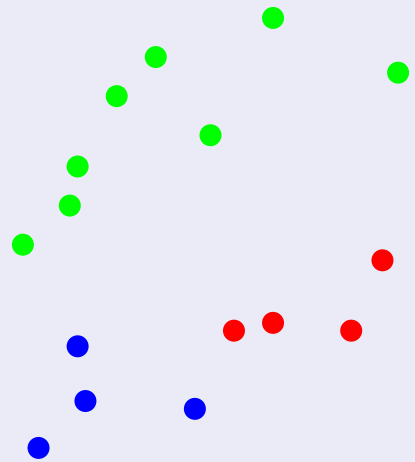


Figure: Source: blue and red; target: green

Histogram propagation over graphs

Example of traffic estimation

- Given: a graph representing roads
- Given: at some nodes, captors allow to compute traffic histograms over a 24h period
- Problem: compute traffic histograms for every node
- Base idea: use a propagation algorithm over the graph
- The similarity should be based both on a similarity between histograms and a similarity on the graph which includes spatial information.

Optimal Transport (OT)

What is it?

A method for comparing probability distributions with the ability to incorporate spatial information

Pros

- Distance between distributions based on a ground distance
- Defined for all distributions: discrete, with density, arbitrary
- Solid mathematical foundations and works well in applications

Cons

- Computing OT is solving an optimization problem
- Mathematics are not easy. Methods and algorithms depend on
  - the measures (often discrete) and dimensions
  - the ground cost: arbitrary, distance, squared distance, geodesic distance
  - the ground space: $\mathbb{R}$, $\mathbb{R}^d$, geodesic

Research on computational OT

Close to Magnet's research problems

- Word mover distance
- Optimal transport for domain adaptation
- Histogram propagation over graphs

But also

- Signal and image processing,
- Wasserstein distances and divergences,
- Efficient computation of (regularized) OT,
- Generative models, Wasserstein GANs,
- among others.

Plan

Intuitions

The goal of OT is to define geometric tools useful to compare probability distributions

Earth Mover Distance

- probability distribution = pile of sand
- move one pile of sand into another one
- local cost: move one grain of sand from one place to another
- OT = minimal global cost

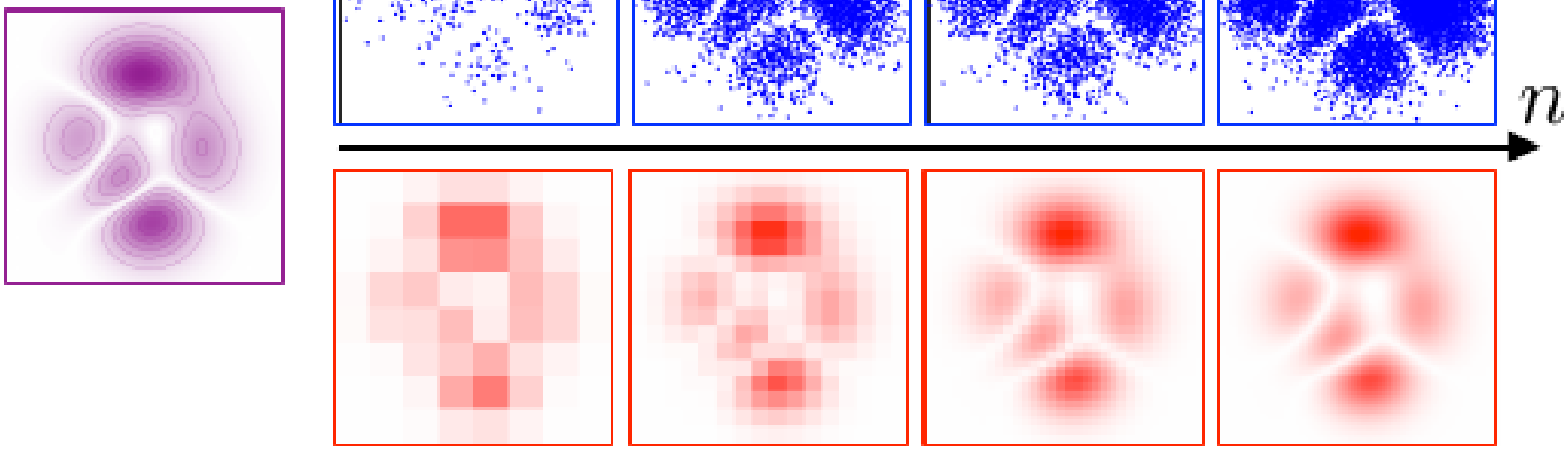
Mines and Factories Problem

- Mines produce ressources across a country
- Factories consume ressources across a country
- Local cost for distributing one ressource from a mine to a factory
- OT = least costly transportation plan from mines into factories

Main Scenarii

Distributions

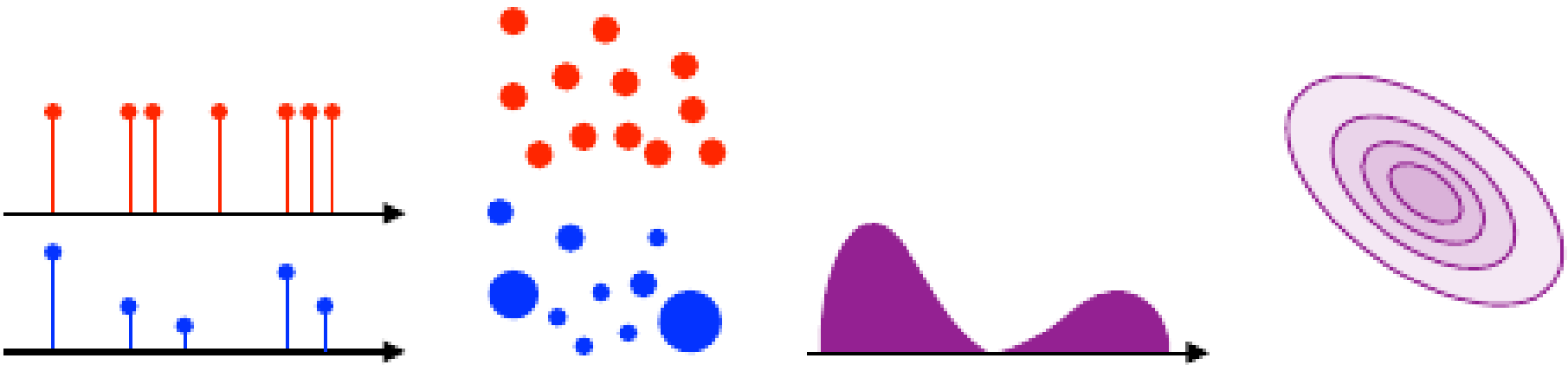
- Discrete measure $\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i}$ where δ_x is the Dirac at x and $\mathbf{a} \in \mathbb{R}_+^n$. Then, for $f \in \mathcal{C}(\mathcal{X})$, $\int_{\mathcal{X}} f(x) d\alpha(x) = \sum_{i=1}^n \mathbf{a}_i f(x_i)$.
  - Probability measure: $\sum_{i=1}^n \mathbf{a}_i = 1$, $\mathbf{a} \in \Sigma_n$ is an histogram
  - Lagrangian (point clouds): $(x_i)_{i=1}^{i=n}$, $\mathbf{a}_i = \frac{1}{n}$
  - Eulerian (histograms): grid, $\mathbf{a}_i$ probability mass at cell i



Main Scenarii

Distributions

- Discrete measure $\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i}$ where δ_x is the Dirac at x and $\mathbf{a} \in \mathbb{R}_+^n$. Then, for $f \in \mathcal{C}(\mathcal{X})$, $\int_{\mathcal{X}} f(x) d\alpha(x) = \sum_{i=1}^n \mathbf{a}_i f(x_i)$.
  - Probability measure: $\sum_{i=1}^n \mathbf{a}_i = 1$, $\mathbf{a} \in \Sigma_n$ is an histogram
  - Lagrangian (point clouds): $(x_i)_{i=1}^{i=n}$, $\mathbf{a}_i = \frac{1}{n}$
  - Eulerian (histograms): grid, $\mathbf{a}_i$ probability mass at cell i
- Measure with density α $d\alpha(x) = \rho_\alpha(x) dx$. Then, for $h \in \mathcal{C}(\mathcal{X})$, $\int_{\mathcal{X}} h(x) d\alpha(x) = \int_{\mathcal{X}} h(x) \rho_\alpha(x) dx$.
- Arbitrary measure



Discrete $d = 1$ Discrete $d = 2$ Density $d = 1$ Density $d = 2$

Main Scenarii

Distributions

- Discrete measure $\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i}$ where δ_x is the Dirac at x and $\mathbf{a} \in \mathbb{R}_+^n$. Then, for $f \in \mathcal{C}(\mathcal{X})$, $\int_{\mathcal{X}} f(x) d\alpha(x) = \sum_{i=1}^n \mathbf{a}_i f(x_i)$.
  - Probability measure: $\sum_{i=1}^n \mathbf{a}_i = 1$, $\mathbf{a} \in \Sigma_n$ is an histogram
  - Lagrangian (point clouds): $(x_i)_{i=1}^{i=n}$, $\mathbf{a}_i = \frac{1}{n}$
  - Eulerian (histograms): grid, $\mathbf{a}_i$ probability mass at cell i
- Measure with density α $d\alpha(x) = \rho_\alpha(x)dx$. Then, for $h \in \mathcal{C}(\mathcal{X})$, $\int_{\mathcal{X}} h(x) d\alpha(x) = \int_{\mathcal{X}} h(x)\rho_\alpha(x)dx$.
- Arbitrary measure

Ground space and ground cost

- $\mathcal{X}$ with a distance and, in general, $\mathcal{X} = \mathbb{R}^d$ with $d = 1$ or $d > 1$
- c is a cost function from $\mathcal{X} \times \mathcal{Y}$ in $\mathbb{R}^+$. When $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, in general, c is a distance d or a squared distance d^2.
  It is a matrix $\mathbf{C}$ in the case of discrete measures.

Monge Problem – Transport Map

Monge Problem for Discrete Measures

- Let $\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i}$ and $\beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j}$ be two discrete measures. Find a map T from $\{x_1, \ldots, x_n\}$ into $\{y_1, \ldots, y_m\}$ such that, $\forall j$, $\mathbf{b}_j = \sum_{\{i | T(x_i) = y_j\}} \mathbf{a}_i$. T defines a pushforward operator $T_\sharp$ between measures s.t. $T_\sharp \alpha = \beta$.
- The transport map should minimize $\sum_i c(x_i, T(x_i))$.

Exemples

- $(x_1, 1)$, $(x_2, 2)$, $(x_3, 3)$, $(x_4, 1)$, $(x_5, 1)$; $(y_1, 4)$, $(y_2, 2)$, $(y_3, 2)$, $c = 1$
- Id with $c = d$, the Euclidean distance
- $(x_1, 2)$, $(x_2, 2)$, $(x_3, 2)$; $(y_1, 1)$, $(y_2, 2)$, $(y_3, 2)$, $(y_4, 1)$, $c = 1$
- $(x_1, 3)$, $(x_2, 2)$; $(y_1, 4)$, $(y_2, 1)$, $c = 1$

Monge problem for histograms

Optimal assignment problem

- Let $n = m$, let $\mathbf{a} = \mathbf{b} = \mathbf{1}_n/n$, and let $\mathbf{C}$ be a cost matrix in $\mathbb{R}_+^n \times \mathbb{R}_+^n$ giving the cost of moving x_i into y_j
- Find σ in the set $Perm(n)$ of permutations of n elements solving

$$\min_{\sigma \in Perm(n)} \frac{1}{n} \sum_{i=1}^{i=n} \mathbf{C}_{i,\sigma(i)}$$

Remarks

- Naive algorithm untractable because $Perm(n)$ has $n!$ elements
- $((2,1),\frac{1}{2})$, $((2,3),\frac{1}{2})$; $((1,2),\frac{1}{2})$, $((3,2),\frac{1}{2})$, Euclidean distance
- $(x_1,\frac{1}{2})$, $(x_2,\frac{1}{2})$, $(y_1,\frac{1}{4})$, $(y_2,\frac{1}{4})$; $(y_3,\frac{1}{4})$, $(y_4,\frac{1}{4})$, $c = 1$

Relaxation of the Monge problem

Limitations of the Monge problem

- Feasible solutions may not exist
- Multiple solutions may exist
- Assignment problem is combinatorial
- Monge problem for arbitrary measures is not convex
- Existence and unicity of the Monge map for c squared Euclidean distance and measures with density (Brenier 91)

Relax the deterministic nature of transportation

- $(x_1, 3)$, $(x_2, 2)$; $(y_1, 4)$, $(y_2, 1)$, $c = 1$.
- Kantorovitch's relaxation: mass splitting from a source towards several targets
- a coupling T maps x_1 into y_1, $\frac{1}{2}$ of the mass in x_2 into y_1 and into y_2
- Where are transported x_1 and x_2?

Kantorovitch's OT problem for discrete measures

Formulation

- Let $\mathbf{a} \in \mathbb{R}_+^n$ and $\mathbf{b} \in \mathbb{R}_+^m$ be two mass vectors for $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ and let $\mathbf{C}$ be the cost matrix
- A coupling is defined by a matrix $\mathbf{P} \in \mathbb{R}_+^{n \times m}$ where $\mathbf{P}_{i,j}$ is the amount of mass flowing from x_i to y_j
- Set of admissible couplings is $\mathbf{U}(\mathbf{a}, \mathbf{b}) = \{\mathbf{P} \mid \mathbf{P}\mathbf{1}_m = \mathbf{a},\ \mathbf{P}^t\mathbf{1}_n = \mathbf{b}\}$. It is bounded and it is a convex polytope.
- Kantorovitch's OT problem is

$$L_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a},\mathbf{b})} < \mathbf{C}, \mathbf{P} >= \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a},\mathbf{b})} \sum_{i,j} \mathbf{C}_{i,j}\mathbf{P}_{i,j}$$

Mines and Factories

Find an optimal transportation plan between mines and factories

Kantorovitch's OT problem

Formulation for arbitrary measures

- Let α and β be two measures, a coupling π is a joint distribution over $\mathcal{X} \times \mathcal{Y}$
- Set of admissible couplings $\mathcal{U}(\alpha, \beta) = \{\pi \mid \mathcal{P}_{\mathcal{X}\sharp}\pi = \alpha \text{ and } \mathcal{P}_{\mathcal{Y}\sharp}\pi = \beta\}$ where $P_{\mathcal{X}\sharp}$ and $P_{\mathcal{Y}\sharp}$ are the push-forward projections.
- Kantorovitch's OT problem for a cost function c is

$$\mathcal{L}_c(\alpha, \beta) = \min_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y)$$

For discrete measures

- Let $\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i}$ and $\beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j}$ be two discrete measures
- Then $\mathcal{L}_c(\alpha, \beta) = L_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$ where $\mathbf{C}$ is the cost matrix defined from c on the support of α and β.

Examples of Kantorovitch's OT problem

$\mathcal{X} = \mathbb{R}$ and Euclidean distance

- $(x_1, 3)$, $(x_2, 2)$; $(y_1, 4)$, $(y_2, 1)$ with $x_1 < x_2$, $y_1 < y_2$
- $(x_1, 3)$, $(x_2, 2)$; $(y_1, 4)$, $(y_2, 1)$ with $x_1 < x_2$, $y_1 > y_2$

Binary cost matrix

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \mathbf{a} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \mathbf{b} = \begin{bmatrix} 2/3 \\ 1/6 \\ 1/6 \end{bmatrix} \mathbf{P} = \begin{bmatrix} ? & ? & ? \\ ? & ? & ? \\ ? & ? & ? \end{bmatrix}$$

Assignment problem on $\mathcal{X} = \mathbb{R}^2$ with the Euclidean distance

$x_1 = (0, 1)$, $x_2 = (0, 2)$, $x_3 = (0, 3)$, $y_1 = (1, \frac{5}{2})$, $y_2 = (1, \frac{3}{2})$, $y_3 = (2, 2)$.

$$\mathbf{C} = \begin{bmatrix} \sqrt{5/2} & \sqrt{5/4} & \sqrt{5} \\ \sqrt{5/4} & \sqrt{5/4} & 2 \\ \sqrt{5/4} & \sqrt{5/2} & \sqrt{5} \end{bmatrix} \mathbf{a} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \mathbf{b} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \mathbf{P} = \begin{bmatrix} ? & ? & ? \\ ? & ? & ? \\ ? & ? & ? \end{bmatrix}$$

Examples of Kantorovitch's OT problem

$\mathcal{X} = \mathbb{R}$ and Euclidean distance

- $(x_1, 3)$, $(x_2, 2)$; $(y_1, 4)$, $(y_2, 1)$; $x_1 < x_2$, $y_1 < y_2$: $x_1 \to y_1$; $1/2x_2 \to y_1$; $1/2x_2 \to y_2$
- $(x_1, 3)$, $(x_2, 2)$; $(y_1, 4)$, $(y_2, 1)$; $x_1 < x_2$, $y_1 > y_2$: $1/3x_1 \to y_2$; $2/3x_1 \to y_1$; $x_2 \to y_1$

Binary cost matrix

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \mathbf{a} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \mathbf{b} = \begin{bmatrix} 2/3 \\ 1/6 \\ 1/6 \end{bmatrix} \mathbf{P} = \begin{bmatrix} 1/3 & 0 & 0 \\ 1/6 & 1/6 & 0 \\ 1/6 & 0 & 1/6 \end{bmatrix}$$

Assignment problem on $\mathcal{X} = \mathbb{R}^2$ with the Euclidean distance

$$\mathbf{C} = \begin{bmatrix} \sqrt{5/2} & \sqrt{5/4} & \sqrt{5} \\ \sqrt{5/4} & \sqrt{5/4} & 2 \\ \sqrt{5/4} & \sqrt{5/2} & \sqrt{5} \end{bmatrix} \mathbf{a} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \mathbf{b} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

Kantorovitch relaxation is tight for assignment problems

Permutation matrices are couplings

- Let $n = m$, let $\mathbf{a} = \mathbf{b} = \mathbf{1}_n/n$, let $\mathbf{C}$ be a cost matrix in $\mathbb{R}_+^{n \times n}$
- Kantorovitch's problem $L_{\mathbf{C}}(\mathbf{1}_n/n, \mathbf{1}_n/n) = \min_{\mathbf{P} \in \mathbf{U}(\mathbf{1}_n/n, \mathbf{1}_n/n)} < \mathbf{C}, \mathbf{P} >$
- For $\sigma \in Perm(n)$, the permutation matrix $\mathbf{P}_\sigma$ is in $\mathbf{U}(\mathbf{1}_n/n, \mathbf{1}_n/n)$

Kantorovitch for matching

- Proposition: There exists an optimal solution of Kantorovitch's problem which is a permutation matrix associated with an optimal permutation for the assignment problem (optimal transport map for Monge problem between histograms).
- Proof: extremal points of $\mathbf{U}(\mathbf{1}_n/n, \mathbf{1}_n/n)$ are permutations (Birkhoff's Theorem) and minimum of a linear objective is reached at extremal points of the polyhedron (Bertsimas and Tsitsiklis).

Plan

Definition of the Wasserstein Distance

OT defines a distance between measures when C satisfies some properties

p-Wasserstein distance between measures

Let $\mathcal{X} = \mathcal{Y}$, let d be a distance on $\mathcal{X}$, let $p \geq 1$, and let $c = d^p$, then $\mathcal{W}_p$ is a distance on $\mathcal{M}_+^1(\mathcal{X})$ where $\mathcal{W}_p(\alpha, \beta) = \mathcal{L}_{d^p}(\alpha, \beta)^{1/p}$

p-Wasserstein Distance between histograms

Let $n = m$, let $\mathcal{X} = \mathcal{Y} = \Sigma_n$, let D be a distance, let $p \geq 1$, and let $\mathbf{C} = D^p$ the matrix of $D_{i,j}^p$,

then W_p is a distance on Σ_n where $W_p(\mathbf{a}, \mathbf{b}) = L_{D^p}(\mathbf{a}, \mathbf{b})^{1/p}$

First Remarks

- W_p (resp. $\mathcal{W}_p$) depends on D (resp. d)
- W_1 (resp. $\mathcal{W}_1$) is the optimal transport cost for $C = D$ (resp. $c = d$)

p-Wasserstein distance is a distance

Proof for histograms

Note that $\mathbf{C} = D^p$ is symmetric and has a null diagonal. Then,

- $W_p(\mathbf{a}, \mathbf{b}) = 0$ if and only if $\mathbf{a} = \mathbf{b}$ easy
- $W_p(\mathbf{a}, \mathbf{b}) = W_p(\mathbf{b}, \mathbf{a})$ easy
- $W_p(\mathbf{a}, \mathbf{c}) \leq W_p(\mathbf{a}, \mathbf{b}) + W_p(\mathbf{b}, \mathbf{c})$
  - Let $\mathbf{S} = \mathbf{P} diag(1/\bar{\mathbf{b}}) \mathbf{Q}$ where $\mathbf{P}$ (resp. $\mathbf{Q}$) is an optimal coupling between $\mathbf{a}$ and $\mathbf{b}$ (resp. between $\mathbf{b}$ and $\mathbf{c}$), and $\bar{\mathbf{b}}$ is $\mathbf{b}$ where null values are set to 1
  - $W_p(\mathbf{a}, \mathbf{c}) \leq (< \mathbf{S}, D^p >)^{1/p}$ because $\mathbf{S} \in \mathbf{U}(\mathbf{a}, \mathbf{c})$
  - Then use the triangular inequality for D^p and the Minkowski inegality to show that $(< \mathbf{S}, D^p >)^{1/p} \leq W_p(\mathbf{a}, \mathbf{b}) + W_p(\mathbf{b}, \mathbf{c})$.

p-Wasserstein distance properties

Geometric intuition

- W_p is a distance while many others are divergences as, for instance, the KL-divergence
- W_p allows to compare singular distributions, for instance discrete ones. While classical distances or divergences do not allow to compare discrete distributions.
- W_p allows to quantify spatial shift between the supports
- Barycenters of distributions can be defined with $\bar{\alpha} = \arg\min_{\alpha} \sum_i \lambda_i \mathcal{W}_p^p(\alpha_i, \alpha)$.
- $W_p(\delta_x, \delta_y) = d(x, y)$ and $W_p(\delta_x, \delta_y) \to 0$ if $x \to y$. This allows to define a more general notion of weak convergence for distributions.

Plan

Special case: binary cost matrix

Binary cost matrix

Let $\mathbf{a}$ and $\mathbf{b}$ be two mass vectors in $\mathbb{R}_+^n$ and let $\mathbf{C}$ be the cost matrix $\mathbf{1}_{n\times n} - \mathbf{I}_{n\times n}$, then

$$L_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \frac{1}{2}\|\mathbf{a} - \mathbf{b}\|_1$$

Kronecker cost function and total variation

Let $\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i}$ and $\beta = \sum_{j=1}^m \mathbf{b}_j \delta_{x_j}$ be two discrete measures, let c be the Kronecker cost function defined by $c(x, y) = 0$ if $x = y$ and $c(x, y) = 1$ otherwise. Then

$$\mathcal{L}_c(\alpha, \beta) = TVD(\alpha, \beta) = \frac{1}{2}\|\alpha - \beta\|_1$$

$TVD(\alpha, \beta) = \sup_i |\mathbf{a}_i - \mathbf{b}_i|$ is the total variation distance between α and β.

Special case: dimension 1

Discrete case

- Let $\mathcal{X} = \mathbb{R}$, let $\alpha = \frac{1}{n}\sum_{i=1}^{i=n} \delta_{x_i}$ and $\beta = \frac{1}{n}\sum_{j=1}^{j=n} \delta_{y_j}$ and let us suppose that $x_1 < x_2 < \dots x_n$ and $y_1 < y_2 < \dots y_n$, then $\mathcal{W}_p$ is the L^p-norm between two vectors of ordered values of α and β.

$$\mathcal{W}_p^p(\alpha, \beta) = \sum_{i=1}^{i=n} |x_i - y_i|^p$$

- $\mathcal{W}_p(\alpha, \beta)$ changes as soon as the order changes
- Can be extended to the case $n \neq m$ with the condition: if $x_i < x_{i'}$, $P_{i,j} \neq 0$ and $P_{i',j'} \neq 0$, then necessarily $y_j \leq y_{j'}$

Arbitrary measure

$$\mathcal{W}_p(\alpha, \beta) = \|\mathcal{C}_\alpha^{-1} - \mathcal{C}_\beta^{-1}\|_{L^p((0,1])}^p$$

Special case: Gaussian distributions

Two Gaussians in $\mathbb{R}$ w.r.t. Euclidean distance

Let $\alpha = \mathcal{N}(m_1, \sigma_1)$ and $\beta = \mathcal{N}(m_2, \sigma_2)$, then

- the optimal transport map is $T(x) = m_2 + (x - m_1)\sqrt{\frac{\sigma_2}{\sigma_1}}$
- $W_2^2(\alpha, \beta) = \|m_2 - m_1\|^2 + \|\sqrt{\sigma_2} - \sqrt{\sigma_1}\|^2$

Two Gaussians in $\mathbb{R}^d$ w.r.t. Euclidean distance

Let $\alpha = \mathcal{N}(\mathbf{m}_\alpha, \mathbf{\Sigma}_\alpha)$ and $\beta = \mathcal{N}(\mathbf{m}_\beta, \mathbf{\Sigma}_\beta)$, then

- the optimal map T is $T(x) = \mathbf{m}_\beta + A(x - \mathbf{m}_\alpha)$ where
  $A = \mathbf{\Sigma}_\alpha^{-\frac{1}{2}}(\mathbf{\Sigma}_\alpha^{\frac{1}{2}}\mathbf{\Sigma}_\beta\mathbf{\Sigma}_\alpha^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{\Sigma}_\alpha^{-\frac{1}{2}} = A^t$
- $W_2^2(\alpha, \beta) = \|m_\alpha - m_\beta\|^2 + tr(\mathbf{\Sigma}_\alpha + \mathbf{\Sigma}_\beta - 2(\mathbf{\Sigma}_\alpha^{\frac{1}{2}}\mathbf{\Sigma}_\beta\mathbf{\Sigma}_\alpha^{\frac{1}{2}})^{\frac{1}{2}})$
- $W_2^2(\alpha, \beta) = \|m_\alpha - m_\beta\|^2 + \|\sqrt{\mathbf{r}} - \sqrt{\mathbf{s}}\|^2$ if $\mathbf{\Sigma}_\alpha = diag(\mathbf{r})$ and $\mathbf{\Sigma}_\beta = diag(\mathbf{s})$

Plan

Reminder

Main cases for discrete measures

- Points in $\mathbb{R}^d$ or cells: $(\mathbf{x}_i)_{i=1}^n$, $(\mathbf{y}_j)_{j=1}^m$
- Discrete measures $\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i}$ and $\beta = \sum_{j=1}^m \mathbf{b}_j \delta_{y_j}$, $\mathbf{a} \in \mathcal{X} = \mathbb{R}_+^n$, $\mathbf{b} \in \mathcal{Y} = \mathbb{R}_+^m$, $\mathbf{a}_i$ mass at $\mathbf{x}_i$ and $\mathbf{b}_j$ mass at $\mathbf{y}_j$
- $\mathcal{X} = \mathbb{R}_+^n$, $\mathcal{X} = \Sigma_n$ (histograms: sum is 1), $\mathbf{a}_i = 1/n$, case $n = m$
- c is a cost, c is a distance, c is the Euclidean distance, c is the squared Euclidean distance, ... $\mathbf{C}$ is the cost matrix of $c_{i,j}$

OT and Wasserstein distance

- $L_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a},\mathbf{b})} < \mathbf{C}, \mathbf{P} > = \min_{\mathbf{P}\mathbf{1}_m = \mathbf{a},\ \mathbf{P}^t \mathbf{1}_n = \mathbf{b}} \sum_{i,j} \mathbf{C}_{i,j} \mathbf{P}_{i,j}$
- Let $n = m$ and $\mathcal{X} = \mathcal{Y} = \Sigma_n$, let d be a distance, let $p \geq 1$, and let $\mathbf{C} = d^p$ the matrix of $d_{i,j}^p$, then W_p is a distance on Σ_n where $W_p(\mathbf{a}, \mathbf{b}) = L_{\mathbf{C}}(\mathbf{a}, \mathbf{b})^{1/p}$. Note that $W_p(\delta_x, \delta_y) = d(x, y)$.

OT is a linear program

Kantorovitch linear program

- Let us recall that Kantorovitch's OT problem is

$$L_{\mathbf{C}}(\mathbf{a},\mathbf{b}) = \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} < \mathbf{C},\mathbf{P} >= \min_{\mathbf{P}\mathbf{1}_m=\mathbf{a},\ \mathbf{P}^t\mathbf{1}_n=\mathbf{b}} \sum_{i,j} \mathbf{C}_{i,j}\mathbf{P}_{i,j}$$

- It can be formulated as

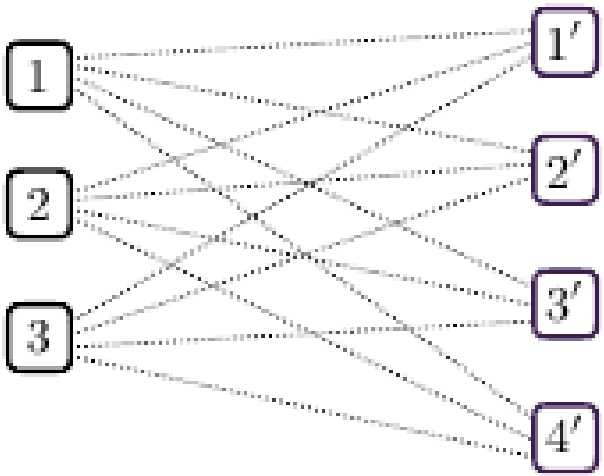
$$L_{\mathbf{C}}(\mathbf{a},\mathbf{b}) = \min_{\mathbf{p}}(\mathbf{c}^t\mathbf{p}) \text{ s.t. } \mathbf{p}\in\mathbb{R}_+^{nm},\ \mathbf{A}\mathbf{p} = \begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix}$$

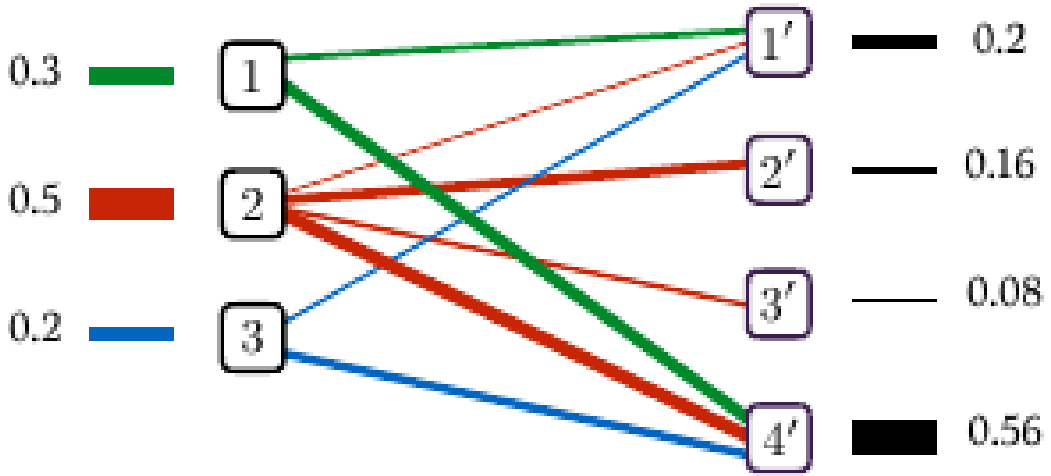
where the $n \times m$ matrices $\mathbf{P}$ and $\mathbf{C}$ have been replaced by nm dimensional vectors $\mathbf{p}$ and $\mathbf{c}$, and admissible couplings are defined with a well chosen matrix $\mathbf{A}$.

Simplex network algorithm for OT

Algorithm – complexity in $O(n^3)$

- Because one can restrict the search to extremal points of the polytope $\mathbf{U}(\mathbf{a}, \mathbf{b})$ and using the structure of matrices $\mathbf{P}$ expressed with bipartite graphs, one can use a network flow solver.
- For matching problems ($n{=}m$, $\mathbf{a}_i = \mathbf{b}_i = \frac{1}{n}$), the auction algorithm runs in $n^3 \|\mathbf{C}\|_\infty / \epsilon$ and the cost of the output is $n\epsilon$ suboptimal





Regularized Optimal Transport

Adding a regularization penalty

The Kantorovitch's regularized OT problem is

$$L_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} < \mathbf{C}, \mathbf{P} > + \lambda \Omega(\mathbf{P})$$

What for?

- Encode prior knowledge
- Better posed problem w.r.t. stability because more dense couplings
- Smooth approximate distance w.r.t. input histogram weights and positions of the Diracs
- Better complexity making the problem convex
- Regularization: quadratic, entropic, group Lasso, KL divergence, ...

Entropic Regularized Optimal Transport

Entropic Regularization

The entropic regularized OT problem (Cuturi 2013) is

$$L_{\mathbf{C}}^{\lambda}(\mathbf{a},\mathbf{b}) = \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} < \mathbf{C},\mathbf{P} > -\lambda H(\mathbf{P})$$

$$\text{I.e. } L_{\mathbf{C}}^{\lambda}(\mathbf{a},\mathbf{b}) = \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} < \mathbf{C},\mathbf{P} > +\lambda \sum_{i,j} \mathbf{P}_{i,j}(\log(\mathbf{P}_{i,j}))$$

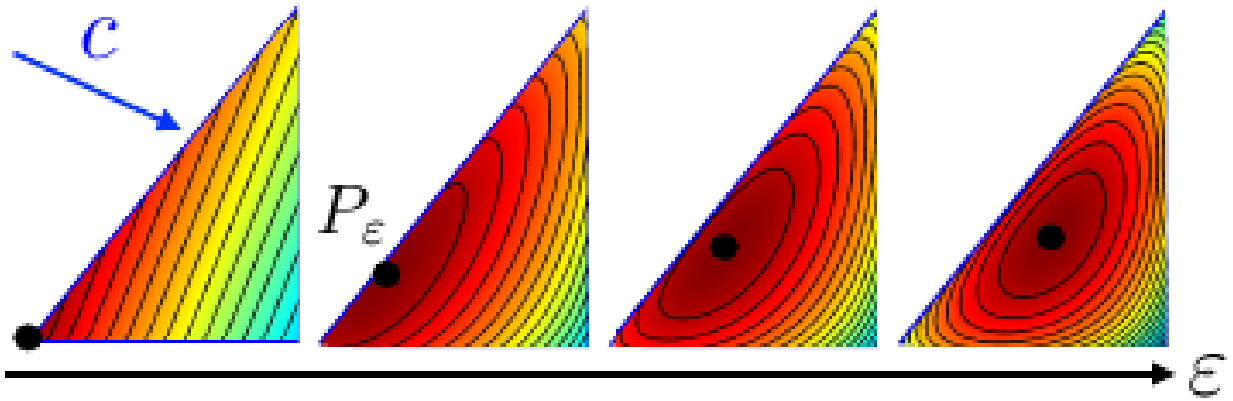


Figure 4.1: Impact of ε on the optimization of a linear function on the simplex, solving $\mathbf{P}_\varepsilon = \text{argmin}_{\mathbf{P}\in\Sigma_3} \langle \mathbf{C}, \mathbf{P}\rangle - \varepsilon\mathbf{H}(\mathbf{P})$ for a varying ε.

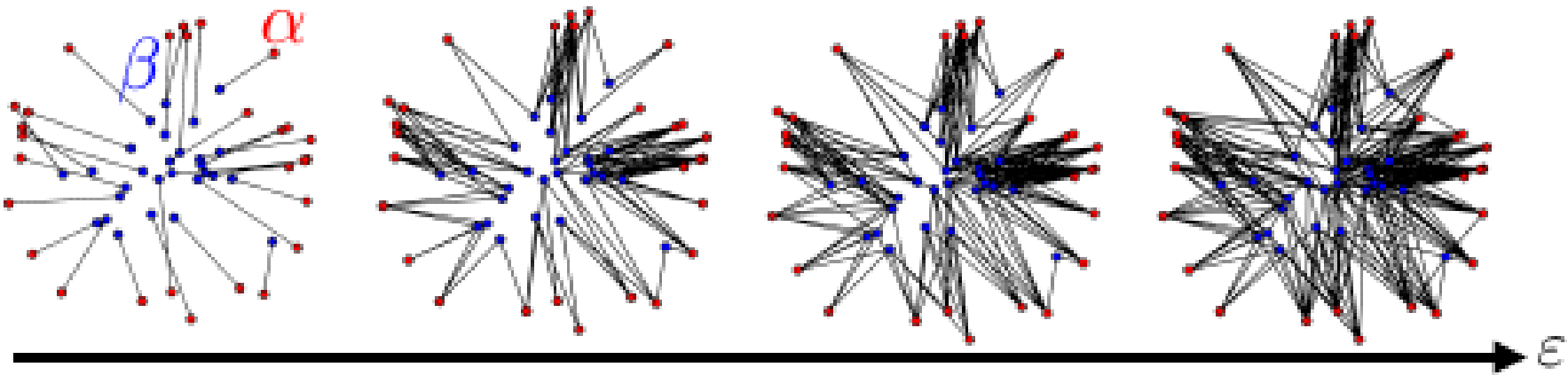
Entropic Regularized Optimal Transport

Entropic Regularization

The entropic regularized OT problem (Cuturi 2013) is

$$L_{\mathbf{C}}^{\lambda}(\mathbf{a},\mathbf{b}) = \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} < \mathbf{C},\mathbf{P} > -\lambda H(\mathbf{P})$$

$$\text{I.e. } L_{\mathbf{C}}^{\lambda}(\mathbf{a},\mathbf{b}) = \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} < \mathbf{C},\mathbf{P} > +\lambda \sum_{i,j} \mathbf{P}_{i,j}(\log(\mathbf{P}_{i,j}))$$



Entropic Regularized Optimal Transport

Entropic Regularization

The entropic regularized OT problem (Cuturi 2013) is

$$L_{\mathbf{C}}^{\lambda}(\mathbf{a},\mathbf{b}) = \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} < \mathbf{C},\mathbf{P} > -\lambda H(\mathbf{P}) \tag{1}$$

$$\text{I.e. } L_{\mathbf{C}}^{\lambda}(\mathbf{a},\mathbf{b}) = \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} < \mathbf{C},\mathbf{P} > +\lambda \sum_{i,j} \mathbf{P}_{i,j}(\log(\mathbf{P}_{i,j}))$$

Convergence with λ

- Entropy is strongly convex. Objective is λ-strongly convex.
- Proposition: the unique solution $\mathbf{P}_\lambda$ of problem 1 converges to the optimal solution with maximal entropy within the set of all optimal solutions of the Kantorovitch's OT problem. In particular,
  $L_{\mathbf{C}}^{\lambda}(\mathbf{a},\mathbf{b}) \overset{\lambda\to 0}{\longrightarrow} L_{\mathbf{C}}(\mathbf{a},\mathbf{b})$.
  Also, $L_{\mathbf{C}}^{\lambda}(\mathbf{a},\mathbf{b}) \overset{\lambda\to\infty}{\longrightarrow} \mathbf{a}\bigotimes\mathbf{b} = \mathbf{a}\mathbf{b}^t = (\mathbf{a}_i\mathbf{b}_j)_{i,j}$

Computing Entropic Regularized Optimal Transport

Entropic regularized OT as matrix scaling

- Introducing dual variables, expressing the Lagrangian of (1), it can be shown that the solution of (1) has the form $\mathbf{P} = diag(\mathbf{u})\mathbf{K}diag(\mathbf{v})$ for (unknown) vectors $\mathbf{u} \in \mathbb{R}_+^n$, $\mathbf{v} \in \mathbb{R}_+^m$, and $\mathbf{K}$ defined by $\mathbf{K}_{i,j} = e^{-\frac{\mathbf{C}_{i,j}}{\lambda}}$
- leads to the Sinkhorn's algorithm:
  Init $\mathbf{v}^{(0)} = \mathbf{1}_m$
  Repeat $\mathbf{u}^{(l+1)} = \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}^{(l)}}$ **and** $\mathbf{v}^{(l+1)} = \frac{\mathbf{b}}{\mathbf{K}^t\mathbf{u}^{(l+1)}}$

Results – complexity in $O(n^2)$

- Only matrix operations
- Convergence but numerical problems and difficulties for small λ
- Complexity in $O(n^2 \log(n)\epsilon^{-3})$
- GPU version when solving multiple OT problems

Conclusion on the first part

Also in Computational Optimal transport, G. Peyré and M. Cuturi

- Semi-discrete OT: one is discrete and one is arbitrary (often with density)
- $\mathcal{W}_1$ OT, i.e. c is a distance
- $\mathcal{W}_2$ OT for a geodesic distance
- Approximating OT with discrete samples (Eulerian or Lagrangian)
- Variational OT, i.e. using Wasserstein distance as a loss function
- Algorithms for computing Wasserstein barycenters

Software for OT

Python Optimal Transport library by Rémy Flamary
`https://github.com/rflamary/POT`

Plan

A new distance between text documents

Base ideas

- Word embeddings: represent every word by a vector
- Word mover's distance (WMD): the distance between two text documents A and B is the minimum cumulative distance that words from document A need to travel to match exactly the point cloud of document B
- Kusner et al study k-NN with the WMD for classifying documents
- Huang et al define metric learning algorithms for the WMD

References

- From word embeddings to document distances, Kusner et al, ICML'15
- Supervised word mover's distance, Huang et al, NIPS'16

From word embeddings to document distances

Representation of text documents

- A vocabulary of size n
- The i-th word w_i is represented by a word vector $\mathbf{x}_i \in \mathbb{R}^d$
- A text document A is represented as an histogram $\mathbf{a} \in \Sigma_n$ defined by $\mathbf{a}_i = \frac{c_i}{\sum_{i=1}^{i=n} c_i}$ where c_i is the word count for w_i in A

Word mover's distance between text documents

- Cost or ground distance is chosen to be the Euclidean distance
- Word mover's distance between text documents A and B is the W_1 distance between $\mathbf{a}$ and $\mathbf{b}$, i.e.

$$WMD(A, B) = W_1(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P}\mathbf{1}_n = \mathbf{a},\ \mathbf{P}^t\mathbf{1}_n = \mathbf{b}} \sum_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|_2 \mathbf{P}_{i,j}$$

WMD for text classification

k-NN with WMD – works well 8-)

- Outperforms known methods on several datasets
- The word2vec embeddings works well on several domains
- Largest runtime because the time complexity of computing the WMD is $O(q^3 \log q)$, where q is the max number of unique words in A or B
- n is too large for GPU computation of multiple WMD with entropy regularization

Prefetch and prune

- The authors introduce two lower bounds for $WMD(A, B)$:
  - Word centroid distance: $WCD(A, B) = \|\mathbf{Xa} - \mathbf{Xb}\|_2$
  - Relaxed WMD: $RWMD = max\{WMD_1(A, B), WMD_2(A, B)\}$
- Use the lower bounds to compute faster an approximation of the k-nearest neighbours for a text query.

Supervised WMD for text classification

WMD to be learned

- Squared generalized Euclidean distance: $c(i,j) = \|\mathbf{A}(\mathbf{x}_i - \mathbf{x}_j)\|_2^2$
- Histogram reweighting: A is represented as $\tilde{\mathbf{a}} = (\mathbf{a} \circ \mathbf{w})/(\mathbf{w}^t\mathbf{a})$
- Then, $WMD_{\mathbf{A},\mathbf{w}} = \min_{\mathbf{P}\mathbf{1}_n=\tilde{\mathbf{a}},\ \mathbf{P}^t\mathbf{1}_n=\tilde{\mathbf{b}}} \sum_{i,j} \|\mathbf{A}(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \mathbf{P}_{i,j}$

Learning the WMD from labeled data

Learn $\mathbf{A} \in \mathbb{R}^{r \times d}$ and $\mathbf{w} \in \mathbb{R}^n$ s.t. $WMD_{\mathbf{A},\mathbf{w}}$ reflects labels. Method:

- Stochastic neighborhood relaxation of the LOO loss as for NCA
- Express the gradient w.r.t. $\mathbf{A}$ and the gradient w.r.t. $\mathbf{w}$
- Compute gradients using entropic regularization ($O(q^2)$ w.r.t. $O(q^3)$) on a subset of neighbors using WCD
- Clever initialization using WCD
- Use batch stochastic gradient descent

Conclusion on WMD for texts

Pros

- Well-defined distance between text documents based on OT
- It works well 8-) for classification with k-NN

Cons = my comments 8-)

- Choice of the ground distance is ad hoc
  - Euclidean distance. Why not the cosine distance?
  - Squared Euclidean distance when gradient computation is needed
- Many tricks for supervised WMD to be solved "efficiently": loss, choice of the neighbors with WCD, initialization, regularization
- Experimental results for supervised WMD not convincing
- Perspective: WMD with Gaussian embeddings; learn Gaussian embeddings with a WMD-based loss.

Plan

Domain adaptation with regularized optimal transport

Source: Courty et al, ECML 2014, IEEE PAMI

Unsupervised domain adaptation

Source: labeled data; target: unlabeled data; problem: classify data in the target domain.

Base idea: transport data from the source domain into the target domain, then learn a classifier.

Figure: Left: source labeled data (blue and red) and target data (green); right: source points are transported and a linear classifier H is learned

Solution of Courty et al

The OT formulation

- Ground distance: squared Euclidean distance
- Entropy regularization for efficiency
- Introducing domain adaptation in OT: the coupling should be s.t. a target point should not receive mass from points with different labels
- This leads to $\mathbf{P}^* = \underset{\mathbf{P} \in \mathbf{U}(\mathbf{a},\mathbf{b})}{\arg\min} < \mathbf{C}, \mathbf{P} > -\lambda H(\mathbf{P}) + \eta L(\mathbf{P})$ where $L(\mathbf{P})$ express group label sparsity of $\mathbf{P}$ with a L^1-norm
- Use an alternating optimization algorithm

Learning in the target domain

- Given a solution $\hat{\mathbf{P}}$, every source point $\mathbf{x}_i$ is transported to the barycenter $\hat{\mathbf{x}}_i$ of its images
- Learn a classifier on the images $\hat{\mathbf{x}}_i$ with the labels of the $\mathbf{x}_i$.

Discussion on domain adaptation with OT

Pros

- Elegant formulation of domain adaptation based on OT
- It works well 8-) for balanced problems
- Other regularizers and other algorithms in long version
- Extended to the semi-supervised case in another paper

Cons = my comments 8-)

- Antagonism between
  - $-\lambda H(\mathbf{P})$ promoting non sparsity
  - $\eta L(\mathbf{P})$ promoting group label sparsity
- Choice of η is not discussed. And choice of λ not so easy.

Mapping estimation for discrete optimal transport

Source: Perrot et al, NIPS'16

Motivations

- let us consider an OT problem in Kantorovitch's formulation for point clouds $(\mathbf{x}_i)_{i=1}^{i=n}$ and $(\mathbf{y}_j)_{j=1}^{j=m}$
- The optimal coupling $\mathbf{P}^*$ allows to define a transportation map T by

$$T(\mathbf{x}_i) = \arg\min_{\mathbf{y}\in\mathcal{Y}} \sum_{j=1}^{j=m} \mathbf{P}(i,j) d(\mathbf{y}, \mathbf{y}_j)$$

- I.e. T maps $\mathbf{x}_i$ into the barycenter of its images. It is the weighted average when d is the Euclidean distance on $\mathcal{Y}$
- But T is defined only for every source point $\mathbf{x}_i$

Idea: learn a transformation T from $\mathcal{X}$ into $\mathcal{Y}$

Mapping estimation for discrete optimal transport

Formulation of Perrot et al

- they propose the following optimization problem

$$\underset{T\in\mathcal{H},\mathbf{P}\in\mathbf{U(a,b)}}{\arg\min} \frac{1}{nd_t}\|T(\mathbf{X_s}) - n\mathbf{P}\mathbf{X_t}\|_{\mathcal{F}}^2 + \frac{\lambda}{max(\mathbf{C})} < \mathbf{P}, \mathbf{C} > + \frac{\gamma}{d_s d_t} R(T)$$

- Problem jointly convex if $\mathcal{H}$ is a convex set of transformations and R is a convex function
- $\mathcal{H}$ si considered to be a set of linear transformations induced by a linear matrix or non linear using kernels

Comments and results

- theoretical bounds are provided but ...
- alternating optimization algorithm
- It works 8-)

Plan

Summary

À retenir 8-)

- OT theory allows to define distances between distributions using spatial information
- For every type of distributions
- Computing OT (or Wasserstein distance) is solving an optimization problem.
- Complexity in $O(n^3)$, reduced to $O(n^2)$ with entropic regularization
- Mathematical foundations but not so easy to understand

OT in the discussed papers

- ad hoc choice of the ground distance
- intricate optimization problems derived from OT
- ad hoc optimization algorithms

But many opportunities to use OT in magnet research problems

Conclusion

For Magnet

- NLP: Mangoes, cosine distance, Gaussian embeddings, applications
- Domain adaptation? Distributed learning?
- For Jan et al, histogram prediction in graphs, Wasserstein propagation for semi-supervised learning, Solomon et al, ICML'14

Some recent papers among others at NIPS'17 and ICLR'18

- Joint Distribution Optimal Transportation for Domain Adaptation
- Near-linear time approximation algorithms for optimal transport ...
- Large Scale Optimal Transport and Mapping Estimation
- Improved Training of Wasserstein GANs
- Learning Wasserstein Embeddings
- Wasserstein Auto-Encoders
- Improving GANs Using Optimal Transport