Advanced Section #2: Optimal Transport

AC 209B: Data Science 2

Javier Zazo Pavlos Protopapas

Lecture Outline

Historical overview

Definitions and formulations

Metric properties about optimal transport

Application I: Supervised learning with Wasserstein Loss

Application II: Domain adaptation

Historical overview

The origins of optimal transport

666. MÉMOIRES DE L'ACADÉMIE ROYALE

MÉMOIRE

SUR LA

THÉORIE DES DÉBLAIS

ET DES REMBLAIS.

Par M. MONGE.

- Gaspard Monge proposed the first idea in 1781.
- How to move dirt from one place (d'eblais) to another (remblais) with minimal effort?
- Enunciated the problem of finding a mapping F between two distributions of mass.
- Optimization with respect to a displacement cost $c(x, y)$.

Transportation problem I

- Formulated by Frank Lauren Hitchcock in 1941.

Factories & warehouses example

- Fixed number of factories, each of which produces good at a fixed output rate.
- Fixed number of warehouses, each of which has a fixed storage capacity.
- There is a cost to transport goods from a factory to a warehouse.
- **Goal:** Find the transportation of goods from factory $\rightarrow$ warehouse with lowest possible cost.

Transportation problem II: Example

Factories:

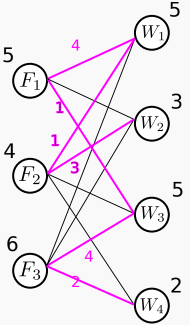
- F_1 makes 5 units.
- F_2 makes 4 units.
- F_3 makes 6 units.

Warehouses:

- W_1 can store 5 units.
- W_2 can store 3 units.
- W_3 can store 5 units.
- W_4 can store 2 units.

Transportation costs:

	W_1	W_2	W_3	W_4
F_1	5	4	7	6
F_2	2	5	3	5
F_3	6	3	4	4



Transportation problem III:

- One factory can transport product to multiple warehouses.
- One warehouse can receive product from multiple factories.
- The Transportation problem can be formulated as an ordinary linear constrained optimization problem (LP):

$$
\begin{aligned}
\min_{x_{ij}} \quad & 5x_{11} + 4x_{12} + 7x_{13} + 6x_{14} + 2x_{21} + 5x_{22} \\
& +3x_{23} + 2x_{24} + 6x_{31} + 3x_{32} + 4x_{33} + 4x_{34} \\
\text{s.t.} \quad & x_{11} + x_{12} + x_{13} + x_{14} = 5 \\
& x_{21} + x_{22} + x_{23} + x_{24} = 4 \\
& x_{31} + x_{32} + x_{33} + x_{34} = 6 \\
& x_{11} + x_{21} + x_{31} \leq 5 \\
& x_{12} + x_{22} + x_{32} \leq 3 \\
& x_{13} + x_{23} + x_{33} \leq 5 \\
& x_{14} + x_{24} + x_{34} \leq 2
\end{aligned}
$$

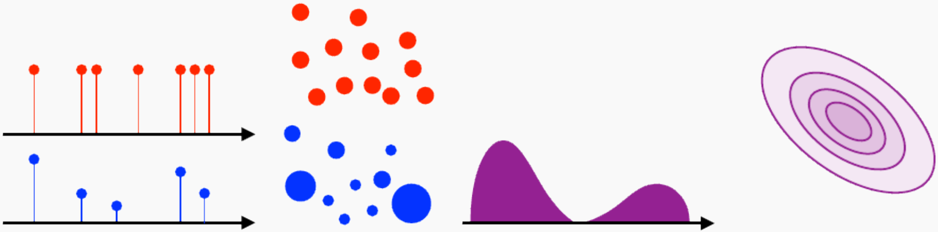
Definitions and formulations

Definitions

- Probability simplex:
$$\Delta_n = \left\{ a_i \in \mathbb{R}_+^n \;\middle|\; \sum_{i=1}^{n} a_i = 1 \right\}$$
- Discrete probability distribution: $\mathbf{p} = (p_1, p_2, \dots, p_n) \in \Delta_n$.
- Space $\mathcal{X}$: support for the distritution (coordinates vector/array, temperature, etc.).
- **Discrete measure:** given weights $\mathbf{p} = (p_1, p_2, \dots, p_n)$ and $\mathbf{x} = (x_1, x_2, \dots, x_n)$ locations,
$$\alpha = \sum_i p_i \delta_{x_i}$$
- **Radon measure:** $\alpha \in \mathcal{M}(\mathcal{X})$,
  - $\mathcal{X}$ is equipped with a distance, integrating it against a continuous function f
$$\int_{\mathcal{X}} f(x) d\alpha(x) \overset{\mathbb{R}^d}{=} \int_{\mathcal{X}} f(x) \rho_\alpha(x) dx$$

More definitions

- Set of positive measures: $\mathcal{M}_+$, such that $\int_{\mathcal{X}} f(x) d\alpha(x) \rightarrow \mathbb{R}_+$.
- **Set of probability measures:** $\mathcal{M}_+^1$, such that $\int_{\mathcal{X}} d\alpha(x) = 1$.

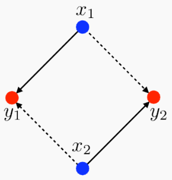


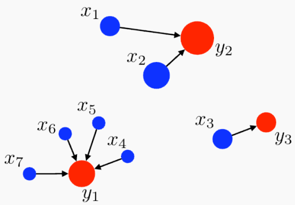
Discrete $d = 1$ Discrete $d = 2$ Density $d = 1$ Density $d = 2$

Assingment and Monge problems

- n origin elements (**factories**),
- $m = n$ destination elements (**warehouses**),
- we look for a permutation (an assignment in the general case) of elements

$$\min_{\sigma \in \mathrm{Perm(n)}} \quad \frac{1}{n} \sum_{i=1}^{n} C_{i,\sigma(i)}$$





- The set of n discrete elements has $n!$ possible permutations.
- Works after Monge, aimed to simplify the problem, such as Hitchcock in 1941, or Kantorovich in 1942.

Kantorovich relaxation

- **Goal:** find a minimal transport plan $\mathbf{F}$ such that

$$\mathbf{F} \in U(\mathbf{p}, \mathbf{q}) = \{\, \mathbf{F} \in \mathbb{R}_{+}^{n \times n} \;\mid\; \mathbf{F}\mathbf{1} = \mathbf{p} \text{ and } \mathbf{F}^T\mathbf{1} = \mathbf{q} \,\}$$

- $\mathbf{F}\mathbf{1} = \mathbf{p}$ sum the rows of $\mathbf{F}$ $\rightarrow$ all goods are transported from $\mathbf{p}$.
- $\mathbf{F}^T\mathbf{1} = \mathbf{q}$ sum the columns of $\mathbf{F}$ $\rightarrow$ all goods are received in $\mathbf{q}$.
- $\mathbf{p}$ and $\mathbf{q}$ are probability distributions $\rightarrow$ mass is conserved and equals 1.

Relation to linear programming

- The Kantorovich problem is an LP:

$$\boxed{\begin{aligned} L_{\mathbf{C}}(\mathbf{p},\mathbf{q}) = \min_{\mathbf{F}\geq 0} \quad & \operatorname{tr}(\mathbf{FC}) \\ & \mathbf{F1} = \mathbf{p}, \quad \mathbf{F}^T\mathbf{1} = \mathbf{q} \end{aligned}} \tag{1}$$

- LP programs can be solved with ***simplex method***, ***interior point methods***, ***dual descent methods***, etc. The problem is **convex**.
- One option is to use LP solvers: Clp, Gurobi, Mosek, SeDuMi, CPLEX, ECOS, etc.
- **Spezialized methods exist** (and Python, C, Julia, etc. libraries)
  - Network simplex
  - Approximate methods: Sinkhorn, smoothed versions, etc.

Kantorovich formulation for arbitrary measures

- Now **C** needs to be a function:

$$c(x,y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$$

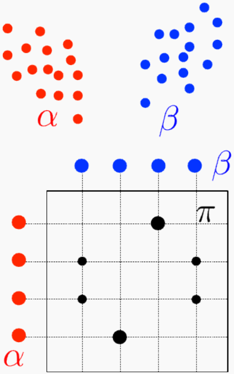
- Discrete measures $\alpha = \sum_i p_i \delta_{x_i}$ and $\beta = \sum_i q_i \delta_{y_i}$:
  - $c(x,y)$ is still a matrix where costs depends on locations of measures.
- For arbitrary probabilistic measures:
  - Define a coupling $\pi \in \mathcal{M}_+^1(\mathcal{X},\mathcal{Y}) \to$ joint probability distribution of $\mathcal{X}$ and $\mathcal{Y}$.

$$U(\alpha,\beta) = \left\{ \pi \in \mathcal{M}_+^1(\mathcal{X},\mathcal{Y}) \,\middle|\, P_{\mathcal{X}\sharp}\pi = \alpha \text{ and } P_{\mathcal{Y}\sharp}\pi = \beta \right\}$$

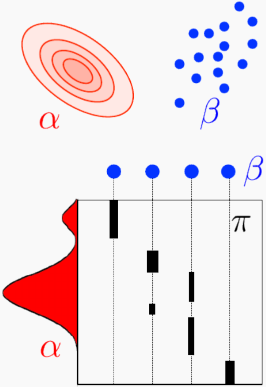
  - The continuous problem:

$$\mathcal{L}_c(\alpha,\beta) = \min_{\pi \in U(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\pi(x,y) = \min_{(X,Y)} \left\{ \mathbb{E}_{(X,Y)}(c(X,Y)) \,\middle|\, X \sim \alpha, Y \sim \beta \right\}$$

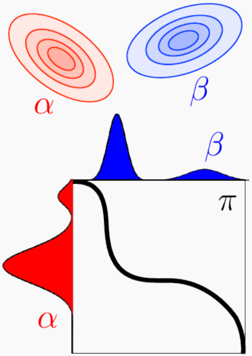
Example of transport maps for arbitrary measures



Discrete



Semidiscrete



Continuous

Metric properties about optimal transport

- Wasserstein distance is also referred as OT, or Earth mover's distance (EMD).

Discrete Wasserstein distance

Consider $\mathbf{p}, \mathbf{q} \in \Delta_n$ and

$$\mathbf{C} \in \mathcal{C}_n = \left\{ \mathbf{C} \in \mathbb{R}_+^{n \times n} \,\middle|\, \mathbf{C} = \mathbf{C}^T, \mathrm{diag}(\mathbf{C}) = 0 \text{ and } \forall (i,j,k) \quad C_{i,j} \leq C_{i,k} + C_{k,j} \right\}.$$

Then,

$$W_p(\mathbf{p}, \mathbf{q}) = L_{\mathbf{C}^p}(\mathbf{p}, \mathbf{q})^{1/p}$$

defines a ***p*-Wasserstein** distance on Δ_n.

- Recall that $L_{\mathbf{C}}(\mathbf{p}, \mathbf{q})$ refers to the discrete Kantorovich problem:

$$L_{\mathbf{C}}(\mathbf{p}, \mathbf{q}) = \left\{ \min \ \mathrm{tr}(\mathbf{F}\mathbf{C}) \,\middle|\, \mathbf{F} \geq 0, \quad \mathbf{F}\mathbf{1} = \mathbf{p}, \quad \mathbf{F}^T\mathbf{1} = \mathbf{q} \right\}$$

Proof that p-Wasserstein constitutes a distance

- We need to show **positivity**, **symmetry** and **triangular inequality**.
- Since $\operatorname{diag}(\mathbf{C}) = 0$, $W_p(\mathbf{p}, \mathbf{p}) = 0$, and $\mathbf{F}^* = \operatorname{diag}(\mathbf{p})$.
- Because of strict positivity of off-diagonal elements, $W_p(\mathbf{p}, \mathbf{q}) = \operatorname{tr}(\mathbf{CF}) > 0$ for $\mathbf{p} \neq \mathbf{q}$.
- Since $W_p(\mathbf{p}, \mathbf{q}) = \operatorname{tr}(\mathbf{CF})$, and $\mathbf{C}$ is symmetric, $W_p(\mathbf{p}, \mathbf{q}) = W_p(\mathbf{q}, \mathbf{p})$.
- For triangularity, define $\mathbf{p}$, $\mathbf{q}$ and $\mathbf{t}$ and

$$\mathbf{F} = \operatorname{sol}(W_p(\mathbf{p}, \mathbf{q})) \qquad \mathbf{G} = \operatorname{sol}(W_p(\mathbf{q}, \mathbf{t})).$$

- For simplicity, assume $\mathbf{q} > 0$ (detailed proof in the lecture notes). Define

$$\mathbf{S} = \mathbf{F} \operatorname{diag}(1/\mathbf{q})\mathbf{G} \in \mathbb{R}_+^{n \times n}.$$

- Note that $\mathbf{F} \in U(\mathbf{p}, \mathbf{t})$, i.e., is a feasible transport plan:

$$\mathbf{S1} = \mathbf{F} \operatorname{diag}(1/\mathbf{q}) \underbrace{\mathbf{G1}}_{\mathbf{q}} = \mathbf{F} \underbrace{\operatorname{diag}(\mathbf{q}/\mathbf{q})}_{\mathbf{1}} = \mathbf{F1} = \mathbf{p}$$

$$\mathbf{S}^T\mathbf{1} = \mathbf{G}^T \operatorname{diag}(1/\mathbf{q}) \underbrace{\mathbf{F}^T\mathbf{1}}_{\mathbf{q}} = \mathbf{G}^T \underbrace{\operatorname{diag}(\mathbf{q}/\mathbf{q})}_{\mathbf{1}} = \mathbf{G}^T\mathbf{1} = \mathbf{t}$$

Wasserstein distance for arbitrary measures

Wasserstein distance for arbitrary measures

Consider $\alpha(x) \in \mathcal{M}_+^1(\mathcal{X}), \beta(y) \in \mathcal{M}_+^1(\mathcal{Y})$, $\mathcal{X} = \mathcal{Y}$, and for some $p \geq 1$,

- $c(x, y) = c(y, x) \geq 0$;
- $c(x, y) = 0$ if and only if $x = y$;
- $\forall (x, y, z) \in \mathcal{X}^3, c(x, y) \leq c(x, z) + c(z, y)$

Then,

$$W_p(\alpha, \beta) = \mathcal{L}_{c^p}(\alpha, \beta)^{1/p}$$

defines a ***p*-*Wasserstein*** distance on $\mathcal{X}$.

- Recall, that the Kantorovich problem for arbitrary measures is given by:

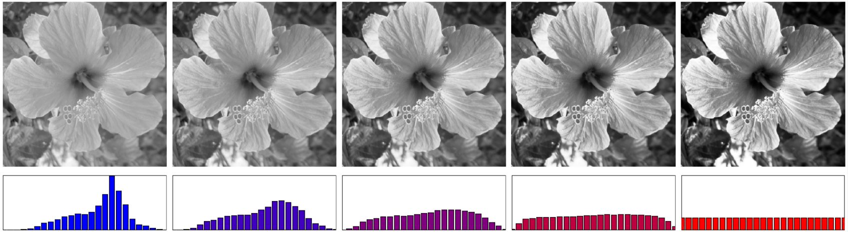
$$\mathcal{L}_c(\alpha, \beta) = \min_{\pi \in U(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y)$$

Special cases I

- Binary cost matrix: If $\mathbf{C} = \mathbf{1}\mathbf{1}^T - \mathbf{I}$, then $L_{\mathbf{C}}(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1$.
- 1D case of empirical measures:
  - $\mathcal{X} = \mathbb{R}$; $\alpha = \frac{1}{n}\sum_i \delta_{x_i}$ $\beta = \frac{1}{n}\sum_i \delta_{y_i}$;
  - $x_1 \leq x_2, \ldots \leq x_n$ and $y_1 \leq y_2, \ldots \leq y_n$ ordered observations.

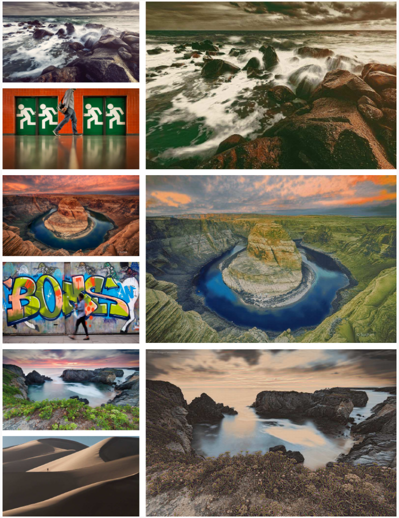
$$W_p(\mathbf{p}, \mathbf{q})^p = \sum_{i=1}^{n} |x_i - y_i|^p$$

- Histogram equalization:



$t = 0$ $t = 0.25$ $t = 0.5$ $t = .75$ $t = 1$

Color transfer

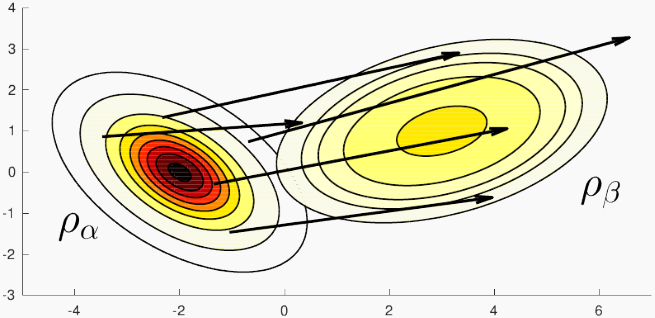


Special cases II: Distance between Gaussians

- If $\alpha = \mathcal{N}(\mathbf{m}_\alpha, \mathbf{\Sigma}_\alpha)$ and $\beta = \mathcal{N}(\mathbf{m}_\beta, \mathbf{\Sigma}_\beta)$ are two gaussians in $\mathbb{R}^d$,
- The following map:

$$T : x \to \mathbf{m}_\beta + A(x - \mathbf{m}_\alpha)$$

  where $A = \mathbf{\Sigma}_\alpha^{-\mathbf{1/2}}(\mathbf{\Sigma}_\alpha^{\mathbf{1/2}}\mathbf{\Sigma}_\beta\mathbf{\Sigma}_\alpha^{\mathbf{1/2}})^{\mathbf{1/2}}\mathbf{\Sigma}_\alpha^{-\mathbf{1/2}}$ constitutes an optimal transport plan.
- Furthermore,$W_2^2(\alpha, \beta) = \|\mathbf{m}_\alpha - \mathbf{m}_\beta\|^2 + \mathrm{tr}(\mathbf{\Sigma}_\alpha + \mathbf{\Sigma}_\beta - \mathbf{2}(\mathbf{\Sigma}_\alpha^{\mathbf{1/2}}\mathbf{\Sigma}_\beta\mathbf{\Sigma}_\alpha^{\mathbf{1/2}})^{\mathbf{1/2}})^{\mathbf{2}}$.



Application I: Supervised learning with Wasserstein Loss

Learning with Wasserstein Loss

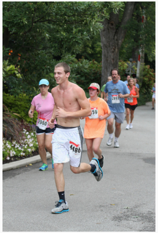
- Natural metric on the outputs that can be used to improve predictions.
- Wasserstein distance provides a natural notion of dissimilarity for probability measures $\longrightarrow$ Can encourage smoothness on the predictions.
  - In ImageNet, 1000 categories may have inherent semantic relationships.
  - Speech recognition systems, output correspond to keywords that also have semantic relations $\rightarrow$ this correlation can be exploited.

Siberian husky   Eskimo dog

Semantic relationships: Flickr dataset

(a) **Flickr user tags**: street, parade, dragon; **our proposals**: people, protest, parade; **baseline proposals**: music, car, band.

(b) **Flickr user tags**: water, boat, reflection, sunshine; **our proposals**: water, river, lake, summer; **baseline proposals**: river, water, club, nature.



(a) **Flickr user tags**: zoo, run, mark; **our proposals**: running, summer, fun; **baseline proposals**: running, country, lake.

(b) **Flickr user tags**: travel, architecture, tourism; **our proposals**: sky, roof, building; **baseline proposals**: art, sky, beach.

(c) **Flickr user tags**: spring, race, training; **our proposals**: road, bike, trail; **baseline proposals**: dog, surf, bike.

Problem setup

- **Goal:** Learn a mapping $\mathcal{X} \subset \mathbb{R}^d \to \mathcal{K} \subset \mathcal{Y} = \mathbb{R}_+^K$, where $|\mathcal{K}| = K$.
- Assume $\mathcal{K}$ possesses a metric $d_{\mathcal{K}}(\cdot, \cdot)$, or ground metric.
- Learning over a hypothessis space $\mathcal{H}$ of predictors: $h_\theta : \mathcal{X} \to \mathcal{Y}$, param. by $\theta \in \Theta$.
  - These can be a logistic regression, output of a NN, etc.
- Empirical risk minimization:

$$\min_{h_\theta \in \mathcal{H}} \quad \mathbb{E}\left\{l(h_\theta(x), y)\right\} \approx \frac{1}{N}\sum_{i=1}^{N} l(h_\theta(x_i), y_i)$$

Discrete Wasserstein loss

- Assuming h_θ outputs a probability measure (or a discrete probability distribution), and $\mathbf{y}_i$ corresponds to the one-hot encoding of the label classes,

$$W_c(\alpha, \beta) = \sum_{i=1}^{N} L_{\mathbf{C}}(h_{\theta(x_i)}, \mathbf{y}_i)$$

  where $\mathbf{C}$ encodes the ground metric given by $c(x, y)$.
- In order to optimize the loss function, how do we compute gradients?
  - Gradients are easy to compute in the dual domain.

Dual problem formulation

1. Construct the Lagrangian:

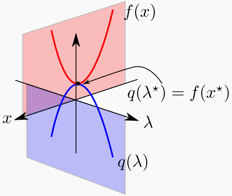
$$L(x, \lambda, \nu) = f(x) + \sum_i \lambda_i g_i(x) + \sum_j \nu_j h_j(x).$$

2. **Dual function**: the minimum of the Lagrangian over x:

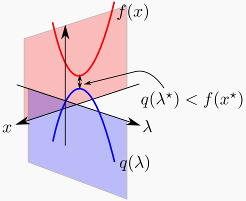
$$q(\lambda, \nu) = \min_x L(x, \lambda, \nu).$$

3. **Dual problem**: maximization of the dual function over $\lambda_i \geq 0$:

$$\begin{aligned} \max_{\lambda \in \mathbb{R}^m, \nu \mathbb{R}^p} \quad & q(\lambda, \nu) \\ \text{s.t.} \quad & \lambda_i \geq 0 \quad \forall i. \end{aligned} \tag{2}$$



strong duality



weak duality

Dual problem of the discrete Kantorovich problem

Dual of the discrete Kantorovich problem

Given $\mathbf{p} \in \mathbb{R}^n$, $\mathbf{q} \in \mathbb{R}^n$ and $\mathbf{C} \in \mathbb{R}^{n \times n}$, the dual of $L_{\mathbf{C}}(\mathbf{p}, \mathbf{q})$ has the following form:

$$\begin{aligned} \max_{\mathbf{r},\mathbf{s}} \quad & \mathbf{p}^T \mathbf{r} + \mathbf{q}^T \mathbf{s} \\ \text{s.t.} \quad & \mathbf{r}\mathbf{1}^T + \mathbf{1}^T \mathbf{s} \leq \mathbf{C} \end{aligned} \tag{3}$$

where $\mathbf{r} \in \mathbb{R}^n$, $\mathbf{s} \in \mathbb{R}^n$.

- Because the primal OT Kantorovich problem is a feasible LP for $\mathbf{p}$ and $\mathbf{q}$ probability distributions, the dual problem is also feasible and strong duality holds.
- The dual problem can play an important part in devising algorithms to solve the Kantorovich problem.
- Interpretation of prices of dual variables.

Dual problem of the discrete Kantorovich problem: Proof

- Semilagrangian of the primal problem:

$$J(\mathbf{F};\mathbf{r},\mathbf{s}) = \operatorname{tr}(\mathbf{C}\mathbf{F}^T) + \mathbf{r}^T(\mathbf{p} - \mathbf{F}\mathbf{1}) + \mathbf{s}^T(\mathbf{q} - \mathbf{F}^T\mathbf{1})$$

- Dual problem:

$$\max_{\mathbf{r},\mathbf{s}} \mathbf{r}^T\mathbf{p} + \mathbf{s}^T\mathbf{q} + \min_{\mathbf{F}\geq 0} \operatorname{tr}(\mathbf{C}\mathbf{F}^T) - \underbrace{\mathbf{r}^T\mathbf{F}\mathbf{1}}_{\operatorname{tr}(\mathbf{F}^T\mathbf{r}\mathbf{1}^T)} - \underbrace{\mathbf{s}^T\mathbf{F}^T\mathbf{1}}_{\mathbf{F}^T\mathbf{1}\mathbf{s}^T}$$

where $\mathbf{Q} = \mathbf{C} - \mathbf{r}\mathbf{1}^T - \mathbf{1}\mathbf{s}^T$

$$\min_{\mathbf{F}\geq 0} \operatorname{tr}(\mathbf{C}\mathbf{F}^T) - \underbrace{\mathbf{r}^T\mathbf{F}\mathbf{1}}_{\operatorname{tr}(\mathbf{F}^T\mathbf{r}\mathbf{1}^T)} - \underbrace{\mathbf{s}^T\mathbf{F}^T\mathbf{1}}_{\mathbf{F}^T\mathbf{1}\mathbf{s}^T} = \begin{cases} 0 & \text{if } \mathbf{Q} \geq 0 \\ -\infty & \text{otherwise} \end{cases}$$

- Giving

$$\begin{aligned} \max_{\mathbf{r},\mathbf{s}} \quad & \mathbf{r}^T\mathbf{p} + \mathbf{s}^T\mathbf{q} \\ \text{s.t.} \quad & \mathbf{r}\mathbf{1}^T + \mathbf{1}^T\mathbf{s} \leq \mathbf{C} \end{aligned}$$

Gradient of the Wasserstein Loss

- Back to the Wasserstein loss function: $\boxed{L_{\mathbf{C}}(h_{\theta(x_i)}, \mathbf{y}_i)}$.
- If we write it in dual form:

$$\begin{aligned} \max_{\mathbf{r},\mathbf{s}} \quad & \mathbf{r}^T h_{\theta(x_i)} + \mathbf{s}^T \mathbf{y}_i \\ \text{s.t.} \quad & \mathbf{r}\mathbf{1}^T + \mathbf{1}^T \mathbf{s} \leq \mathbf{C}. \end{aligned}$$

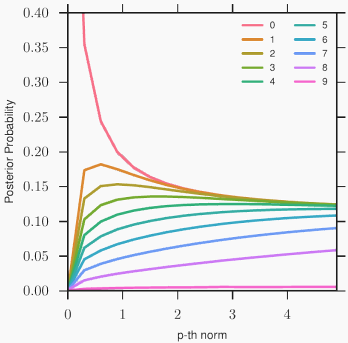
- We can take conditional subgradient w.r.t. $h_\theta(x)$:

$$\frac{d}{dh_\theta(x)} W_p(h_\theta(x), y) = \mathbf{r}$$

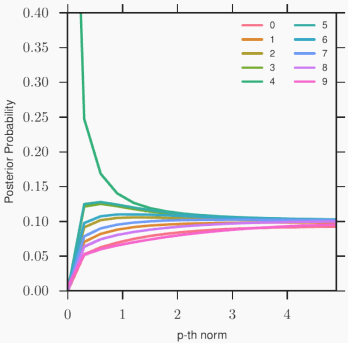
- Note that the Wasserstein loss is subdifferientiable.
- Computing the Wasserstein loss for N examples can be costly in high dimensions...
- Once we have the subgradient, we can backpropagate to update θ with SGD.

Effects of the ground metric I

- Authors compare discriminative power of W_p for different p norm values.



(a) Posterior prediction for images of digit 0.

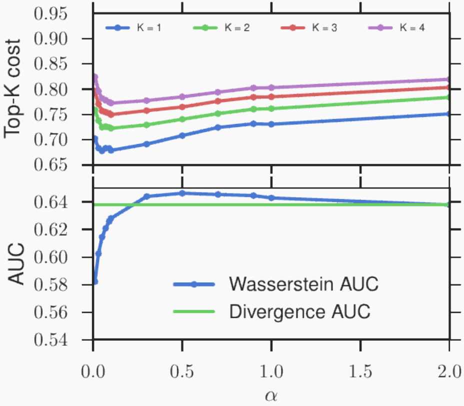


(b) Posterior prediction for images of digit 4.

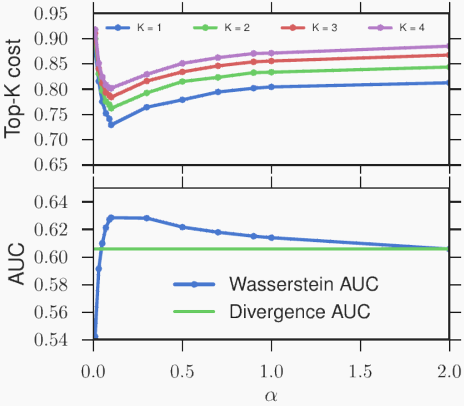
Effects of the ground metric II

- KL loss vs. Wasserstein loss on the Flickr database:

$$l(x_i, y_i) = W_p(h_\theta(x_i), y_i) + \alpha KL$$



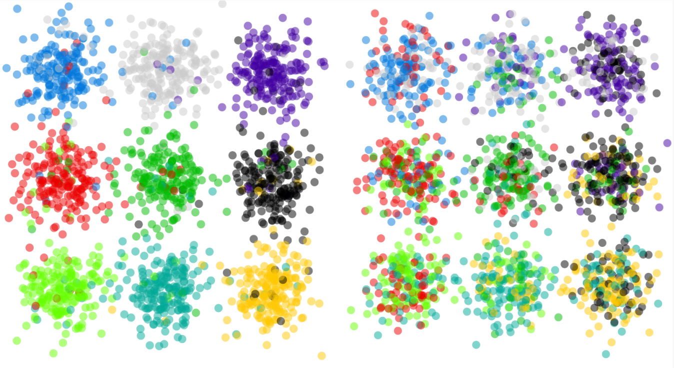
(a) Original Flickr tags dataset.



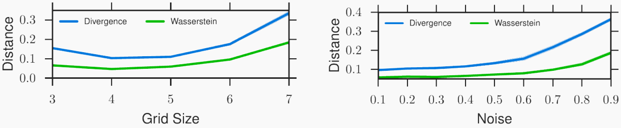
(b) Reduced-redundancy Flickr tags dataset.

Homework proposal

- Train a Wasserstein loss classifier on the plane with semantic classes.



(a) Noise level 0.1 (b) Noise level 0.5



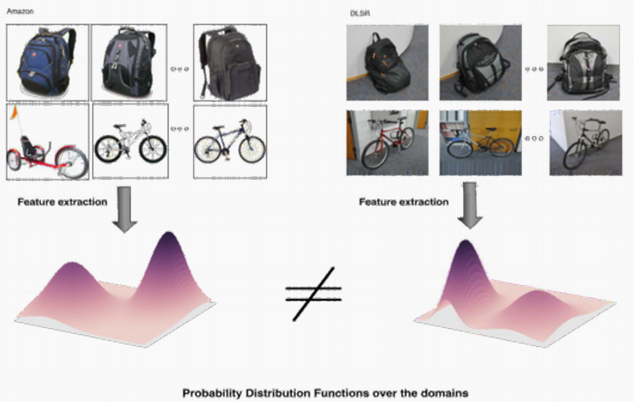
Thank you for listening!

- There are more things I wanted to talk about.

1. **Approximate methods** such as Sinkhorn, or smooth OT, to scale problem dimensions.
2. **Domain adaptation** transport a database of unlabelled data, to a domain where such labels exist, according to a Wasserstein transport plan.
3. **Ground metric learning** allows to learn the cost matrix from data, potentially improving performance compared to a p-Wasserstein loss as we have seen in examples.
4. Barycenter estimation: for clustering, or interpolation between histograms.
5. Transfer learning.
6. Unbalanced optimal transport.
7. Wasserstein discriminant analysis.
8. Etc.

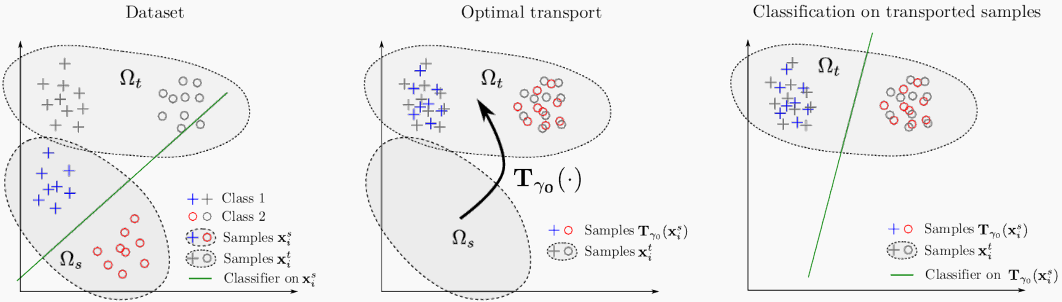
Application II: Domain adaptation

Problem intuition



- We consider unsupervised domain adaptation $\longrightarrow$ labels only in source domain.
- **Assumption**: data is processed to make the domains similar.
- Transformation follows a **least effort principle**.

Procedure



1. Estimate the marginals μ_s and μ_t from source and target sample distributions.
2. Find a transport map T from μ_s to μ_t.
3. Use T to transport labeled samples $\mathbf{x}_s$ and train a classifier from them.

Related work

- The approach defines a local transformation for each sample in the domain.
- It can be seen as a graph matching problem $\longrightarrow$ marginal distribution conservation.
- **Related work**:
  1. Projection methods: inner products, region transformation, extraction of common features.
  2. Unsupervised: common latent space representations; feature extraction is key.
  3. Gradual alignment of feature representation: kernel methods.

Problem description

- $\mathcal{K}$ set of possible labels; only available for $\mathcal{X}$.
- Source sample data: $((\mathbf{x}_i^s)_i^N, (y_i)_i^N)$.
- Target sample data: $((\mathbf{x}_i^s)_i^N)$.
- Joint probability distribution in source: $P_s(\mathbf{x}^s, y)$
- Marginal over x: μ_s.
- Joint probability distribution in target: $P_t(\mathbf{x}^t, y)$.
- Marginal over x: μ_t.

Assumptions of the transportation

- The domain drift is to an unknown, possibly nonlinear transformation of the linear space
$$T : \mathcal{X} \to \mathcal{Y}$$
- From probabilistic perspective, T transforms μ_s into μ_t, i.e.,
$$T\sharp\mu_s : \mathcal{M}_+^1 \to \mathcal{M}_+^1 = \mu_t$$
X_t are drawn from same pdf as $T\sharp\mu_s$.
- Transformation preserves conditional distribution, i.e.,
$$P_s(y|\mathbf{x}^s) = P_t(y|\mathbf{x}^t) \quad \Longleftrightarrow f_t(T(\mathbf{x}^s)) = f_s(\mathbf{x}^s)$$

Problem formulation

- Empirical distributions:

$$\mu_s = \sum_{i=1}^{N_s} p_i^s \delta_{x_i^s}, \qquad \mu_t = \sum_{i=1}^{N_t} p_i^t \delta_{x_i^t}$$

- Transport problem:

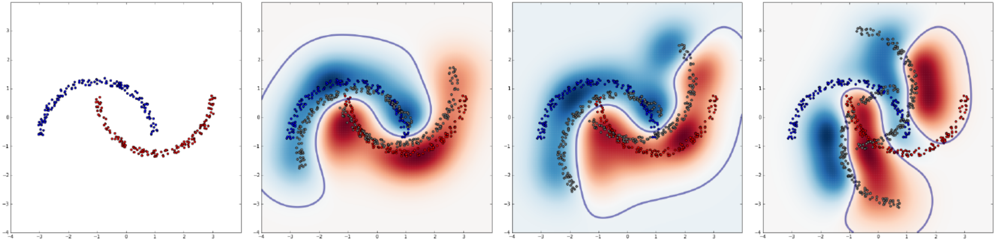
$$\mathbf{F} = \underset{\mathbf{F} \in U(\mu_s, \mu_t)}{\arg\min} \ \mathrm{tr}(\mathbf{FC})$$

  where $C_{ij} = \|\mathbf{x}_s - \mathbf{x}_t\|^2$.

- When $N_s = N_t = N$ and forall i, $p_i^s = p_i^t = 1/N$, $\mathbf{F}$ is simply a permutation matrix, which makes a correspondence of one to one from source to target domain.

Results

- Once we have the transport plan, we can bring features with labels to the target domain and train a classifier.
- Regularization can be induced to improve results using labels
- Results:



(a) source domain (b) rotation=20° (c) rotation=40° (d) rotation=90°

Thanks again

Questions?