ON THE COMPUTATION OF KANTOROVICH-WASSERSTEIN DISTANCES BETWEEN 2D-HISTOGRAMS BY UNCAPACITATED MINIMUM COST FLOWS

July 7, 2019

Georgios Michas (sdi1400109)
Maria - Despoina Siampou (sdi1600151)

National and Kapodistrian Universiry of Athens
Department of Informatics and Telecommunications

Contents

0.1 Introduction 2
0.2 Notations and Definitions 2
0.3 Flow on Reduce Graphs and Relaxations 5
0.4 Efficient Computation of W1 distance between 2D-Histograms 7

0.1 Introduction

To start with, recent works in Computer Vision and Machine Learning have shown the benefits of measuring Wasserstein distances of order one between histograms with N bins, by solving a classical **transportation problem** on (very large) complete bipartite graphs with N nodes and N^2 edges. The key idea of the paper is that we can transform the original **transportation problem** to an **uncapacitated min cost flow problem** on a reduced flow network of size $O(N)$. Bin distances (from bin centers) will be measured with $1-norm, 2-norm$ and $\infty-norm$.

0.2 Notations and Definitions

1. **Wasserstein/Kantorovich/Rubinstein Distance/EMD**
   The basic idea of the distance is intuitively, if we have a set of distributions and we view each one as a unit amount of "dirt" piled on **(metric space) M**, the metric is the minimum "cost" of turning one pile into the other, which is assumed to be the amount of dirt that needs to be moved times the mean distance it has to be moved. Because of this analogy, the metric is known in computer science as the **Earth Mover's Distance (EMD)**.
   Now we will define this distance in the discrete setting. Let $X = \{x_1, ..., x_n\}$ and $Y = \{y_1, ..., y_m\}$ be two discrete spaces. Given two probability vectors on X and Y, say $\mu = \{\mu(x_1), ..., \mu(x_n)\}$ and $\nu = \{\nu(y_1), ..., \nu(y_m)\}$, and a cost function such as $c : X \times Y \to \mathbb{R}_+$ the distance is :

   $$W_c(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \sum_{(x,y) \in X \times Y} c(x, y)\pi(x, y). \quad (1)$$

   with

   $$\sum_{y \in Y} \pi(x, y) = \mu(x) \quad and \quad \sum_{x \in X} \pi(x, y) = \nu(y) \quad \forall (x, y) \in X \times Y$$

   where $\Pi(\mu, \nu)$ is the set of all the probability measures on $X \times Y$ with marginals μ and ν (Such probability measures are sometimes called transportation plans or couplings for μ and ν.). if $X = Y$ and as a cost function we apply the distance on X then we have the **Wasserstein Distance of order one**.

2. **Linear Programming on Earth Movers Distance**
   The Kantorovich-Rubinshtein transport problem in the discrete setting can be seen

as a special case of the following Linear Programming problem, where we assume now that μ and ν are generic vectors of dimension n and m, with positive components,

$$\text{(P) minimize} \quad \sum_{x\in X}\sum_{y\in Y} c(x,y)\pi(x,y) \quad (2)$$

$$\text{subject to} \quad \begin{array}{ll} \sum_{y\in Y}\pi(x,y) \le \mu(x) & \forall x \in X \quad (3) \\ \sum_{x\in X}\pi(x,y) \ge \nu(y) & \forall y \in Y \quad (4) \\ \pi(x,y) \ge 0. \quad (5) & \end{array}$$

Now we will determine when the problem is balanced and when the problem is unbalanced with the following :

- if $\sum_x \mu(x) = \sum_y \nu(y)$, we have a **balanced transportation problem**.
- if $\sum_x \mu(x) \neq \sum_y \nu(y)$, we have an **unbalanced transportation problem**.

We can easy define the Duality of this Problem (P), remember that any feasible solution of a dual problem gives a valid lower bound to the primal problem but it isn't necessary to show it now, you can find it at page 5 in the original paper.
From problem (P) now we can define the **EMD** distance as a **linear programming problem** with the following changes:

- let $X, Y \subset \mathbf{R^d}$
- let i,j be the data clusters
- let x_i be the center of data cluster i and y_j be the center of data cluster j
- let $\mu(x_i)$ be the total points of cluster i and $\nu(y_j)$ be the total points of cluster j
- let $c(x_i, y_j)$ be some measure of dissimilarity between x_i and y_j

With this notations now we can solve the linear program (P).As a result we have the optimal transport π^*.
EMD for signatures $(x_i, \mu(x_i))$ and $(y_i, \nu(y_i))$ is :

$$EMD(\mu,\nu) = \frac{\sum_{x\in X}\sum_{y\in Y} c(x,y)\pi^*(x,y)}{\sum_{x\in X}\sum_{y\in Y}\pi^*(x,y)}.$$

As we see to calculate the EMD we have to solve a linear program.

3. **Uncapacitated minimum cost flow problem on a graph (network)**
The key idea is that we can transform a Linear Programming Problem to a minimum cost flow problem to solve it more efficient (also we can reduce the graph with some techniques which i will mention, this is what we will do in the end).
Let $G = (V, E)$ be a directed network graph with no-self loops, V is the Vertex Set and E is the Edge Set.
Define a cost function $c : E \to [0, +\infty)$, a function $b : V \to \mathbb{R}$ such that $\sum_{u \in V} b(u) = 0$ and a flow function (or b-flow) $f : E \to [0, +\infty)$ (can be seen as the flow (units) we transfer from vertex u to v), we denote $F(G, b)$ the class of all b-flows on G.
For $b(u)$ we have the following :

   - $b(u) > 0 \implies$ total flow $\uparrow$ (u is a supply node)
   - $b(u) < 0 \implies$ total flow $\downarrow$ (u is a demand node)
   - $b(u) = 0 \implies$ total flow maintained

The definition of this linear program is the following :

$$
\begin{aligned}
\text{(P2) minimize} \quad & \sum_{(u,v) \in E} c(u,v) f(u,v) \quad (6) \\
\text{subject to} \quad & \textstyle\sum_{u:(u,v) \in E} f(u,v) - \sum_{u:(u,v) \in E} f(v,u) = b(u) \quad \forall u \in V \quad (7) \\
& f(u,v) \geq 0. \quad (8)
\end{aligned}
$$

You can find the Duality of P2 problem at page 7 in the original paper.
Now we set :

$$F_{G,c}(b) := \min_{f \in F(G,b)} \sum_{(u,v) \in E} c(u,v) f(u,v).$$

with f be the optimal flow function.
We can easy see that if we have a graph G and a sub graph G_s (of G) :

$$F_{G_s,c}(b) \geq F_{G,c}(b). \quad (9)$$

4. **Wasserstein distance of order one as a minimum cost flow problem**
This part is the most important of all because we will transform the $W_c(\mu, \nu)$ to a minimum cost flow problem to apply network reduction.
For this purpose we will use a **bipartite graph**. A bipartite graph (or bigraph) is a graph whose vertices can be divided into two disjoint and independent sets U and

V such that every edge connects a vertex in U to one in V. Vertex sets U and V are usually called the parts of the graph.
So let's consider the following graph definition :

$$G_{X\to Y} = (X \cup Y, E_{X\to Y})$$

with $E_{X\to Y} = \{(x, y) : x \in X, y \in Y\}$. Also we define the following :

$$b(u) := \begin{cases} \mu(u) & \text{if } u \in X, \\ -\nu(u) & \text{if } u \in Y. \end{cases}$$

According to this definition we have the following :

- $u \in X \implies \mu(u) = b(u) = \sum_{u:(u,v)\in E} f(u,v) - \sum_{u:(u,v)\in E} f(v,u) = \sum_{u\in Y} f(u,v)$
- $u \in Y \implies -\nu(u) = b(u) = \sum_{u:(u,v)\in E} f(u,v) - \sum_{u:(u,v)\in E} f(v,u) = -\sum_{u\in X} f(u,v)$

So with the vector of the form :

$$b(u) = (\mu(x_1), ..., \mu(x_n) ..., -\nu(y_1), ..., -\nu(y_m))$$

we have the following equality :

$$F_{X\to Y,c}(b) = W_c(\mu, \nu). \quad (10)$$

As a result we can calculate the Wasserstein distance of order one by solving an uncapacitated minimum cost flow problem on the previous bipartite graph.

For a network $G = (V, E)$, with $n = |V|$ and $k = |E|$ we can compute $W_c(\mu, \nu)$ with time complexity of $O((n+m)^3 \log(n+m))$ as mentioned in the paper.

0.3 Flow on Reduce Graphs and Relaxations

1. **Introduction**
   As the paper proposed if $X = Y$ and c be a distance, we want with some way to reduce the main graph ($G_{X\to X}$). If we accomplish this then the computation of the W_c will be faster.

2. **Reduced Graph**
   For the Reduced Graph we have the following :

   - $b = \mu - \nu$.
   - $\mu(x) > 0$ and $\nu(x) > 0$, $\forall x \in X$.
   - The graph has set of nodes $V = X$, with cardinality equal to n while the first graph ($G_{X \to Y}$) had cardinality equal to $2n$.
   - For the **reduced graph** we choose for set of Edges, the set of all possible directed links on X, that is $G = K_n$, K_n being the complete (directed) graph on X (without self-loops).

   Although, we reduced the graph we still have a large number of edges, as mentioned in the paper the number is $\frac{n(n-1)}{2}$ but as they propose this will still be useful.

   As mentioned in the paper now we will define **an auxiliary bipartite graph** G_b as :

   - Set of Nodes : $V_b := S_b \cup D_b \subset X$, with $S_b := \{x : b(x) > 0\}$ and $D_b := \{x : b(x) < 0\}$
   - Set of Edges : $E_b := \{(x, y) \in S_b \times D_b\}$
   - The Minimum Cost Flow Problem for this Graph :

   $$F_{G_b,c}(b) := \min_{f \in F(G_b,b)} \sum_{x \in S_b, y \in D_b} c(x, y) f(x, y).$$

   - The G_b has cardinality $E_b = \alpha^+ \alpha^- n^2$, where α^+- is the fractions of vertexes x with $b(x) \lesseqgtr 0$. As a result $\alpha^+ \alpha^- < 1$, this means the cardinality of G_b is smaller from $G_{X \to X}$.

   The main inequality for the previous graphs and the minimum cost flow of them is the following :

   $$F_{G_b,c}(\mu - \nu) \geq W_c(\mu, \nu) \geq F_{K_n,c}(\mu - \nu) \quad (11)$$

   We can compute EMD for two normalize measues μ and ν with $c = d$ (where d is a ground distance) by solving the minimum cost flow $F_{G_b,d}(\mu - \nu)$.
   As mentioned before we know that the graph G_b is smaller from the graph $G_{X \to X}$. The algorithm which has been proposed to solve the minimum cost flow problem $F_{G_b,d}(b)$ has same order complexity ($0(n^3 log(n))$ as the algorithm used to solve the minimum cost flow problem $F_{X \to Y,d}(b)$.

3. **Relaxation and Error Bounds**
   Our main goal is for certain costs (c) to reduce our main graph to a simpler one. This means the new graph will be a relaxation of the main graph, this will lead to a faster computation of the distances we want.
   For a graph $G = (V, E) \subset K_n$ such as $F_{G,c}(b) \geq F_{K_n,c}(b)$ we have :

$$0 \leq \frac{F_{G,c}(b) - F_{K_n,c}(b)}{F_{G,c}(b)} \leq 1$$

   We define the relative error of approximation between the original minimum and its relaxation as :

$$\mathscr{E}_G(b) = \frac{F_{G,c}(b) - F_{K_n,c}(b)}{F_{G,c}(b)}$$

   So as a result if one can find a subgraph $G \subset K_n$ for which $\mathscr{E}_G(b) = 0$, one can reduce the original problem of computing $F_{K_n,c}(b)$ to a simpler minimum cost flow problem $F_{G,c}(b)$. Even if this isn't always possible we can approximate the $F_{K_n,c}(b)$ via a subgraph G.
   Moreover, in the paper they introduce an upper bound for the $\mathscr{E}_G(b)$. This upper bound is based only on the geometry (G,c) and not on the specific b.

0.4 Efficient Computation of W1 distance between 2D-Histograms

Histograms can be seen as discrete measures on a finite set of points in R^d. Considering the case of d=2, we try to represent histograms of N × N equally spaced bins. We assume without loss of generality that :

$$X = LN := \{i = (i_1, i_2) : \ i_1 = 0, 1, .., N-1, \ i_2 = 0, 1, .., N-1\}$$

where $(i_1,\ i_2)$ can be considered as the center of a bin.

In this case, $L_N = N^2$ the complete directed graph $K = K_{N^2}$ has $N^2(N^2 - 1) = N^4$ edges.

Given i = $(i_1, i_2) \in$ Z × Z and $j = (j_1,\ j_2) \in$ Z × Z, we also **define** $i + j = (i_1 + j_1, i_2 + j_2)$.
If G = (V, E) is a graph and i,j $\in V$, we denote the directed edge connecting i to j by (i,j) $\in E$. Let $N_0 := N \setminus \{0\}$ and $Z_0 := Z \setminus \{0\}$. For L $\in N_0$, we define the following sets of

vertices:

$$V_0 \ := \ \{(1,0),(0,1),(-1,0),(0,-1)\}$$

$$V_L \ := \ V_0 \ \cup \ \{(i_1,i_2) \in Z_0 \times Z_0 : |\, i_1 \,|\leq L, \ |\, i_2 \,|\leq L\},$$

where $|\, i_1 \,|, |\, i_2 \,|$ implies that any point of coordinates $i \in V_L$ is visible from the origin (0,0).

1. **L_1 ground distance**

   Considering $d_1(i,j) = |\, i_1 - j_1 \,| + |\, i_2 - j_2 \,|$ as ground distance, we just need to choose a subgraph K:

   $$G_0 \ := \ (L_N, E_0)$$

   $$\text{where } E_0 \ := \ \{(i, i+j) : \ i \in L_N, \ j \in V_0, \ i+j \in L_N\}, \ 2(N-1)N = O(N^2).$$

   In that case, it can be easily assumed that $\Gamma_{G_0,d_1} = 0$, but also, $\Gamma_{G_0,d_1}(\mathrm{b}) = \Gamma_{K,d_1}(\mathrm{b})$.

   According to these, it can be also assumed that, $W_{d_1}(\mu,\nu) = F_{G_0,d_1}(\mu,\nu)$ for every couple of probability measures μ and ν on L_N, meaning that the computation of $W_{d_1}(\mu, \ \nu)$ between two normalized histograms μ and ν can be done without the cause of any error, by solving a minimum cost flow problem.

2. **L_∞ ground distance**

   Following the same procedure as above, we consider $d_\infty(i,j) = \max(|\, i_1 - j_1 \,| + |\, i_2 - j_2 \,|)$ as ground distance and we a subgraph K:

   $$G_1 \ := \ (L_N, E_1)$$

   $$\text{where } E_1 \ := \ \{(i, i+j) : \ i \in L_N, \ j \in V_1, \ i+j \in L_N\}, \ 2(N-1)N = O(N^2).$$

   Following the same assumptions with previous section, we realize that also the computation of $W_{d_\infty}(\mu, \ \nu)$ between two normalized histograms μ and ν can be done without the cause of any error, by solving a minimum cost flow problem.

3. **L_2 ground distance**

   Considering $d_2(i, j) = \sqrt{| i_1 - j_1 |^2 + | i_2 - j_2 |^2}$ as ground distance, we need to choose a suitable subgraph K:

$$G_L := (L_N, E_L)$$

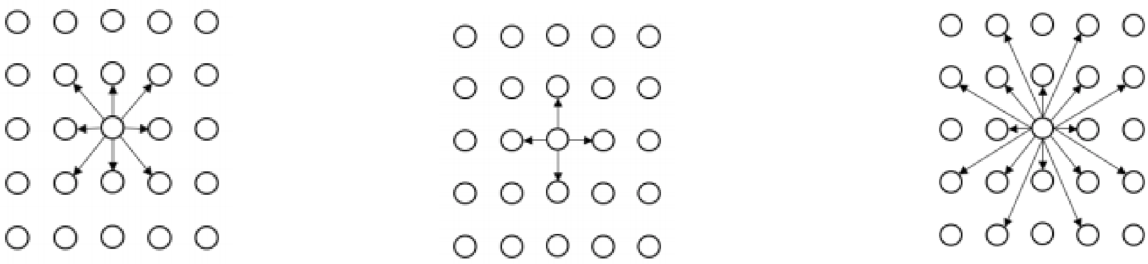
$$\text{where } E_L := \{(i, i+j) : i \in L_N, \ j \in V_L, \ i+j \in L_N\}, \ \forall \ 1 \le L \le N-1 \text{ and}$$

$$V_L := \{(i_1, i_2) \in V_L : i_1 \ge i_2 \ge 0\},$$

where V_L is the graph inducing G_L.

   The only directions considered here, are those characterized by angles in ranging between 0 and $\frac{\pi}{4}$. All other directions can be obtained from V_L by rotations of k$\frac{\pi}{4}$, $\forall \ 1 \le k \le 7$.

   As a result, in the specific case of G_{N-1} subgraph the approximation error is zero, meaning that the computation of $W_{d_2}(\mu, \nu)$ between two normalized histograms μ and ν can be done without the cause of any error, by solving a minimum cost flow problem which is strictly contained in the complete graph K. However, in this case, the number of required edges is of order N^4. In other cases there is a high possibility for an error to occur due to the approximation of K with G_L, $\forall \ L < N-1$.



(a) L_1 **(b)** L_∞ **(c)** L_2

Figure 1: Subgraphs of K corresponding to the previous sections