Approximation of Wasserstein distance with Transshipment
Nicolas Papadakis

To cite this version:
Nicolas Papadakis. Approximation of Wasserstein distance with Transshipment. 2019. hal-01997598

HAL Id: hal-01997598
https://hal.archives-ouvertes.fr/hal-01997598
Submitted on 22 Jul 2019
Approximation of Wasserstein distance with Transshipment

Nicolas Papadakis

1CNRS, Univ. Bordeaux, IMB, F-33400 Talence, France

Abstract

An algorithm for approximating the $p$-Wasserstein distance between histograms defined on unstructured discrete grids is presented. It is based on the computation of a barycenter constrained to be supported on a low dimensional subspace, which corresponds to a transshipment problem. A multi-scale strategy is also considered. The method provides sparse transport matrices and can be applied to large and non structured data.

1 Introduction

The computation of optimal transportation between two discrete normalized measures $\mu_x$ and $\mu_y$ defined from $x \in \Omega \subset \mathbb{R}^d$ to $[0; 1]$ remains a challenging problem when an accurate discretization of the domain $\Omega$ is considered. Optimal transportation requires to define a ground distance between points $x, y \in \mathbb{R}^d$ of the domain $\Omega$. This ground metric is then used to measure how much it costs to move $\mu_x(x)$ to $\mu_y(y)$. Ground distances $||x - y||^p$ are here considered for $p \geq 1$, which leads to the $p$-Wasserstein distance between $\mu_x$ and $\mu_y$ (see [45, 36, 33] for more detailed introductions). Such distances give robust metrics in retrieval applications for $1 \leq p < 2$ [35, 31, 32, 21]. The underlying sparse transport matrix is also a useful tool for interpolation and transfer purposes [34, 17, 42].

Computing Wasserstein distances The computation of Wasserstein distances is only explicit for $d = 1$. A standard approach to estimate Wasserstein distances when $d > 1$ consists in pre-computing a cost matrix $||x - y||^p$ for $x, y \in \Omega$ and then estimating the whole transport matrix which dimension grows quadratically with $n = |\Omega|$. Linear programming or transportation simplex [25, 19] can be applied to estimate $p$-Wasserstein distance for $p \geq 1$ but they are limited in practice to low dimensional problems, i.e. small values of $n$, for complexity and storage issues. The implementation provided in [11] nevertheless allows to tackle problems of interesting dimensions, i.e. $\Omega$ discretized with more than $n = 10^4$ points. By exploiting the sparsity of the transport map, multi-scale strategy [28] or grid refinement [38] can deal with larger problems by solving iteratively sparse low dimensional problems with linear programming. More efficient algorithms can be considered in the specific case of $1$-Wasserstein distances [23].

Approximation of Wasserstein distance For large scale problems, approximated sparse transport matrices and Wasserstein distances can be obtained by considering successive one dimensional problems with the so-called sliced Wasserstein distance [34, 10]. The entropic regularization of the transport map proposed in [14] is another relevant way to deal with problems of high dimension. Given a regularization parameter $\gamma > 0$, it approximates the true Wasserstein distance (that corresponds in this setting to $\gamma = 0$) with the well-known Sinkhorn algorithm. The estimated transport matrices are nevertheless dense and should be truncated with care for interpolation purposes. When data are discretized on an uniform grid, the estimation of the distance can be obtained through iterative convolutions [42] with a kernel $K_p = \exp(-||x - y||^p/\gamma)$. This leads to very fast algorithm
for the $p-$Wasserstein distance as the Kernel is separable when considering the $L^p$ norm $||.||_p$, and only convolutions and storage of $d$ one-dimensional kernels are necessary. For small values of $\gamma$, numerical instabilities nevertheless arise and dedicated attention must be given to the implementation by considering for instance decaying values of $\gamma$, stabilization in log domain or multi-resolution approaches [39]. The numerical convergence is also reduced with low values of $\gamma$, and overrelaxation [37, 13] or greedy coordinate descents [3] approaches have been proposed to tackle this issue. When the kernel $K_p$ is not separable, it must be carefully truncated to zero to store sparse matrices. If not considering more complex and adaptive truncations [39], it limits the possible amplitude of the transport, which may be annoying in case of large displacements. Other regularizations can then be of interest [16, 9, 11], namely when it is suitable to recover sparse transport matrices.

However, as mentioned before, this kind of techniques only leads to efficient implementations for densities discretized on structured grids. Moreover, if large displacements are involved in the data, the Wasserstein distance can not be accurately approximated, since numerical instabilities arises with $||x - y||^p/\gamma$ when $\gamma$ goes to 0.

**Wasserstein barycenters and Transshipment** The $p-$Wasserstein barycenter, as introduced in [1], can be used to approximate the $p-$Wasserstein distance between measures involving a large number $n$ of dirac masses. As proposed in [46] for clustering problems, the discrete barycenter between 2 discrete measures can be parameterized with a weighted sum of $\kappa << n$ diracs to obtain a low dimensional problem that corresponds to the transshipment problem of $n$ resources with $\kappa$ intermediate locations. The sum of the distances between each data and the barycenter then gives an approximation of the effective distance, as it has been proposed for $1-$Wasserstein distances [5]. Statistical properties of such a method have been later studied in [18, 30], where it has been underlined the robustness of this low rank regularization of the transport matrix to data outliers. These ideas have been extended to discrete approximation of barycenter between continuous measures in [12].

**Content** The use of low dimensional barycenters is the point of view adopted in this note to propose fast approximation of Wasserstein distances involving sparse transport matrices. As in [46], at each iteration of the presented algorithm, the $\kappa$ locations of the barycenter are updated and a linear program of dimension $2\kappa n$ is solved. Compared to the dimension $n^2$ for the classical Wasserstein distance, the overall complexity of the algorithm becomes attractive for high dimensional data. In [46] and [18, 30], this low dimensional barycenter problem is respectively solved with the Alternating Direction Method of Multipliers or Sinkhorn iterations [15]. The barycenter problem between two discrete densities is here seen as a transshipment problem and solved up to numerical accuracy with an efficient network simplex graph algorithm based on the work of [11]. Then $\kappa$ transportation subproblems are solved in parallel to recover a sparse transport matrix. Following [39, 23, 24], this leads to the design of a multi-scale barycenter scheme, to iteratively refine the transport matrix and the associated approximation of the Wasserstein distance.

The theoretical computer science community has recently provided improved bounds for graph based algorithms solving approximate transportation problems up to $\epsilon$ additive or $1+\epsilon$ multiplicative errors. This namely includes the optimal transportation problem [8], its entropic regularization [2] or the transshipment problem [6]. As for the sliced method [34], the presented approach does not have such theoretical guarantees but very good performances are observed in practice. More precisely, approximate $p-$Wasserstein distances between cloud points of $10^3$ elements are obtained in a few minutes without involving prohibitive memory storage issues. The proposed empirical algorithm can also be directly applied to non structured data. Thanks to the multi-scale refinement approach, a sparse transport map is provided which can be of interest for interpolation purposes.
Outline Section 2 details how to approximate the Wasserstein distance when computing the barycenter between two discrete measures. The strategy of [46] is recalled in the general context of $p$–Wasserstein distance. A multi-scale algorithm for recovering sparse transport matrices is finally presented. The performances of the algorithm are discussed in section 3 through extensive experiments realized on the benchmark [40]. Numerical results show that this whole empirical process is efficient, namely when one of the two input data is spatially regular.

2 Approximate Wasserstein distance from barycenter estimation

Let $\mu_x$ and $\mu_y$ be two discrete measures defined on $\Omega \subset \mathbb{R}^d$, $d \geq 1$: $\mu_x = \sum_{i=1}^{m} w_i^x \delta_{x_i}$ and $\mu_y = \sum_{j=1}^{n} w_j^y \delta_{y_j}$. These measures are supported at positions $\{x_i\}_{i=1}^{m}$ and $\{y_j\}_{j=1}^{n}$, $x_i, y_j \in \Omega \subset \mathbb{R}^d$.

They have normalized positive weights vectors $w^x \in S_m$ and $w^y \in S_n$, where $S_n$ is the simplex of size $n$ defined as $S_n = \{w \in \mathbb{R}^n_+, \text{ s.t } \sum_{i=1}^{n} w_i = 1\}$. For $p \geq 1$, let $c^{xy} \in \mathbb{R}^{m \times n}$ be the ground cost matrix over $\Omega$ defined as $c_{ij}^{xy} = \|x_i - y_j\|^p$, which corresponds to the power $p$ of the distance related to a given norm on $\mathbb{R}^d$. Then the $p$–Wasserstein distance between discrete measures $\mu_x$ and $\mu_y$ is

$$W_p(\mu_x, \mu_y) = \min_{\gamma \in \mathcal{P}(w^x, w^y)} \gamma, c^{xy} := \sum_{ij} \gamma_{ij} c_{ij}^{xy} ,$$

with the set of admissible transport matrices

$$\mathcal{P}(w^x, w^y) = \{ \gamma \in \mathbb{R}_+^{m \times n}, \text{ s.t } \gamma 1_n = w^x, \gamma^\top 1_m = w^y\},$$

and where $1_n$ the vector full of ones in $\mathbb{R}^n$. This problem can be efficiently solved with linear programming. It can also be formulated through a directed graph containing $m + n$ nodes and $mn$ vertices. The final transport matrix $\gamma$ is very sparse in practice (at most $m + n - 1$ non null entries) but the involved complexity and memory storage scale with the product of data dimensions $nm$. For latter purpose, Algorithm 1 details the function estimating the distance $W_p(\mu_x, \mu_y)$.

### Algorithm 1

1: procedure $W_p(x, w^x, y, w^y, p)$
2: Set $c_{ij}^{xy} = \|x_i - y_j\|^p$
3: Solve problem (1) under the constraints (2) to get $\gamma$
4: Set $W = (\gamma, c^{xy})$
5: return $W, \gamma$

2.1 Interpolation and barycenters

Let $\gamma^{xy}$ be an optimal transport matrix solution of (1) and $\mu_t$ be the following interpolation between measures $\mu_x$ and $\mu_y$ for $t \in [0; 1]$

$$\mu_t = \sum_{ij} \gamma_{ij}^{xy} \delta_{x_i + t(y_j - x_i)} := \sum_{k} w_k^t \delta_{z_k}.$$  

This interpolation is the discrete analogue [33] to the geodesic between $\mu_x$ and $\mu_y$ defined by the McCann’s interpolation [26]. Discrete versions of some results in [36] can now be expressed.

**Proposition 1** For $p \geq 1$, $\gamma^{xy}$ a solution of (1), $\mu_t$ defined in (3) and $t \in [0; 1]$, the following relations hold:

$$W_p(\mu_x, \mu_t) = t W_p(\mu_x, \mu_y) \quad W_p(\mu_y, \mu_t) = (1-t) W_p(\mu_x, \mu_y),$$

so that

$$W_p(\mu_x, \mu_y) = W_p(\mu_x, \mu_t) + W_p(\mu_y, \mu_t).$$
Proof. Observing that
\[
W_p^p(\mu_x, \mu_t) = \min_{\gamma \in \mathcal{P}(\mu_x, \mu_t)} \sum_{i} \gamma_{ik} \|x_i - z_k\|^p \leq \sum_{ij} \gamma_{ij}^p \|t(x_i - y_j)\|^p,
\]
leads to the upper bound \(W_p^p(\mu_x, \mu_t) \leq t W_p^p(\mu_x, \mu_y), \forall t \in [0; 1]\). It can be shown in the same way that \(W_p^p(\mu_y, \mu_t) \leq (1 - t) W_p^p(\mu_x, \mu_y)\). Since \(W_p\) is a distance, the triangle inequality \(W_p^p(\mu_x, \mu_y) \leq W_p^p(\mu_x, \mu_t) + W_p^p(\mu_t, \mu_y)\) involves that the previous relations are in fact equalities. □

Following [11], it can be shown that the mid interpolation \(\mu_{1/2}\) is solution of the p-Wasserstein barycenter problem between \(\mu_x\) and \(\mu_y\) with weights \((1/2, 1/2)\).

**Proposition 2** For \(p \geq 1\) and the interpolation \(\mu_{1/2}\) defined in [3], it holds that
\[
\frac{1}{2} W^p_p(\mu_x, \mu_{1/2}) + \frac{1}{2} W^p_p(\mu_y, \mu_{1/2}) = \left(\frac{1}{2} W^p_p(\mu_x, \mu_y)\right)^p
\]
and \(\mu_{1/2}\) is a solution of the \(p\)-Wasserstein barycenter problem:
\[
\mu_{1/2} \in \arg \min_{\mu} \frac{1}{2} W^p_p(\mu_x, \mu) + \frac{1}{2} W^p_p(\mu_y, \mu).
\]

Proof. From [4], it can first be noticed that:
\[
\frac{1}{2} \left(W^p_p(\mu_x, \mu_{1/2}) + W^p_p(\mu_y, \mu_{1/2})\right) = \frac{1}{2} \left(2 \left(\frac{1}{2} W^p_p(\mu_x, \mu_y)\right)^p\right) = \left(\frac{1}{2} W^p_p(\mu_x, \mu_y)\right)^p.
\]

Since \(W_p\) is a distance and the function \(|.|^p\), is convex for \(p \geq 1\), it can next be observed that \(\forall \mu:\)
\[
\left(\frac{1}{2} W^p_p(\mu_x, \mu_y)\right)^p \leq \left(\frac{1}{2} (W^p_p(\mu_x, \mu) + W^p_p(\mu_y, \mu))\right)^p \leq \frac{1}{2} (W^p_p(\mu_x, \mu) + W^p_p(\mu_y, \mu)).
\]
Combining relations [8] and [9] implies that \(\mu_{1/2}\) is a solution of the barycenter problem [7]. □

Existence (and uniqueness for \(p > 1\)) of Wasserstein barycenters have been deeply studied in [11][20]. As stated in the following proposition, the \(p\)-Wasserstein distance can be obtained for any \(p \geq 1\) through the resolution of the barycenter problem [7].

**Proposition 3** Let \(\bar{\mu}\) be a solution of the \(p\)-Wasserstein barycenter problem [7], then
\[
W_p^p(\mu_x, \mu_y) = W_p^p(\mu_x, \bar{\mu}) + W_p^p(\mu_y, \bar{\mu})
\]
and \(W_p^p(\mu_x, \bar{\mu}) = W_p^p(\mu_y, \bar{\mu})\) for \(p > 1\).

Proof. Assuming, without loss of generality, that \(W_p^p(\mu_y, \bar{\mu}) = \alpha W_p^p(\mu_x, \bar{\mu})\), with \(\alpha \in [0; 1]\), it holds from Proposition 2 that \(\bar{\mu}\) is a solution of [7] iff \(2^{p-1}(1 + \alpha^p) W_p^p(\mu_x, \bar{\mu}) = W_p^p(\mu_x, \mu_y)\). Assuming that \(W_p^p(\mu_x, \bar{\mu}) + W_p^p(\mu_y, \bar{\mu}) = (1 + \alpha) W_p^p(\mu_x, \bar{\mu}) > W_p^p(\mu_x, \mu_y)\), it involves that
\[
2^{p-1}(1 + \alpha^p) < (1 + \alpha)^p < (1 + \alpha) 2^{p-1} \quad \alpha^p < \alpha,
\]
which is impossible \(\forall \alpha \in [0; 1]\) and \(p \geq 1\). Thus relation [10] holds by contradiction and \(2^{p-1}(1 + \alpha^p) = (1 + \alpha)^p\). With the very same arguments than in [11], it leads to \(\alpha = 1\) for \(p > 1\). □
2.2 Barycenter computation

A discrete barycenter $\mu_{1/2}$ between $\mu_x$ and $\mu_y$ can be obtained by solving (7) with the distance $\|\cdot\|_2$. As done in [15] with an additional entropic regularization, the mid barycenter $\mu_{1/2}$ between $\mu_x$ and $\mu_y$ can be constrained to be supported on a set of $\kappa$ dirac masses, i.e. $\mu_{1/2} = \sum_{k=1}^{\kappa} w_k \delta_{z_k}$, with positions $z_k \in \mathbb{R}^d$ and weights $w_z \in S_\kappa$. The barycenter problem can be rewritten as:

$$\min_{\{z_k\}_{k=1}^{\kappa} \in \Omega^\kappa \atop w_z \in S_\kappa} W_p^p(\mu_x, \mu_z) + W_p^p(\mu_y, \mu_z) = \min_{\{z_k\}_{k=1}^{\kappa} \in \Omega^\kappa} (\gamma^x, \gamma^x) + (\gamma^y, \gamma^y),$$

with the admissible set of matrices

$$\tilde{P}(w^x, w^y) = \{\gamma^x \in \mathbb{R}_+^{m \times \kappa}, \gamma^y \in \mathbb{R}_+^{n \times \kappa}, \text{ s.t. } \gamma^x 1_\kappa = w^x, \gamma^y 1_\kappa = w^y, (\gamma^x)^\top 1_m - (\gamma^y)^\top 1_n = 0_\kappa\}$$

and the cost matrices $c_{ik}^x = \|x_i - z_k\|^p$ and $c_{jk}^y = \|y_j - z_k\|^p$. Notice that the weight vector of the barycenter $w_z \in \mathbb{R}^\kappa$ is implicitly included in the set of constraints: $(\gamma^x)^\top 1_m = (\gamma^y)^\top 1_n = w_z$. The positions $z_k$ here act as intermediate locations where the mass has to transit from $x$ to $y$. Following [46] and as illustrated in Figure 1 the idea behind this modeling is to consider a limited number of transhipment locations $\kappa$ to speed up the computation.

The problem [12] is separately convex with respect its variables $\gamma^x$, $\gamma^y$ and $z_k$. For $p > 1$, the coupling terms are differentiable and alternate minimization over transport matrices $(\gamma^x, \gamma^y)$ and dirac positions $z_k$ converges to a saddle point. Following [10], the Wasserstein distance $W_p(\mu_x, \mu_y)$ can be approximated with $\tilde{W}_p(\mu_x, \mu_y)$ obtained from transport matrices $\gamma^x$ and $\gamma^y$ solutions of the problem [12] as

$$W_p(\mu_x, \mu_y) \leq \tilde{W}_p(\mu_x, \mu_y) = W_p(\mu_x, \mu_{1/2}) + W_p(\mu_y, \mu_{1/2}) = (\gamma^x, c^x)^{1/p} + (\gamma^y, c^y)^{1/p}.$$

The alternate optimization steps for solving [12] are now detailed and the process is summed up in Algorithm 2.

**Remark 1** The norm $\|\cdot\|_2$ is usually taken as reference for the $p-$Wasserstein distance. From the equivalence of norms in finite dimensions and as can be done with entropic regularization to make the problem more tractable numerically, the $L^p$ norm $\|x_i - y_j\|^p = \sum_{s=1}^{d} (x_s^i - y_s^j)^p$ is here considered for computing the $p-$Wasserstein distance.
2.2.1 Update of positions

To update positions $z_k \in \mathbb{R}^d$, the problem (12) is solved for fixed transport matrices $\gamma^x$ and $\gamma^y$. For $p \geq 1$, this leads to $d$ convex problems, that can be solved in parallel for each space dimension $s = 1 \ldots d$. Dimension indexes $s$ are thus omitted in the following and the problem writes

$$
\min_{\{z_k\}_{k=1}^\kappa} \sum_{i=1}^M \sum_{k=1}^\kappa \gamma^x_{ik} |x_i - z_k|^p + \sum_{j=1}^N \sum_{k=1}^\kappa \gamma^y_{jk} |y_j - z_k|^p,
$$

(13)

Different strategies are considered according to $p$. Some locations $z_k$ may become useless if $\sum_i \gamma^x_{ik} = \sum_j \gamma^y_{jk} = 0$. In this case the corresponding $z_k$ are removed and $\kappa$ is decreased. When considering large scale problems and few transshipment locations $\kappa$, this almost never happens.

**Case W2.** For $p = 2$, there exists an explicit update formula of the dirac positions to find the unique minimizer of (13) with respect to $z$:

$$
z_k = \frac{(\gamma^x)^\top x + (\gamma^y)^\top y)_k}{(\gamma^x)^\top 1_m + (\gamma^y)^\top 1_n)_k},
$$

(14)

This step acts like the cluster position update in a $\kappa$-mean algorithm. It realizes for each $z_k$ a weighted mean of the positions $x_i$’s and $y_j$’s according to $\gamma^x_{ik}$ and $\gamma^y_{jk}$.

**Case Wp, $p > 2$.** For $p > 2$ the problem (13) admits a unique minimizer and is twice differentiable. Newton’s method can be considered to approximate the solution:

$$
z_k^{t+1} = z_k^t - \frac{\left( (\gamma^x \otimes |z^w|^{|p-2|} \otimes c^{xz})^\top 1_m + (\gamma^y \otimes |z^w|^{|p-2|} \otimes c^{yz})^\top 1_n \right)_k}{(p-1) \left( (\gamma^x \otimes |z^w|^{|p-2|})^\top 1_m + (\gamma^y \otimes |z^w|^{|p-2|})^\top 1_n \right)_k},
$$

where $c_{ik}^{xz} = z_k^t - x_i$ and $c_{jk}^{yz} = z_k^t - y_j$, while $\otimes$ denotes the elementwise product between matrices and the power $p - 2$ is also element-wise.

**Case W1.** When $p = 1$, a global optimum of (13) can be obtained by taking each $z_k$ as a weighted median of the positions $x_i$ and $y_j$ with respect to the weights $\gamma^x_{ik}$ and $\gamma^y_{jk}$. This operation mainly requires to sort the value of the $x_i$’s and $y_j$’s along each dimension. Notice that alternate minimization on problem (12) may not converge when $p = 1$.

**Case Wp, $p \in [1;2]$.** The problem does not admit a second derivative. An iterative scheme is then considered by decomposing $|z_k - x_i|^p = |z_k - x_i|^{p-1}|z_k - x_i|$ and solving successive weighted median problems between the positions $x_i$ and $y_j$ with the weights $\gamma^x_{ik}|z_k^t - x_i|^{p-1}$ and $\gamma^y_{jk}|z_k^t - y_j|^{p-1}$.
2.2.2 Update of transport matrices through Transshipment

For fixed \( z_k \), problem \([12]\) can be solved with classic linear programming optimization tools. The interesting point is that this barycenter problem is a transshipment problem with \( \kappa \) intermediate locations. It can therefore be formulated in terms of a directed graph with \( m + \kappa + n \) vertices (i.e. \( x_i, z_k \) and \( y_j \)) and \((m\kappa + \kappa n)\) edges \( e_{ik} \) and \( c_{kj} \) that correspond to the transport matrices \( \gamma^x_{ik} \) and \( \gamma^y_{jk} \). The following cost function is then minimized

\[
(\gamma^x, (\gamma^y)^\top) \in \underset{e=(\{e_{ik}\}, \{c_{kj}\})}{\text{argmin}} \sum_{i=1}^{m} \sum_{k=1}^{\kappa} e_{ik} e_{ik}^x + \sum_{k=1}^{\kappa} \sum_{j=1}^{n} c_{kj} c_{kj}^{yz},
\]

under the set of constraint \( \tilde{P}(w^x, w^y) \) that translates into:

\[
\begin{align*}
    e_{ik}, c_{kj} & \geq 0 & i = 1 \cdots m, & k = 1 \cdots \kappa, & j = 1 \cdots n \\
    \sum_{k=1}^{\kappa} e_{ik} & = w^x_i & i = 1 \cdots m \\
    \sum_{i=1}^{m} e_{ik} & = 0 & k = 1 \cdots \kappa \\
    \sum_{j=1}^{n} e_{kj} & = w^y_j & j = 1 \cdots n
\end{align*}
\]

This problem can be efficiently solved with the network simplex algorithm \([29]\). An extension of the original non sparse implementation proposed in \([11]\) is here considered.

2.2.3 Discussion

It is well known \([1, 4]\) that if \( \mu_x \) and \( \mu_y \) are respectively supported by \( m \) and \( n \) dirac masses, then their exists a barycenter supported by up to \( m + \kappa + n \) dirac masses. Hence, the approximation of the barycenter from a set of \( \kappa < \min(m,n) \) dirac masses seems to be a bad choice at first sight. Notice however that, as pointed out in \([18]\), a barycenter supported by a low dimensional space allows the approximate distance to be more robust to data outliers. Next, contrary to entropic regularization of OT, the obtained approximate transport map is here sparse, which allows efficient storage and can be directly used for interpolation purposes without any complex post-processing like sharpening.

Finally, it is worth noting that the approximation depends on the regularity of the data. As underlined in \([11]\), when computing the barycenter of a set of densities \( \mu_i \) in the continuous case, if one of the input data \( \mu_i \) is absolutely continuous with respect to the Lebesgue measure, so does the barycenter. Such observation gives an interesting insight of the experiments presented in this paper: when at least one of the two data \( \mu_x \) or \( \mu_y \) is smooth, then the obtained approximate Wasserstein distance \( \tilde{W}_p \) can be very close to the true one \( W_p \) for small values of \( \kappa \). Similar behaviour can be observed with semi-discrete optimal transport models where voronoi cells act like barycenters \([27]\). This point is illustrated in Figure 2 with the comparison of relative errors between true and approximate distances obtained with increasing values of \( \kappa \) for different scenarios involving “random” or “smooth” 2D data of dimensions \( m = n = 10^4 \). When at least one “regular” data is involved, then a small relative error \( \frac{\tilde{W}_p - W_p}{W_p} < 10^{-3} \) (i.e. 0.1%) is observed for \( \kappa/n = 0.015 \).

![Figure 2](image-url)
2.3 Transport refinement and Multi-scale approach

As illustrated in Figure 2, the barycenter approach is not sufficient to produce an accurate approximation of the $p$–Wasserstein distance between any data. First notice that the approximation $\hat{W}_p$ of the Wasserstein distance can be easily improved using the following result that directly considers the transport matrix between $\mu_x$ and $\mu_y$.

**Proposition 4** Let $\mu_z = \sum_{k=1}^n w_k^z \delta_{z_k}$ and $\gamma^x$ (resp. $\gamma^y$) be an optimal transport matrix from $\mu_x$ (resp. $\mu_y$) to $\mu_z$. Let also $\hat{\gamma}$ be the transport matrix between $\mu_x$ and $\mu_y$ defined as $\hat{\gamma} = \gamma^x D(\gamma^y)^\top$, with the rescaling $D^{-1} = \text{diag}(w^z)$ given by the weights $w_k^z = \sum_i \gamma^x_{ik} = \sum_j \gamma^y_{jk}$. Then the following relation holds

$$
\left( \sum_{ij} \hat{\gamma}_{ij} |x_i - y_j|^p \right)^{1/p} \leq \hat{W}_p(\mu_x, \mu_z) + W_p(\mu_y, \mu_z).
$$

\( (17) \)

**Proof.** The matrix $\hat{\gamma}_{ij} = \sum_k \gamma^x_{ik} \gamma^y_{jk} / w_k^z$ is an admissible transport matrix between $\mu_x$ and $\mu_y$, since $\sum_j \hat{\gamma}_{ij} = \sum_k \gamma^x_{ik} = w_i^x$ and $\sum_i \hat{\gamma}_{ij} = \sum_k \gamma^y_{jk} = w_j^y$. Following the proof of Minkowski’s inequality:

$$
\begin{align*}
\sum_{ij} \hat{\gamma}_{ij} |x_i - y_j|^p &\leq \sum_{ijk} \gamma^x_{ik} \gamma^y_{jk} / w_k^z |x_i - y_j|^p \leq \sum_{ij} \gamma^x_{ik} \gamma^y_{jk} / w_k^z |x_i - y_j|^{p-1} |x_i - z_k| + |y_j - z_k|) \\
&\leq \left( \sum_{ijk} \gamma^x_{ik} \gamma^y_{jk} / w_k^z |x_i - y_j|^p \right)^{(p-1)/p} \left( \left( \sum_{ijk} \gamma^x_{ik} \gamma^y_{jk} / w_k^z |x_i - z_k|^p \right)^{1/p} + \left( \sum_{ijk} \gamma^x_{ik} \gamma^y_{jk} / w_k^z |y_j - z_k|^p \right)^{1/p} \right) \\
&\leq \left( \sum_{ij} \hat{\gamma}_{ij} |x_i - y_j|^p \right)^{(p-1)/p} \left( \left( \sum_{ik} \gamma^x_{ik} |x_i - z_k|^p \right)^{1/p} + \left( \sum_{jk} \gamma^y_{jk} |y_j - z_k|^p \right)^{1/p} \right),
\end{align*}
$$

then gives \( (17) \). Notice that the proof of the triangle inequality \[13\] is here also obtained, since $W_p(\mu_x, \mu_y) \leq (\sum_{ij} \hat{\gamma}_{ij} |x_i - y_j|^p)^{1/p}$. □

From this proposition, the approximation $\hat{W}_p(\mu_x, \mu_y)$ corresponding to a barycenter $\mu_{1/2} = \sum_{k=1}^n w_k^z \delta_{z_k}$ solution of \[12\] with transport matrices $\gamma^x$ and $\gamma^y$ is defined as

$$
\hat{W}_p(\mu_x, \mu_y) \leq \hat{W}_p(\mu_x, \mu_y) = \langle \gamma^x \text{diag}(w^z)^{-1}\gamma^y \rangle^\top \in \hat{W}_p(\mu_x, \mu_y). \tag{18}
$$

For matching or interpolation purposes, $\hat{\gamma} = \gamma^x \text{diag}(w^z)^{-1}\gamma^y$ gives a sparse approximation of the optimal transport matrix. However, as illustrated in Figure 3, such approach maps all locations $x_i$ and $y_j$ that transit by $z_k$, resulting in a poor block transport matrix for small values of $\kappa$.

![Illustration of the block transport matrices $\hat{\gamma}$ estimated from data $\mu_x$ and $\mu_y$ through transshipment for increasing values of $\kappa$.](image)

Figure 3: Illustration of the block transport matrices $\hat{\gamma}$ estimated from data $\mu_x$ and $\mu_y$ through transshipment for increasing values of $\kappa$. 

8
In order to get a sparser and more accurate approximation of the optimal transport matrix and associated distance, it is necessary to “untie” the links between locations passing through $z_k$. To do so, a solution is to consider $\kappa$ optimal transportation sub-problems, by refining the transport of the mass transshipped through $z_k$. The barycenter approach then acts as a clustering and for each intermediate location $z_k$, the $p-$Wasserstein distance between the following partial discrete densities is computed:

$$\mu_k^x = \sum_{i=1}^{m_k} \gamma_{ik}^x \delta_{x_i}, \quad \mu_k^y = \sum_{j=1}^{n_k} \gamma_{jk}^y \delta_{y_j},$$

(19)

where the number of active dimensions are expected to be reduced: $m_k = \#\{\gamma_{ik}^x > 0\} < m$ and $n_k = \#\{\gamma_{jk}^y > 0\} < n$. The multi-scale approach is then performed as follows. The Wasserstein distance $W_p(\mu_k^x, \mu_k^y)$ is estimated exactly with network simplex [11] if the sum of $m_k$ and $n_k$ is small enough. Otherwise the barycenter approach is recursively applied to the subproblem. From numerical experiments, the threshold $n_k + m_k < 2000$ has been chosen to reach the best compromise between numerical accuracy and computational cost. The whole process is illustrated in Figure 4 and detailed in Algorithm 3.

![Figure 4: Illustration of transportation obtained from refinement of transshipment with $\kappa = 4$ and $\kappa = 8$ intermediate locations. edges indicate there is a mass transport (i.e. $\gamma_{ij} > 0$) between locations $x_i$ (in blue) and $y_j$ (in red). Optimal Transportation is here recovered when refining the transshipment solution obtained with $\kappa = 8$.](image-url)
The relative errors are very low. In all other cases, exact and approximate distances have been computed between all (i.e. \(32 \times 32 \times 32 = 16384\)) possible pairs of images. This has been done for different values of \(\kappa\) and a threshold of \(T = 2000\) in Algorithm 3. For each experiment, the mean and median relative errors between approximate and exact methods are computed and presented in Table 1.

### Table 1: Mean and median relative errors between approximate and true EMD on the Benchmark

<table>
<thead>
<tr>
<th>(\kappa) = 4</th>
<th>(\kappa) = 16</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n = 32 \times 32 = 1024)</td>
<td>Mean: 2.70% Median: 2.18%</td>
</tr>
<tr>
<td>(n = 64 \times 64 = 4096)</td>
<td>Mean: 3.56% Median: 2.61%</td>
</tr>
<tr>
<td>(n = 128 \times 128 = 16384)</td>
<td>Mean: 3.49% Median: 2.34%</td>
</tr>
</tbody>
</table>

A more detailed presentation is then given in Figure 5, with the mean relative errors over intra and inter classes experiments in the case of \(128 \times 128\) images. As can be observed, significant errors are obtained when unstructured random data (classes 4 and 10) are involved. In all other cases, the relative errors are very low.

---

1 The code is available at [https://www.math.u-bordeaux.fr/~npapadak/GOTMI/codes.php](https://www.math.u-bordeaux.fr/~npapadak/GOTMI/codes.php)
4 Conclusion

This paper presents an empirical method for approximating Wasserstein distances. It is based on existing concepts used in parallel works \cite{13, 5, 30}. The contribution is to provide an efficient multi-scale implementation able to deal with unstructured point clouds while providing sparse transport matrices. Numerical experiments demonstrate the accuracy of the computed approximate distance, while the involved computational cost are improved with respect to the literature. As a perspective, several intermediary transshipment levels could be considered, in relation to branched transport \cite{7} It would also be of interest to add constraints encouraging an homogeneous repartition of the number of points transiting by each location \(z_k\) of the barycenter, or at least a more uniform distribution of the barycenter weights \(w_k\). The \(\kappa\) multi-scale sub-problems that can be solved in parallel would have similar dimensions, and theoretical guarantees on the overall running time could be given.
Figure 6: Comparison of running times for computing $2-W$ Wasserstein distances for different values of $n$: C++ network simplex [11], its multi-threaded extension and the proposed multi-scale approximation (matlab C++ mex), that is the only one being able to handle high scale problems.

Acknowledgements This study has been carried out with financial support from the French State, managed by the French National Research Agency (ANR GOTMI) (ANR-16-CE33-0010-01). The project has also received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 777826.

References


