Approximating the Earth Mover's Distance between sets of points and line segments[*]

Marc van Kreveld[1], Frank Staals[1], Amir Vaxman[1], and Jordi L. Vermeulen[1]

1 **Department of Information and Computing Sciences, Utrecht University**
`{m.j.vankreveld;f.staals;a.vaxman;j.l.vermeulen}@uu.nl`

Abstract

We show that a $(1+\varepsilon)$-approximation algorithm exists for the Earth Mover's Distance between a set of n points and set of n line segments with equal total weight. Our algorithm runs in $O\left(\frac{n^6}{\varepsilon^2}\log^2\left(\frac{1}{\varepsilon}\right)\log^2\left(\frac{n^2}{\varepsilon}\log\frac{1}{\varepsilon}\right)\right)$ time.

1 Introduction

The Earth Mover's Distance (EMD) is a metric that is widely used in fields such as image retrieval [13], shape matching [5, 8, 16] and mesh reconstruction [3]. It models two sets A and B as distributions of mass, and takes their distance $\mathcal{D}(A,B)$ to be the minimum cost of transforming one distribution into the other, where cost is measured by the amount of mass moved multiplied by the distance over which it is moved. More formally,

$$\mathcal{D}(A,B) = \inf_{\mu\in M}\int_A\int_B d(a,b)\cdot\mu(a,b)\;da\,db \qquad (1)$$

where M is the set of all mappings of mass between A and B. In the case where A and B are sets of (weighted) points, we can rewrite this as

$$\mathcal{D}(A,B) = \min_{\mu\in M}\sum_{a\in A}\sum_{b\in B} d(a,b)\cdot\mu(a,b) \qquad (2)$$

For unweighted point sets, the solution can be obtained by solving an assignment problem; for weighted point sets, this is an instance of a min cost max flow problem.

In this work, we consider the case where A is a set of weighted points and B is a set of line segments in $\mathbb{R}^2$. We provide a polynomial-time algorithm that gives a $(1+\varepsilon)$-approximation to the Earth Mover's Distance between A and B, and also gives an assignment of mass that realises this cost. To our knowledge, this is the first combinatorial algorithm with a provable approximation ratio for the Earth Mover's Distance when the objects are continuous rather than discrete points.

2 Related work

The general problem of optimally moving a distribution of mass was first described by Monge in 1781 [11], and was reformulated by Kantorovich in 1942 [6]. It is known as the Earth Mover's Distance due to the analogy of moving piles of dirt around; it is also known as the 1-Wasserstein distance, and is a special case of the more general optimal transport problem.

[*] This research was supported by the Netherlands Organisation for Scientific Research under project number 612.001.651.

35th European Workshop on Computational Geometry, Utrecht, The Netherlands, March 18–20, 2019.
This is an extended abstract of a presentation given at EuroCG'19. It has been made public for the benefit of the community and should be considered a preprint rather than a formally reviewed paper. Thus, this work is expected to appear eventually in more final form at a conference with formal proceedings and/or in a journal.

For a full treatment of the problem's history and connections to other areas of mathematics, the reader is referred to Villani's book [17].

The Earth Mover's Distance has been studied in many geometric contexts. Cabello et al. [1] give a $(2 + \varepsilon)$-approximation to minimising the EMD between two point sets under rigid transformations. For continuous distributions, rather than discrete point sets, many numerical algorithms are known (see e.g. De Goes et al. [2], Lavenant et al. [7], Mérigot [9, 10] and Solomon et al. [15]). For the case where one set contains weighted points and the other is a bounded set $C \subset \mathbb{R}^d$, Geiß et al. [4] give a geometric proof that there exists an additively weighted Voronoi diagram such that transporting mass from each point p to the part of C contained in its Voronoi cell is optimal. The weights of this Voronoi diagram can be determined numerically.

De Goes et al. [3] discuss a problem similar to our own in the context of the reconstruction and simplification of 2D shapes. Given a set of points, they want to reconstruct a simplicial complex of a given number of vertices that closely represents the shape of the point set. They start with computing the Delaunay triangulation of the point set, then iteratively collapse the edge that minimises the increase in the EMD between the point set and the triangulation. They use a variant of the EMD in which the cost is proportional to the square of the distance (2-Wasserstein distance). This allows them to calculate this variant of the EMD between a given set of points and a given edge of the triangulation exactly, as the squared distance can be decomposed into a normal and a tangential component. However, they determine the assignment of points to edges heuristically. In this work, we show how to obtain a $(1 + \varepsilon)$-approximation to the true optimal solution.

3 Approximating the Earth Mover's Distance

We are given a set of points $P = \{p_1, \ldots, p_n\}$ with weights $\{w_1, \ldots, w_n\}$ and a set of segments $S = \{s_1, \ldots, s_n\}$, with lengths $\{l_1, \ldots, l_n\}$. We assume the mass associated with a segment is equal to its length, and that this mass is distributed uniformly over each segment. Given that $\sum w_i = \sum l_j = n$, we want to compute a "transport plan" of mass from P to S that minimises the cost according to the Earth Mover's Distance. We define for each pair $(p_i, s_j) \in P \times S$ a function $\mu_{i,j}(t)$, with $t \in [0, l_j]$, that describes the density of mass being moved from p_i to the point $s_j(t)$. All these functions together describe the function μ used in the definition of $\mathcal{D}(A, B)$. Such a set of functions needs to satisfy the following conditions to be a valid transport plan:

$$0 \leq \mu_{i,j}(t) \leq 1 \tag{3}$$

$$\forall i : \sum_{j=1}^{n} \int_t \mu_{i,j}(t)\, dt = w_i \tag{4}$$

$$\forall j, t : \sum_{i=1}^{n} \mu_{i,j}(t) = 1 \tag{5}$$

We can then define the cost of a given transport plan as

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \int_0^{l_j} \mu_{i,j}(t) \cdot d(p_i, s_j(t))\, dt \tag{6}$$

where $s_j(t)$ is the point on s_j associated with value t and $d(\cdot, \cdot)$ is any metric. Our problem is now to find a transport plan with minimal cost.

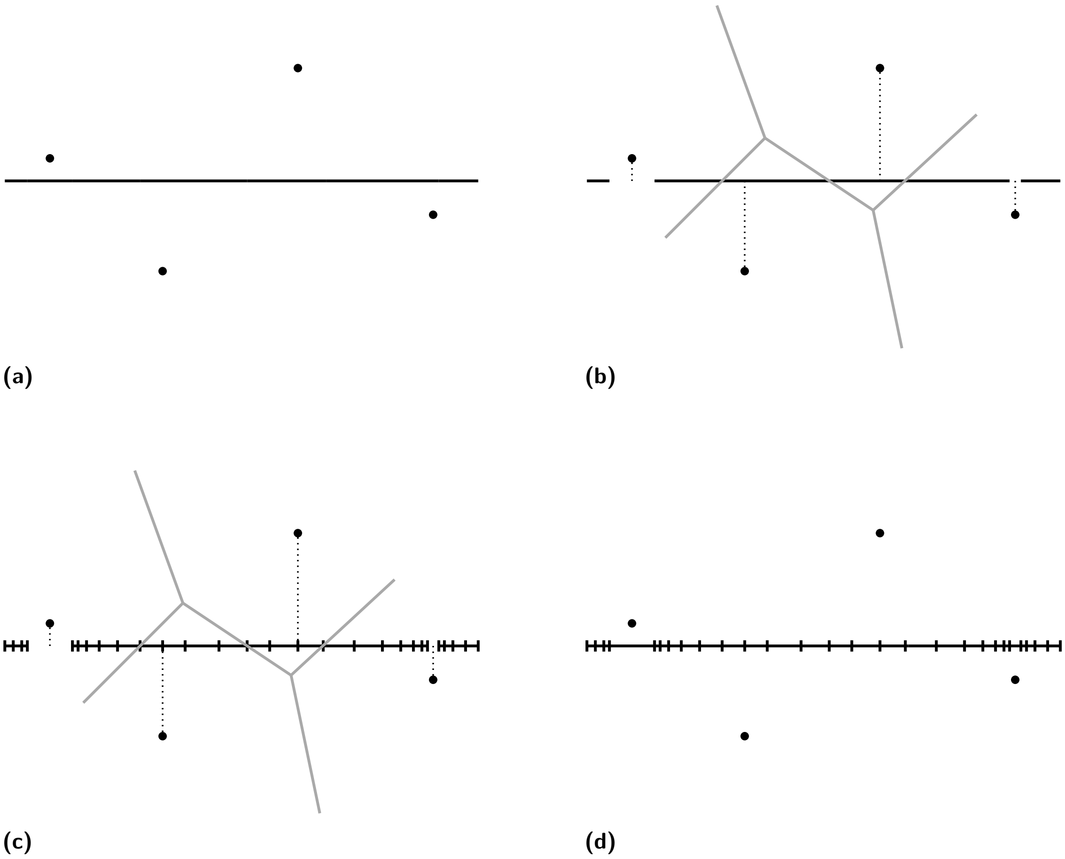

Figure 1 Our construction of Q for a single line segment under the Euclidean metric. (a) shows the input. (b) shows the Voronoi diagram of the points and the parts of the segment with distance at least δ/n. (c) shows the generated subsegments, and (d) all of Q, with the parts with distance less than δ/n added back in.

We now describe a polynomial-time algorithm that finds a transport plan with a cost that is at most $1+\varepsilon$ times the cost of the optimal transport plan. Our strategy is as follows: we subdivide each segment such that for each subsegment s' the ratio of the distance to the closest and furthest point on s' for any $p_i \in P$ is at most $1+\delta$. We then solve a min cost max flow problem on a bipartite graph between P and the subdivided segments, where the cost of any edge is equal to the shortest distance between a point and a subsegment. Finally, we use the solution to this flow problem to build a discrete transport plan. For an appropriate choice of δ, this gives a $(1+\varepsilon)$-approximation.

We begin by subdividing our segments as follows. First, we remove all parts of segments that lie within distance δ/n of any point in P (we will consider these parts separately later). Call the remaining set of segments S'; we subdivide S' by performing the following procedure: for each point in $p \in P$, we consider the part of S' that is within its Voronoi cell. Call this part S'_p. Now consider the closest point to p of any segment $s \in S'_p$, call their distance d. We create a subsegment s', starting at the closest point of s to p, with length $d \cdot (1+\delta)$ (or the length of s, if that is smaller). We remove this subsegment s' from S'_p, and iterate until S' is empty. Note that this way, the subsegments increase in size in both directions from the closest point to p. Call the set of all s' and all parts of the segments that lie within distance δ/n Q; this is our subdivision of S.

▶ **Lemma 1.** *Q has $O\left(\frac{n^2}{\delta}\log\frac{1}{\delta}\right)$ subsegments.*

Proof. Consider any segment $s_j \in S$. As the subsegments are created based on the closest point, we define two variables r_i and ℓ_i that denote the length of the part of s_j contained in p_i's Voronoi cell on either side of the perpendicular line from p_i to the supporting line of s_j. We also have at most one subsegment per point for the part that is within distance δ/n. The number of subsegments generated by p_i on s_j can then be expressed as $g(r_i,\ell_i) \leq \lceil\log_{1+\delta}(\frac{r_i}{\delta/n})\rceil + \lceil\log_{1+\delta}(\frac{\ell_i}{\delta/n})\rceil + 1 \leq \log_{1+\delta}(\frac{r_i}{\delta/n}) + \log_{1+\delta}(\frac{\ell_i}{\delta/n}) + 3$. Here we are counting the number of times we need to multiply the starting distance of δ/n by $1+\delta$ in order to reach length r_i or ℓ_i.

We are interested in the worst-case number of subsegments, so we want to find the maximum value of $\sum g(r_i,\ell_i)$ subject to the constraint that $\sum(r_i+\ell_i) \leq l_j$. As $\sum g(r_i,\ell_i)$ is a sum of logarithms, this is the same as maximising the product of all $g(r_i,\ell_i)$, which is achieved when all r_i and ℓ_i are equal (i.e. all are $l_j/2n$). By the same argument, the worst number of subsegments generated on all of S is upper bounded by making all l_j equal (i.e. all are 1). This gives us an upper bound on the total number of subsegments of

$$\sum_{i=1}^{n}\sum_{j=1}^{n} 2\cdot\log_{1+\delta}\left(\frac{1/2n}{\delta/n}\right) + 6 = 2n^2\cdot\frac{\ln\left(\frac{1}{2\delta}\right)}{\ln(1+\delta)} + 6n^2 = O\left(\frac{n^2}{\delta}\ \log\ \frac{1}{\delta}\right) \tag{7}$$

Note that we use the fact that $\ln(1+\delta) \approx \delta$ for small values of δ. We get that our number of subsegments is $O\left(\frac{n^2}{\delta}\log\frac{1}{\delta}\right)$ in the worst case. ◀

We now define a complete bipartite graph $G = (P \cup Q, E)$, with edges between each point-subsegment pair. The cost of each edge will simply be the shortest distance between the point and segment it connects. A solution to a flow problem in G can be transformed into a transport plan by assigning a piece of subsegment to a point with length equal to the amount of flow along the corresponding edge. We will show that the EMD between P and S is approximated by the cost of any transport plan derived from a min cost max flow in G.

First note the following general lower bound on the cost of an optimal solution:

▶ **Lemma 2.** *The Earth Mover's Distance between P and Q is bounded from below by the cost $\|\mathcal{W}\|$ of a min cost max flow $\mathcal{W}$ in G.*

Proof. Consider any transport plan that minimises the Earth Mover's Distance; call the cost associated with this plan OPT. If instead of spreading the mass equally over the whole segment, we move all the mass to the closest point on the segment, we obtain a plan with cost $\text{OPT}^* \leq \text{OPT}$. Such a plan is a solution to a maximum flow problem in G, as it moved all available mass. It follows that the cost $\|\mathcal{W}\|$ of a minimum cost maximum flow $\mathcal{W}$ in G satisfies $\|\mathcal{W}\| \leq \text{OPT}$. ◀

We also note the following lower bound on the value of $\|\mathcal{W}\|$:

▶ **Lemma 3.** $\|\mathcal{W}\| \geq \delta - 2\delta^2$.

Proof. For each point-segment pair, the part of the segment that has distance at most δ/n has length at most $2\delta/n$. The total length over all point-segment pairs is then $2\delta n$. This leaves $n - 2\delta n$ length with distance of at least δ/n, which gives a minimum cost of $(n-2\delta n)\cdot\delta/n = \delta - 2\delta^2$. Due to our construction of Q, we know that no subsegment crosses the distance boundary of δ/n. It follows that $\delta - 2\delta^2 \leq \|\mathcal{W}\|$. ◀

Using the lower bound from Lemma 2 and the way we constructed Q, we can derive a lower and upper bound on the solution obtained by the flow problem.

▶ **Lemma 4.** *For any transport plan $\mathcal{T}$ derived from $\mathcal{W}$ we have that its cost $\|\mathcal{T}\| \leq (1+\delta)\|\mathcal{W}\| + 2\delta^2$.*

Proof. We can upper bound $\|\mathcal{T}\|$ by measuring all distances to the furthest point in each subsegment. We constructed Q such that the ratio of the closest and furthest distance between any point-subsegment pair was $1+\delta$ when the closest distance was at least δ/n. We can therefore bound all parts of $\mathcal{T}$ where the distance is at least δ/n by $(1+\delta)\|\mathcal{W}\|$. The total mass being moved over a distance at most δ/n in $\mathcal{T}$ is at most δn, giving a cost of $2\delta^2$. The total cost when measuring to the furthest point is therefore $(1+\delta)\|\mathcal{W}\| + 2\delta^2$. ◀

▶ **Corollary 5.** $\|\mathcal{W}\| \leq \mathrm{OPT} \leq (1+\delta)\|\mathcal{W}\| + 2\delta^2$.

Putting this all together, we can show that $\|\mathcal{T}\|$ approximates OPT.

▶ **Theorem 6.** *The cost of any transport plan $\mathcal{T}$ derived from $\mathcal{W}$ is a $(1+5\delta)$-approximation to the Earth Mover's Distance between P and S for $0 < \delta \leq \frac{1}{4}$.*

Proof. The ratio between the upper and lower bound on $\|\mathcal{T}\|$ is

$$\frac{(1+\delta)\|\mathcal{W}\| + 2\delta^2}{\|\mathcal{W}\|}$$

This ratio is the largest for small values of $\|\mathcal{W}\|$, so we plug in the lower bound from Lemma 3:

$$\begin{aligned} & \frac{(1+\delta)\|\mathcal{W}\| + 2\delta^2}{\|\mathcal{W}\|} \\ \leq\ & \frac{(1+\delta)(\delta-2\delta^2)+2\delta^2}{\delta-2\delta^2} \\ =\ & \frac{1+\delta-2\delta^2}{1-2\delta} \\ =\ & 1+\frac{3\delta-2\delta^2}{1-2\delta} \\ =\ & 1+\delta+\frac{2\delta}{1-2\delta} \\ \leq\ & 1+5\delta \qquad \text{(assuming } \delta \leq \tfrac{1}{4}\text{)} \end{aligned}$$

As $\|\mathcal{W}\|$ is also a lower bound for OPT, and $\mathcal{T}$ can obviously not have lower cost than the optimal transport plan, this gives a $(1+5\delta)$-approximation. ◀

Setting $\delta = \varepsilon/5$ gives a $(1+\varepsilon)$-approximation.

3.1 Analysis

We can calculate $\mathcal{W}$ in $O(|E|\log|V|((|E|+|V|)\log|V|))$ time using Orlin's algorithm for minimum cost maximum flows [12]. In our case, $|V| = O\left(\frac{n^2}{\varepsilon}\log\frac{1}{\varepsilon}\right)$ and $|E| = O\left(\frac{n^3}{\varepsilon}\log\frac{1}{\varepsilon}\right)$; as $|V| \in O(|E|)$, we can simplify the running time to $O(|E|^2\log^2|V|)$. This gives us a total running time of $O\left(\frac{n^6}{\varepsilon^2}\log^2\left(\frac{1}{\varepsilon}\right)\log^2\left(\frac{n^2}{\varepsilon}\log\frac{1}{\varepsilon}\right)\right)$.

This time can be improved when the lengths of the segments in S are divisible by δ^2/n, by making all subsegments have the same length. When this is the case, $\mathcal{W}$ becomes a minimum cost matching rather than a minimum cost maximum flow. We can then use the algorithm by Sharathkumar and Agarwal to find a $(1+\delta)$-approximate bipartite matching in $O(|V| \text{ poly}(\log|V|, \frac{1}{\delta}))$ time [14]. We pay for this approximate rather than optimal matching by an extra 2δ in our approximation factor, giving a $(1+7\delta)$-approximation. Using this subdivision, the number of subsegments is n^2/δ^2, giving a total running time of $O\left(\frac{n^2}{\varepsilon^2} \text{ poly}\left(\log\frac{n^2}{\varepsilon^2}, \frac{1}{\varepsilon}\right)\right) = O\left(n^2 \text{ poly}\left(\log\frac{n}{\varepsilon}, \frac{1}{\varepsilon}\right)\right)$.

References

1 S. Cabello, P. Giannopoulos, C. Knauer, and G. Rote. Matching point sets with respect to the Earth Mover's Distance. *Computational Geometry*, 39(2):118–133, 2008.

2 F. de Goes, K. Breeden, V. Ostromoukhov, and M. Desbrun. Blue noise through optimal transport. *ACM Transactions on Graphics*, 31(6):171, 2012.

3 F. de Goes, D. Cohen-Steiner, P. Alliez, and M. Desbrun. An optimal transport approach to robust reconstruction and simplification of 2D shapes. *Computer Graphics Forum*, 30(5):1593–1602, 2011.

4 D. Geiß, R. Klein, R. Penninger, and G. Rote. Optimally solving a transportation problem using voronoi diagrams. *Computational Geometry*, 46(8):1009 – 1016, 2013.

5 K. Grauman and T. Darrell. Fast contour matching using approximate Earth Mover's Distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 220–227, 2004.

6 L. V. Kantorovich. On the translocation of masses. *Doklady Akademii Nauk*, 37:199–201, 1942.

7 H. Lavenant, S. Claici, E. Chien, and J. Solomon. Dynamical optimal transport on discrete surfaces. In *SIGGRAPH Asia 2018 Technical Papers*, pages 250:1–250:16, 2018.

8 F. Mémoli. Spectral Gromov-Wasserstein distances for shape matching. In *Proceedings of the 12th IEEE International Conference on Computer Vision Workshops*, pages 256–263, 2009.

9 Q. Mérigot. A multiscale approach to optimal transport. *Computer Graphics Forum*, 30(5):1583–1592, 2011.

10 Q. Mérigot, J. Meyron, and B. Thibert. An algorithm for optimal transport between a simplex soup and a point cloud. *SIAM Journal on Imaging Sciences*, 11(2):1363–1389, 2018.

11 G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

12 J. B. Orlin. A faster strongly polynomial minimum cost flow algorithm. *Operations Research*, 41(2):338–350, 1993.

13 Y. Rubner, C. Tomasi, and L. J. Guibas. The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

14 R. Sharathkumar and P. K. Agarwal. A near-linear time ε-approximation algorithm for geometric bipartite matching. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing*, pages 385–394, 2012.

15 J. Solomon, F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics*, 34(4):66, 2015.

16 Z. Su, Y. Wang, R. Shi, W. Zeng, J. Sun, F. Luo, and X. Gu. Optimal mass transport for shape matching and comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11):2246–2259, 2015.

17 C. Villani. *Optimal Transport: Old and New.* Springer Verlag, Heidelberg, 2008.