A Near Linear Time Constant Factor Approximation for Euclidean Bichromatic Matching (Cost)*

Piotr Indyk
MIT

Abstract

We give an $N \log^{O(1)} N$-time randomized $O(1)$-approximation algorithm for computing the cost of minimum bichromatic matching between two planar point-sets of size N.

1 Introduction

Consider two multisets A, B of points in $\mathbb{R}^2$, $|A| = |B| = N$. We define $EMD(A, B)$ to be the minimum cost of a perfect matching with edges between A and B, i.e., $EMD(A, B) = \min_{\pi: A \to B} \sum_{a \in A} \|a - \pi(a)\|_1$,[1] where π ranges over all one-to-one mappings. We are interested in efficient algorithms for computing $EMD(A, B)$.

The problem is of significant importance in applied areas, e.g., in computer vision [RTG00, IT03]. For general (i.e., non-planar) distances, it can be solved in time $O(N^3)$, using the "Hungarian" method [Law76]. That algorithm works even for *multisets*, i.e., when points have weights[2]. The results for the Euclidean version of the problem are given in the following table.[3]

Paper	Approx.	Time
[Vai89]	1	$N^{5/2} \log^{O(1)} N$
[AES95]	1	$N^{2+\delta}, \delta > 0$
[AV99]	$1 + \epsilon$	$N^{3/2}(\log N + 1/\epsilon)^{O(1)}$
[Cha02, IT03]	$\log n$	$N \log n$
[AV04]	$\log(1/\delta)$	$N^{1+\delta} \log^{O(1)} N, \delta > 0$
This paper	$O(1)$	$N \log^{O(1)} N$

The main result of this paper is a constant-factor approximation algorithm with running time $N \log^{O(1)} N$. As a consequence, we also obtain a constant-factor approximation algorithm for the problem of finding a translation $T \in \mathbb{R}^2$ which minimizes $EMD(T(A), B)$. This is because, as shown in [KV05, CGKR05], simply aligning the centroids of A and B produces a "good enough" translation.

1.1 Techniques

The algorithm is obtained in two steps. Firstly, we show that the matching between the whole sets A and B can be "decomposed" into several matchings between subsets of those sets. The decomposition is done is such a way that the cost of the total matching is well-approximated by the sum of the costs of (sub)-matchings computed for the subsets. The sum can be then, in principle, approximated via random sampling. However, the costs of the submatchings can vary; in particular, it is possible that the cost of one submatching dominates the whole sum. Thus, the sampling of the submatchings needs to be done by using random distribution where the probability of choosing a submatching is roughly proportional to its cost. To compute the distribution, we compute a rough (logarithmic) approximation of the cost of each submatching; this can be done very quickly. Then, a small random sample of submatchings suffices to estimate the total cost.

It should be noted that the algorithm does *not* produce an approximately optimal matching, but only estimates its cost. This drawback seems to be inherent to the random sampling-based approach. Also, unlike the algorithms of [Cha02, IT03, NS06], our algorithm cannot be transformed into an embedding of the EMD metric into l_1. In fact, a recent result of [NS06] shows that any such embedding must have distortion at least $\sqrt{\log n}$, while our algorithm provides constant approximation factor.

2 Preliminaries

In the following, all set operations (union, selection, etc) are performed on multisets.

*This work was supported in part by NSF CAREER award CCR-0133849, David and Lucille Packard Fellowship and Alfred P. Sloan Fellowship.

[1]Note that for planar pointsets, the $\|\cdot\|_1$ and $\|\cdot\|_2$ norms differ only by a factor of $\sqrt{2}$.

[2]The weight of a point can be thought of as the number of times the point appears in the multiset.

[3]Note that some of the algorithms require that the pointsets A, B are discrete, that is, they are subsets of $[n]^2$ for an integer n. As we see in Preliminaries, we can make this assumption to hold (essentially) without loss of generality, for n close to N.

2.1 Setup Let $\epsilon > 0$ be some small constant. In the following we apply simple (and standard) transformations to the input to make it more manageable. The transformations will change the cost of the objective function by at most a factor of $1 \pm \epsilon$.

We start by producing an estimate $T > 0$ such that $T \leq EMD(A, B) \leq \lambda T$, for some $\lambda > 0$. One way of doing it is to use the algorithm of [AV04] (as in the table) with $\delta = 1/\log N$; this gives $\lambda = \log \log N$. Alternatively, we can use a simpler algorithm of [Cha02, IT03]; however, its running time and approximation factor depend logarithmically on the aspect ratio n.

By multiplying all coordinates by $(2N/\epsilon)/T$, we can assume $T = 2N/\epsilon$. Moreover, if we round each coordinate to its nearest integer, the EMD between the sets changes by at most $\pm 2N$, i.e., $\pm \epsilon T$. Thus, we can assume that the points in A and B have integer coordinates, and this changes the objective function by at most $(1 + \epsilon)$.

Consider now a grid G with side length $n = 2T\lambda$. Impose this grid on the plane, shifted at random. Since any pair of points a, b is "cut" by the grid with probability $\|a - b\|_1/n$, it follows that the probability that any edge of the minimal matching between A and B is cut is at most

$$EMD(A, B)/n \leq \lambda T/n \leq 1/2$$

Thus, with probability $1/2$, the problem decomposes into several bi-chromatic matching subproblems, which can be solved separately. The sets in each subproblem are subsets of $[n]^2$.

2.2 Importance sampling Importance sampling is a statistical variance reduction technique for sampling-based estimation (see [MR95], p. 312 for further description and some applications to algorithm design). The idea is as follows. Assume we want to estimate a sum $Z = \sum_{i=1}^{s} Z_i$, but we do not want to compute all Z_i's. One way of doing it is by sampling. For example, we could choose an index i uniformly at random from $[s]$, and use a random variable $S = sZ_i$ to estimate Z. In particular, $E[S] = Z$. However, the variance of S could be quite large, if there are few "large" elements Z_i. The idea of importance sampling is to assign higher probability mass to such Z_i's to ensure they are more likely to be picked.

A specific version of importance sampling that we use is defined in the following lemma.

LEMMA 2.1. *Consider a probability distribution defined by $p_1 \dots p_s \geq 0$, and values $Z_1 \dots Z_s \geq 0$. Let $Z = \sum_i Z_i$ and $q_i = Z_i/Z$. Assume that $q_i \leq \lambda p_i$ for some $\lambda \geq 1$. Consider a random variable S such that $\Pr[S = Z_i/p_i] = p_i$; note that $E[S] = Z$. Then, the variance of S is at most $Z^2\lambda = \lambda E^2[Z]$.*

Proof. The variance of S is at most

$$E[S^2] = \sum_i p_i (Z_i/p_i)^2 = Z^2 \sum_i q_i \cdot q_i/p_i \leq Z^2 \lambda$$

The above lemma enables us to use standard (Chebyshev) bounds to show that, for any $\epsilon > 0$, $O(\lambda/\epsilon^2)$ samples suffices to estimate Z up to $(1 \pm \epsilon)$ with constant probability.

2.3 Probabilistic embeddings Consider a metric space (X, D), and a distribution $\mathcal{D}$ over pairs $[(X', D'), f]$, where (X', D') is a metric and $f : X \to X'$. Following [Bar96], we say that (X, D) probabilistically embeds into $\mathcal{D}$ with distortion c if for every $a, b \in X$:

- For every $[(X', D'), f]$ in $\mathcal{D}$, $D(a, b) \leq D'(f(a), f(b))$
- $E_{\mathcal{D}}[D'(f(a), f(b))] \leq cD(a, b)$

3 Algorithm

We start by extending EMD to sets of non-equal size. Consider multisets $A, B \subset [n]^2$, $|A|, |B| \leq N$. Define

$$\begin{aligned} EEMD_n(A, B) &= \min_{S \subset A, S' \subset B, |S| = |S'|} [EMD(S, S') \\ &+ n(|A - S| + |B - S'|)] \end{aligned}$$

If n is clear from the context, we skip the subscript.

Note that $2n$ is equal to the diameter of $[n]^2$. As a result, the minimum is always realized for $|S| = |S'| = \min(|A|, |B|)$; otherwise, there is a pair of points $a \in A - S$, $b \in B - S'$, which can be matched at a cost at most $2n = n + n$, so they can be as well included in S and S', respectively. As a result, if $|A| \leq |B|$, we can alternatively define $EEMD(A, B)$ as

$$EEMD_n(A, B) = \min_{S \subset B, |S| = |A|} EMD(A, S) + n|B - A|$$

LEMMA 3.1. *$EEMD_n(\cdot, \cdot)$ is a metric.*

Proof. Consider an (extension) set $X = [n]^2 \cup [N]$. We extend the l_1 metric over $[n]^2$ to X by defining $D(a, b) = n$ if either $a \in [N]$ or $b \in [N]$, and $D(a, b) = \|a - b\|_1$ otherwise. For each set A, we define $\tilde{A} = A \cup [N - |A|]$. It can be easily verified that $EMD(\tilde{A}, \tilde{B}) = EEMD_n(A, B)$. Thus, $EEMD_n(A, B)$ is a metric.

Now we show how to decompose $EEMD_n$ into a sum of metrics $EEMD_m$ for $m << n$. The decomposition induces some (constant) distortion, and is later used in the algorithm. The ideas used here are not terribly new - the algorithm of [AV04] used a similar partitioning of the plane using randomly shifted grids ([Cha02, IT03] used a simpler version of such partitioning as well). In those papers, as well as here, the partitioning enables us to reduce the original problem over "large" grid to several subproblems over smaller grids. However, for our purpose, we need to ensure that the subproblems are constructed *independently* from each other. This is because the final estimation is performed by a (biased) sampling of the subproblems. Our decomposition result can be phrased (Theorem 3.1) as a low-distortion probabilistic embedding of $EEMD_n$ into a weighted sum of $EEMD_m$'s, where m is much smaller than n.

The decomposition procedure is as follows. Consider an (arbitrary shifted) grid G, with cell side length m, imposed over $[n]^2$. Formally, we will interpret G as a set of cells, naturally associated with $[k]^2$ for $k \leq \lceil n/m \rceil + 1 \leq 2n/m$ (assuming n,m are large enough); this also induces the l_1 metric on G. For any $a \in [n]^2$, $G(a)$ denotes the cell containing a. For any *multiset* $A \subset [n]^2$, $G(A)$ is a *multiset* consisting of all points $G(a)$, $a \in A$. For any cell $c \in G$, we define $A_c = \{a \in A : G(a) = c\}$. We can think about A_c as (multi)subsets of $[m]^2$.

The grid naturally decomposes $EEMD_n$. That is:

Lemma 3.2. *For any $A, B \subset [n]^2$, we have*

$$\begin{aligned} EEMD_n(A,B) &\leq \sum_{c \in G} EEMD_m(A_c, B_c) \\ &+ mEEMD_k(G(A), G(B)) \end{aligned}$$

Proof. Assume $|A| \leq |B|$. We will construct a matching between A and a subset $S \subset B$ as follows. Firstly we construct the matching within cells. For a given cell c, assume $|A_c| \leq |B_c|$. We match points in A_c with a subset $S_c \subset B_c$; this has cost $EMD(A_c, S_c)$. The remaining points $B_c - S_c$ for all c are matched between different cells, or not matched at all. Each of the latter adds a cost of n, charged to the $mk|G(A) - G(B)|$ term of $mEEMD(G(A), G(B))$ (note that $n \leq mk$). To match the points between cells, observe that a point $a \in A_c$ and $b \in B_{c'}$ can be matched by a path that goes from a to the center of c, then to the center of c', then to b. The cost of this connection is at most $m + m\|c - c'\|_1 + m$. We charge the first term to the $m|A_c - B_c|$ term of $EEMD(A_c, B_c)$, the second term to $mEMD(G(A), G(B))$, and the third term to the $m|A_{c'} - B_{c'}|$ term of $EEMD(A_{c'}, B_{c'})$,

The above inequality holds for any placement of the grid G on $[n]^2$. The following two lemmas will show that, if the grid G is shifted at random, then an approximate version of the *reverse* inequality holds as well (in the expectation). Specifically, we have the following two lemmas.

Lemma 3.3. *Consider a random variable $Z = \sum_{c \in G} EEMD_m(A_c, B_c)$. We have*

$$E[Z] \leq 2 \cdot EEMD_m(A, B).$$

Proof. Assume without loss of generality that $|A| \leq |B|$. Consider any matching between A and $S \subset B$, $|A| = |S|$, and consider any of its edges (a, b). The probability that this edge is cut by a randomly shifted grid is at most $p(a, b) = \frac{\|a-b\|_1}{m}$. If the edge is cut we add (at most) m to Z, otherwise we add $\|a - b\|_1$ to Z. Thus,

$$\begin{aligned} E[Z] &\leq m|A - B| + \sum_{(a,b)} (p(a,b)m + \|a-b\|_1) \\ &= m|A - B| + \sum_{(a,b)} (\|a-b\|_1 + \|a-b\|_1) \\ &\leq m|A - B| + 2EMD(A, S) \end{aligned}$$

Lemma 3.4.

$$E[mEEMD(G(A), G(B))] \leq EEMD(A, B)$$

Proof. Follows from similar argument to the proof of Lemma 3.3

We will now apply the above lemmas in a recursive manner. That is, we impose a grid G_1 on $[n]^2$, then a grid G_2 on G_1, and so on. The last grid is denoted by G_t. All grids have cell side length m. As a result G_t has dimensions $M_t \times M_t$ where $M_t \leq n2^t/m^t$. For $\delta > 0$ we will choose the parameters $t = O(1/\delta)$ and $m = O(n^\delta)$ so that $M_t \leq m$.

Consider any $i = 1 \ldots t$. Define $G^i(A) = G_i(G_{i-1}(\ldots G_1(A)))$. That is, $G^i(A)$ is "representation" of A using the grid G_i. Define

$$X_i = m^i \sum_{c \in G^i} EEMD(G^{i-1}(A)_c, G^{i-1}(B)_c)$$

and

$$Y = m^t EEMD(G^t(A), G^t(B))$$

By applying Lemma 3.4 ($i-1$ times) and Lemma 3.3 (once) we get $E[X_i] \leq 2 \cdot EEMD(A, B)$. By applying Lemma 3.4 (t times) we get $E[Y] \leq EEMD(A, B)$. Therefore

$$E[\sum_{i=1}^{t} X_i + Y] \quad \leq \quad (2t + 1)EEMD(A, B)$$

At the same time, by applying Lemma 3.2 in an analogous fashion, we get $EEMD(A, B) \leq \sum_{i=1}^{t} X_i + Y$ Therefore, we have the following.

THEOREM 3.1. *For any $\delta > 0$, $EEMD_n$ can be probabilistically embedded into a weighted sum of metrics $EEMD_m$, $m = n^\delta$ (with non-negative weights), with distortion $O(1/\delta)$. That is, there is a distribution $\mathcal{D}$ over T-tuples of mappings $< f_1, \ldots, f_T >$, $f_i : [n]^2 \rightarrow [m]^2$, and weights $< w_1, \ldots, w_T >$, such that for any $A, B \subset [n]^2$:*

- $EEMD_n(A,B) \leq \sum_i w_i EEMD_m(f_i(A), f_i(B))$ *with probability* 1
- $E[\sum_i w_i EEMD_m(f_i(A), f_i(B))] \leq O(1/\delta) \cdot EEMD_n(A,B)$

Moreover, all weights and images $f_i(A)$ can be computed in time $O(|A|/\delta)$; in particular, $\sum_i |f_i(A)| = O(|A|/\delta)$.

Note: in our mapping the weights w_i are *fixed*.

It remains to show an efficient algorithm for the EEMD approximation. Firstly, we generate the weights w_i and mappings f_i as per Theorem 3.1. Let $S = \sum_i w_i EEMD(f_i(A), f_i(B))$. By Markov inequality we get that $\Pr[S \geq 4 \cdot O(1/\delta) \cdot EEMD(A,B)] \leq 1/4$. Also, $S \geq EEMD(A,B)$. Thus, it suffices to estimate S. For simplicity of notation, we (conceptually) replicate each $EEMD$ metric several times, so that we have $w_i = 1$ for all i. Let $Z_i = EEMD(A_i, B_i)$, then $S = \sum_i Z_i$.

Our estimation algorithm uses *importance sampling*. As described in Preliminaries, we compute estimations E_i of Z_i such that $E_i \leq Z_i \leq \lambda E_i$. This takes $O(N \log^{O(1)} N)$ time.

Define $E = \sum_i E_i$, $p_i = E_i/E$. We use Lemma 2.1 to estimate $\sum_i Z_i$ up to a factor of $(1+\epsilon)$, for some small constant $\epsilon > 0$, using probabilities p_i, with probability of correctness greater than $1 - 1/5$. Since each value Z_i can be evaluated exactly in time $O(m^6)$ (using the Hungarian algorithm), it follows that the estimation of Z_i can be done in time $O(\lambda m^6)$. We can choose the constant δ so that $\lambda m^6 = o(N)$. Thus, the total running time is $N \log^{O(1)} N$.

THEOREM 3.2. *There is an algorithm that, given $A, B \subset [n]^2$, $|A| = |B| = N$, in time $O(N \log^{O(1)} N)$ outputs an estimate C such that $C \leq EEMD(A,B) \leq O(C)$ with probability at least $1 - 1/5 - 1/4 > 1/2$.*

4 Conclusions

In this paper we presented a constant-factor approximation for computing the cost of minimum bi-chromatic matching in the plane. As its predecessors [IT03, AV04] it can be easily extended to $\mathbb{R}^d$ for constant d.

The algorithm uses a combination of two ideas: a decomposition of EMD metric into several metrics over smaller domains, and calculating the total cost by sampling the metrics, using probabilities which approximate the costs of individual metrics. This combination could be useful for other problems as well.

Acknowledgments

The author would like to thank Tasos Sidiropoulos, Kasturi Varadarajan, Julia Chuzhoy and Sariel Har-Peled for helpful comments on this paper.

References

[AES95] P. K. Agarwal, A. Efrat, and M. Sharir. Vertical decomposition of shallow levels in 3-dimensional arrangements and its applications. *Proceedings of the ACM Symposium on Computational Geometry*, 1995.

[AV99] P.K. Agarwal and K. Varadarajan. Approximation algorithms for bipartite and non-bipartite matching in the plane. *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 1999.

[AV04] P. Agarwal and K. Varadarajan. A near-linear constant factor approximation for euclidean matching ? *Proceedings of the ACM Symposium on Computational Geometry*, 2004.

[Bar96] Y. Bartal. Probabilistic approximation of metric spaces and its algorithmic applications. *Proceedings of the Symposium on Foundations of Computer Science*, 1996.

[CGKR05] S. Cabello, P. Giannopoulos, C. Knauer, and G. Rote. Matching point sets with respect to the earth mover's distance. *Proceedings of the European Symposium on Algorithms*, pages 520–531, 2005.

[Cha02] M. Charikar. Similarity estimation techniques from rounding. *Proceedings of the Symposium on Theory of Computing*, 2002.

[IT03] P. Indyk and N. Thaper. Fast color image retrieval via embeddings. *Workshop on Statistical and Computational Theories of Vision (at ICCV)*, 2003.

[KV05] O. Klein and R. C. Veltkamp. Approximation algorithms for computing the earth mover's distance under transformations. *Proceedings of the 16th Annual Symposium on Algorithms and Computation*, 2005.

[Law76] E. Lawler. *Combinatorial optimization: Networks and Matroids*. Holt, Rinehart and Winston, 1976.

[MR95] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[NS06] A. Naor and G. Schechtman. Planar earthmover is not in l_1. *Proceedings of the Symposium on Foundations of Computer Science*, 2006.

[RTG00] Y. Rubner, C. Tomassi, 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.

[Vai89] P. Vaidya. Geometry helps in matching. *SIAM Journal on Computing*, 18:1201–1225, 1989.