Space-Efficient Approximation Scheme for Circular Earth Mover Distance[*]

Joshua Brody[1], Hongyu Liang[2], and Xiaoming Sun[3]

[1] Aarhus University, Aarhus, Denmark joshua.e.brody@gmail.com
[2] Institute for Interdisciplinary Information Sciences, Tsinghua University
lianghy08@mails.tsinghua.edu.cn
[3] Institute of Computing Technology, Chinese Academy of Sciences
xiaoming.sun@gmail.com

Abstract. The Earth Mover Distance (EMD) between point sets A and B is the minimum cost of a bipartite matching between A and B. EMD is an important measure for estimating similarities between objects with quantifiable features and has important applications in several areas including computer vision. The streaming complexity of approximating EMD between point sets in a two-dimensional discretized grid is an important open problem proposed in [8, 9].

We study the problem of approximating EMD in the streaming model, when points lie on a discretized circle. Computing the EMD in this setting has applications to computer vision [13] and can be seen as a special case of computing EMD in on a discretized grid. We achieve a $(1 \pm \varepsilon)$ approximation for EMD in $\tilde{O}(\varepsilon^{-3})$ space, for every $0 < \varepsilon < 1$. To our knowledge, this is the first streaming algorithm for a natural and widely applied EMD model that matches the space bound asked in [9].

1 Introduction

For two multisets A, B of points of equal sizes in a space $\mathcal{S}$, the *Earth Mover Distance (EMD)* between A and B is defined as the minimum cost of a perfect matching between points in A and B, where the cost function is identical to the distance function equipped with the space $\mathcal{S}$.

[*] This work was supported in part by the National Basic Research Program of China Grant 2007CB807900, 2007CB807901, and the National Natural Science Foundation of China Grant 61033001, 61073174. The authors acknowledge support from the Danish National Research Foundation and The National Science Foundation of China (under the grant 61061130540) for the Sino-Danish Center for the Theory of Interactive Computation, within which this work was performed. The authors also acknowledge support from the Center for Research in Foundations of Electronic Markets (CFEM), supported by the Danish Strategic Research Council. Joshua Brody is also supported in part by the Danish Council for Independent Research grant LOBO 438146. Xiaoming Sun is also supported in part by the National Natural Science Foundation of China Grant 61170062 and Tsinghua University Initiative Scientific Research Program 2009THZ02120

When restricted on specific spaces, the Earth Mover Distance becomes a natural measure for estimating the similarity between two objects with quantifiable features, and thus has found important applications in various areas. Starting with the work of [16, 17], the Earth Mover Distance has attracted significant attention and interest in the area of computer vision. This is because an image, in different contexts, can be represented as a collection of representative features, such as pixels in a color space [17], object contours [4], hue histograms [18], SIFT-like descriptors [7], circular histograms [13], and others [5]. The Earth Mover Distance is thus an appropriate measure of similarity between images. The considered point spaces can vary according to different applications. In many situations, the space is a d-dimensional integer grid $[\varDelta]^d$ for some integers $\varDelta$ and d, with ℓ_1-distance being the distance metric.[4] For example, an image can be represented as a set of pixels each of which is a point in the 3-dimensional color space [17]. Another important application of EMD in computer vision is to compare one-dimensional circular histograms [13], where the point space is a (discretized) circle in the 2D Euclidean space and the distance between two points on the circle is the length of the shortest *arc* connecting them on the circle. Due to its particular structure, the EMD over such space is also called the *Circular Earth Mover Distance (CEMD)* [13].

Since the computation of EMD can be easily reduced to the weighted bipartite matching problem, it can be solved optimally in $O(n^3)$ time and $O(n^2)$ space, where n is the size of the point-sets. Nevertheless, in many applications, the sizes of the point-sets are very large, and we may need to select a large number (sometimes millions) of feature sets and compute all the corresponding EMD's. Thus, the commonly used matching algorithm is not satisfactory. This motivates the exploration of *approximation algorithms* for EMD that run faster or use less working space. When considering space-bounded computation, an extensively-studied algorithmic setting is the streaming model, in which the input data are given in a streaming fashion and only limited working and storing space is allowed. This model dates back to [10] and was popularized by Alon, Matias and Szegedy [1]. For a survey of related results we refer the readers to [11].

An important open problem in the streaming literature, proposed in [8], is whether EMD over 2-dimensional integer grids $[\varDelta]^2$ with ℓ_1-distance admits a constant factor approximation algorithm in the one-pass streaming model that uses $\log^{O(1)}(n\varDelta)$ space, where n is the size of the given point-sets. Currently the best known algorithm, due to Andoni et al. [2], can maintain an $O(1/\varepsilon)$-approximation of the exact value of EMD between two point-sets in $[\varDelta]^2$ using $O(\varDelta^{\varepsilon}\log^{O(1)}(n\varDelta))$ space and update time for $0 < \varepsilon < 1$. This amount of space still has a $\varDelta^{\Theta(1)}$ gap from the conjectured bound in [8, 9]. Furthermore, Naor and Schechtman [12] showed that any ℓ_1 embedding of EMD on $[\varDelta]^2$ incurs distortion $\varOmega(\sqrt{\varDelta})$, suggesting that embeddings alone are unlikely to produce space-efficient $O(1)$-approximations of EMD. On the other hand, things get much easier when dealing with 1-dimensional grids $[\varDelta]^1$. It is folklore that the EMD between two point-sets in $[\varDelta]^1$ is equal to the ℓ_1 distance between two

[4] We use $[\varDelta]$ to denote the set $\{0, 1, \ldots, \varDelta - 1\}$.

corresponding vectors in $[\Delta]^n$. In the streaming model, we can reduce EMD to problem of estimating the ℓ_1-norm of a vector in the turnstile model [1] (in which input tokens stand for update operations on the coordinates of the vector), and by [6] this implies that EMD over $[\Delta]^1$ allows a $(1\pm\varepsilon)$-approximation streaming algorithm using $O(\varepsilon^{-2}\log(n\Delta))$ space (see Section 2 for more details). Note that this space complexity meets the bound asked in [8]. As little progress has been made towards the 2-dimensional case during these years, a natural target is to find an "intermediate" space that "lies between" $[\Delta]^1$ and $[\Delta]^2$, on which the EMD problem has space-efficient constant factor approximation algorithms.

In this paper we study the streaming complexity of Circular Earth Mover Distance (CEMD) mentioned before. In the traditional algorithmic setting, the complexity of this problem has already been well understood. It is shown in [20, 3] that the problem can be solved in $O(n\log n)$ time where n is the size of the point-sets, and can even be solved in $O(n)$ time if the points are sorted on the circle in advance. However, neither this approach nor the ones in [20, 3] is space efficient; they all require $\Omega(n)$ space when converted to a (one-pass) streaming algorithm.

Our Contributions. We present a $(1\pm\varepsilon)$-approximation one-pass streaming algorithm for CEMD that uses $\tilde{O}(\varepsilon^{-3}\log(n\Delta))$ space and succeeds with probability 0.99, for every $0<\varepsilon<1$. To our knowledge, this is the first streaming algorithm for a natural and widely applicable EMD model that matches the space bound asked in [8]. It is also not difficult to see that the circle space, in some sense, lies between the 1-dimensional and 2-dimensional spaces.

The central part of our results is a theorem establishing the quality of matchings obtained from a random cut approach. Specifically, for every $0<\epsilon<1$, by cutting the circle at a point chosen uniformly at random, the matching induced by the obtained line segment is a $(1+\varepsilon)$-approximation with probability $\Omega(\varepsilon)$ (see Theorem 3). By repeating this process $O(\varepsilon^{-1})$ times independently and returning the minimum estimate, we get a $(1+\varepsilon)$-approximation with probability 0.99. This, combined with the streaming algorithm for ℓ_1-distance in the turnstile model given by [6], yields a streaming algorithm for CEMD (Theorem 4).

2 Preliminaries

A *metric space* $\mathcal{S}$ is a pair (S,d_S) where S is a set of elements (or points) and $d_S : S\times S\rightarrow[0,\infty)$ is a symmetric distance function defined on pairs of points in S. Given a space $\mathcal{S}=(S,d_S)$ and two finite, equal-sized (multi-)sets $A,B\subseteq S$, the *Earth Mover Distance (EMD)* between A and B (over $\mathcal{S}$) is defined as:

$$EMD_{\mathcal{S}}(A,B):=\min_{\phi:A\rightarrow B}\sum_{p\in A}d_S(p,\phi(p)),$$

where the minimum is taken over all bijections ϕ between A and B.

In the streaming version of the Earth Mover Distance problem, the input stream consists of $2n$ tokens (C,p), where $C\in\{A,B\}$ and $p\in\mathcal{S}$. A token (C,p) means $p\in C$. The goal is to compute the Earth Mover Distance between A and B specified by the tokens. We assume that the $2n$ tokens can come in an *arbitrary order*, which makes the problem harder and makes our result stronger.

One-Dimensional EMD. Consider the *1-dimensional grid space* $[\Delta]^1 = ([\Delta], d)$, where Δ is a positive integer, and $d(a, b) := |a - b|$ for all $a, b \in [\Delta]$. Let A and B be two equal-sized subsets of $[\Delta]$. Suppose $A = \{a_1, a_2, \ldots, a_n\}$ and $B = \{b_1, b_2, \ldots, b_n\}$, where $a_1 \leq a_2 \leq \ldots \leq a_n$ and $b_1 \leq b_2 \leq \ldots \leq b_n$. By [19] (or simple observations) we have

$$EMD_{[\Delta]^1}(A, B) = \sum_{i=1}^{n} |a_i - b_i|,$$

and this is achieved when a_i is matched with b_i for every $1 \leq i \leq n$. Such matching will be called the *canonical matching* between A and B. Using the result of [6] for ℓ_1-norm estimation, we obtain[5]:

Theorem 1 ([6]). *For any $0 < \varepsilon, \delta < 1$, there is a one-pass streaming algorithm that $(1 \pm \varepsilon)$-approximates 1-dimensional EMD with probability at least $1 - \delta$ using $O(\varepsilon^{-2} \log(n\Delta) \log(1/\delta))$ space.*

Circular EMD. Let Δ be a positive integer. For any integer a, define $(a)_\Delta := a \bmod \Delta$. Let $\mathscr{C} := ([\Delta], d_\Delta)$ where d_Δ is defined as:

$$\text{For all } p_1, p_2 \in [\Delta],\ d_{[\Delta]}(p_1, p_2) := \min\{(p_1 - p_2)_\Delta, (p_2 - p_1)_\Delta\}.$$

We can imagine that the Δ points in $[\Delta]$ are drawn clockwisely on a circle of circumference Δ, in the order $0, 1, 2, \ldots, \Delta - 1$, such that every two adjacent points have distance 1 on the circle. Then $d_\Delta(p_1, p_2)$ is just the length of the shortest arc connecting p_1 and p_2 on the circle. (See Figure 1 for an example with $\Delta = 8$.) Hereinafter we will always use this circle realization of the space $\mathscr{C}$.

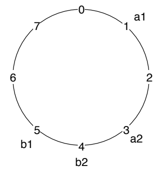

Fig. 1: An example of $\Delta = 8$

Let $A = \{a_1, a_2, \ldots, a_n\} \subseteq [\Delta]$ and $B = \{b_1, b_2, \ldots, b_n\} \subseteq [\Delta]$ be two subsets of $[\Delta]$ of size n (which can be multisets). The points in

[5] When points from $A \cup B$ appear on the stream in arbitrary (instead of sorted) order, there is a subtle issue in mapping the EMD input to an appropriate input for the ℓ_1-norm estimator. The solution is easy and appears to be folklore; we leave a complete discussion to the full version of the paper.

A and B are also called *A-points* and *B-points*, respectively. Let OPT denote the Earth Mover Distance between A and B over $\mathscr{C}$, i.e., $OPT := EMD_{\mathscr{C}}(A, B)$. Throughout this paper, an instance of the circular EMD problem consists of the space $\mathscr{C}$ (specified entirely by Δ) and the two sets A, B. The goal is to compute OPT. (See Figure 1 for an example where $n = 2$, $A = \{1, 3\}$ and $B = \{5, 4\}$, in which case $OPT = 5$.)

We need some more notations. For the simplicity of expressions and without loss of generality, we assume that $A \cup B$ is not a multiset, i.e., A and B are simple sets and $A \cap B = \emptyset$. This assumption can be made without loss of generality; we explain in the full version of this paper how to easily obtain the same results for the general case.

Cutting Points. For any point $p \in [\Delta]$, let $\mathscr{C}_p$ denote the space $([\Delta], d_p)$, where d_p is defined as follows:

$$d_p(p_1, p_2) = \begin{cases} (p_2 - p_1)_\Delta & \text{if } p, p_1, p_2 \text{ appear clockwisely;} \\ (p_1 - p_2)_\Delta & \text{otherwise.} \end{cases}$$

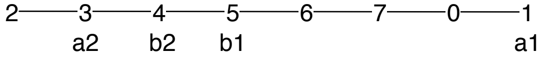

Fig. 2: An example of $\mathscr{C}_2$

Notice that $\mathscr{C}_p$ also has an intuitive realization as follows: We cut the circle $\mathscr{C}$ at the point p, and then "straighten" it to obtain a line segment, ensuring that p is the leftmost point. Then for $p_1, p_2 \in [\Delta]$, $d_p(p_1, p_2)$ is exactly the (normal) distance between p_1 and p_2 on this line segment. (See Figure 2 for $\mathscr{C}_2$ where the original space $\mathscr{C}$ is specified by Figure 1.) Hereafter we shall identify $\mathscr{C}_p$ with the corresponding line segment. In this sense, p is also called the *cutting point* of $\mathscr{C}_p$. Clearly $\mathscr{C}_p$ is isomorphic to $[\Delta]^1$. To ease notation, we write $EMD(\mathscr{C}_p) := EMD_{\mathscr{C}_p}(A, B)$, the EMD between A and B over $\mathscr{C}_p$. Crucial to our results is the following theorem in [14] (whose full proof can be found in [15]):

Theorem 2 (Equation (2.4) in [14]). $OPT = \min_{p \in A \cup B}\{EMD(\mathscr{C}_p)\}$.

Note that Theorem 2 holds for the case where $A \cup B$ can be a multiset. Using cutting points allows us to leverage known space-efficient approximations for $EMD_{[\Delta]^1}(A, B)$ (Theorem 1), as shown in the next section.

3 A Streaming Algorithm for Circular EMD

In this section, we develop an efficient streaming algorithm for CEMD that maintains a $(1 \pm \varepsilon)$-approximation with high probability. As mentioned in the introduction, we do this by randomly selecting a set of cut points, and estimating the Earth Mover Distance on the each resulting line segment using known approximation algorithms.

Indeed, an intuitive explanation is the following: viewing the optimal matching between A and B as a series of directed edges from $a \in A$ to $\phi(a) \in B$, it is easy to see that if no arc is cut when we cut the circle at p, then $EMD(\mathscr{C}_p) = OPT$. The proof of Theorem 2 in [14] shows that it is always possible to find $p \in A \cup B$ whose best matching has no arc across p; thus, computing $EMD(\mathscr{C}_p)$ for each p suffices to compute OPT. Unfortunately, we do not have enough space to even approximate EMD for all n points in $A \cup B$. Instead we take a few random cut points. Our key technical contribution is a result showing that the EMD of a random cut $\mathscr{C}_p$ gives $(1+\varepsilon)$-approximation to $CEMD(A, B)$ with nontrivial probability. This result is captured in the following theorem, whose technical proof we defer until Section 4.

Theorem 3. *Choose a cutting point $p \in [\varDelta]$ uniformly at random. Then, for every ε such that $0 < \varepsilon < 1/6$, $\mathbf{Pr}[EMD(\mathscr{C}_p) \le (1 + 10\varepsilon)OPT] \ge \varepsilon$.*

Theorem 4. *For any $0 < \varepsilon, \delta < 1$, there is a one-pass streaming algorithm for $(1 \pm \varepsilon)$-approximating CEMD that uses $O(\varepsilon^{-3} \log^2(1/(\varepsilon\delta)) \log(n\varDelta))$ space and succeeds with probability at least $1 - \delta$.*

Proof. Fix $0 < \varepsilon, \delta < 1$. Our algorithm first chooses $k := \lceil 100\varepsilon^{-1} \ln(2/\delta) \rceil$ points from $[\varDelta]$ with repetition, say $p_1, p_2, \ldots, p_k$, and stores them in memory. This initial step takes $O(k \log \varDelta)$ space. Then we apply the algorithm in Theorem 1, using parameters $\varepsilon' = \varepsilon/3$ and $\delta' = \delta/2k$, to estimate $EMD(\mathscr{C}_{p_i})$ for all $1 \le i \le k$ in parallel. The space used during this process is at most k times that of approximating 1-dimensional EMD using Theorem 1. Let the k estimated distances be $E_1, \ldots, E_k$. We take the minimum of them as our estimation of OPT. For each $1 \le i \le k$, we know that $EMD(\mathscr{C}_{p_i}) \ge OPT$ always holds, and, by Theorem 3,

$$\mathbf{Pr}[EMD(\mathscr{C}_{p_i}) \le (1 + \varepsilon/3)OPT] \ge \varepsilon/30. \tag{1}$$

From Theorem 1 and our choice of ε' and δ', we have

$$\mathbf{Pr}[(1 - \varepsilon/3)EMD(\mathscr{C}_{p_i}) \le E_i \le (1 + \varepsilon/3)EMD(\mathscr{C}_{p_i})] \ge 1 - \delta/2k. \tag{2}$$

Therefore, for each $1 \le i \le k$,

$$\mathbf{Pr}[E_i < (1 - \varepsilon/3)OPT] \le \mathbf{Pr}[E_i < (1 - \varepsilon/3)EMD(\mathscr{C}_{p_i})] \le \delta/2k.$$

This holds for any $1 \le i \le k$, so by the union bound, we have

$$\mathbf{Pr}[\min\{E_i \mid 1 \le i \le k\} < (1 - \varepsilon/3)OPT] \le k \cdot \delta/2k = \delta/2. \tag{3}$$

From (1) and (2), the fact that $(1 + \varepsilon/3)^2 < 1 + \varepsilon$ for all $0 < \varepsilon < 1$, and another union bound, we have

$$\varepsilon/60 \le \varepsilon/30 - \delta/2k \le \mathbf{Pr}[E_i \le (1+\varepsilon/3)^2 OPT] \le \mathbf{Pr}[E_i \le (1+\varepsilon)OPT] .$$

It follows that

$$\mathbf{Pr}[\min\{E_i \mid 1 \le i \le k\} > (1 + \varepsilon)OPT] \le (1 - \varepsilon/60)^k \le \delta/2. \tag{4}$$

By (3) and (4) we obtain

$$\mathbf{Pr}[(1 - \varepsilon/3)OPT \le \min\{E_i \mid 1 \le i \le k\} \le (1 + \varepsilon)OPT] \ge 1 - \delta.$$

The total used space is at most

$$O(k \log \varDelta) + k \cdot O((\varepsilon')^{-2} \log(1/\delta') \log(n\varDelta)) = O(\varepsilon^{-3} \log^2(1/(\varepsilon\delta)) \log(n\varDelta)).$$

4 $(1+\varepsilon)$-approximation of OPT

In this section we prove our main lemma stating that a simple solution can $(1+\varepsilon)$-approximate OPT with probability $\Omega(\varepsilon)$. A key component of our analysis breaks the circle into a series of *intervals* and analyzes how much a matching moves points from A to B *across* each interval. Before getting to the proof, some definitions are required.

Intervals. Let $p_1, p_2 \in [\Delta]$. The *interval* $[p_1, p_2]$ is the set of points obtained by starting at p_1 and travelling in a clockwise fashion until reaching p_2. A *left-open interval* $(p_1, p_2]$ is defined similarly, except p_1 is not included. We define the *length* of an interval to be its size and write $len(I) := |I|$. Unless otherwise specified (e.g., "an interval $[a, b]$"), we assume an interval I to be a left-open. For any interval $I = (p_1, p_2]$, let $l(I) := p_1$ and $r(I) := p_2$ denote the *left endpoint* and *right endpoint* of I, respectively.

Definition 1. *An interval I is* simple *if*

- $l(I) \in A \cup B$;
- $I \cap (A \cup B) = \{r(I)\}$.

Thus, the endpoints of a simple interval are both points in $A \cup B$, and there are no other A- or B-points lying inside the interval. Let $\mathcal{I}$ denote the set of all simple intervals. Since $|A \cup B| = 2n$, it is clear that $|\mathcal{I}| = 2n$. For example, in Figure 1, we have $\mathcal{I} = \{(a_1, a_2], (a_2, b_2], (b_2, b_1], (b_1, a_1]\}$. Note that $\mathcal{I}$ partitions $[\Delta]$.

Matchings and Coefficients. Let $p \in [\Delta]$. The *canonical matching* between A and B over $\mathscr{C}_p$, which is (one of) the matching(s) having cost $EMD(\mathscr{C}_p)$, naturally induces n intervals whose endpoints are pairs of matched A- and B-points. Let $\mathcal{M}_p$ denote the set of these n intervals associated with $\mathscr{C}_p$. By definition we have $EMD(\mathscr{C}_p) = \sum_{I \in \mathcal{M}_p} len(I)$. For any simple interval $I \in \mathcal{I}$, the *coefficient of I in $EMD(\mathscr{C}_p)$*, denoted by $c_p(I)$, is defined to be the number of intervals in $\mathcal{M}_p$ that contain I, i.e., $c_p(I) := |\{J \mid I \subseteq J \in \mathcal{M}_p\}|$. It is clear that

$$EMD(\mathscr{C}_p) = \sum_{I \in \mathcal{I}} c_p(I) \cdot len(I). \tag{5}$$

We start with the following lemma that relates the coefficient of a simple interval with the numbers of A- and B-points in a corresponding set.

Lemma 1. *For every $p \in [\Delta]$ and every simple interval $I \in \mathcal{I}$,*

$$c_p(I) = \left| |[p, l(I)] \cap A| - |[p, l(I)] \cap B| \right|.$$

That is, the coefficient of I in $EMD(\mathscr{C}_p)$ equals the (absolute) difference between the number of A-points and that of B-points in $[p, l(I)]$.

Proof. Fix $p \in [\Delta]$ and $I \in \mathcal{I}$. Assume without loss of generality that in the canonical matching between A and B over $\mathscr{C}_p$, a_j is matched with b_j and the corresponding interval is $[a_j, b_j]$, for every $1 \leq j \leq n$. (If, for some $1 \leq j \leq n$, the interval is $[b_j, a_j]$ instead of $[a_j, b_j]$, we can simply switch the roles of a_j and b_j in the following argument when dealing with this j.) Since I is a simple interval, it holds that for every $1 \leq j \leq n$,

$$I \subseteq [a_j, b_j] \text{ if and only if } a_j \in [p, l(I)] \text{ and } b_j \notin [p, l(I)]. \tag{6}$$

We consider two cases. First suppose $c_p(I) = 0$, i.e., no interval $[a_j, b_j]$ contains I. According to (6), for each $1 \leq j \leq n$, either a_j and b_j are both in $[p, l(I)]$, or they are both in $[\Delta] \setminus [p, l(I)]$. Therefore the numbers of A-points and B-points in $[p, l(I)]$ are equal, implying that $c_p(I) = 0 = ||[p, l(I)] \cap A| - |[p, l(I)] \cap B||$, which proves the first case. Next suppose that $c_p(I) \geq 1$. Let $S = \{j \mid I \subseteq [a_j, b_j]\}$. Then by definition we have $c_p(I) = |S|$. Let j_1 be the smallest index in S. Due to (6) we have $a_{j_1} \in [p, l(I)]$ and $b_{j_1} \notin [p, l(I)]$. By the definition of the canonical matching, we have $b_j \notin [p, l(I)]$ for all $j \in S$, and hence $a_j \in [p, l(I)]$ for all $j \in S$. From (6) we know that for all $j \notin S$, a_j and b_j are either both in $[p, l(I)]$ or both in $[\Delta] \setminus [p, l(I)]$. Thus, the difference between the numbers of A-points and B-points in $[p, l(I)]$ is exactly $|S|$, which is equal to $c_p(I)$. This finishes the proof of Lemma 1.

Based on Lemma 1, we further give some definitions and prove some useful lemmas. Let $p^* \in A \cup B$ be such that $OPT = EMD(\mathscr{C}_{p^*})$. (The existence of p^* is ensured by Theorem 2). For any integer $i \in \mathbb{Z}$, define

$$\mathcal{T}_i := \{I \in \mathcal{I} \mid |[p^*, l(I)] \cap A| - |[p^*, l(I)] \cap B| = i\}$$

By Lemma 1 we have that for any interval $I \in \mathcal{I}$, $c_{p^*}(I) = i$ if and only if $I \in \mathcal{T}_i \cup \mathcal{T}_{-i}$. Let $t = \max\{| i | \mid \mathcal{T}_i \neq \emptyset\}$. Clearly $1 \leq t \leq n$; it is also easy to see that $\{\mathcal{T}_i\}$ partition $\mathcal{I}$. The next lemma is less obvious.

Lemma 2. *If $\mathcal{T}_i = \emptyset$ for some $i \geq 0$, then $\mathcal{T}_j = \emptyset$ for all $j \geq i$. If $\mathcal{T}_i = \emptyset$ for some $i \leq 0$, then $\mathcal{T}_j = \emptyset$ for all $j \leq i$.*

Proof. Assume that $i \geq 0$ (the case where $i \leq 0$ is handled in the same manner), and assume for some $j \geq i$ holds that $\mathcal{T}_i = \emptyset$ and $\mathcal{T}_j \neq \emptyset$. Let I^* be the simple interval containing p^*, i.e., $p^* \in (l(I^*), r(I^*)]$. Then $c_{p^*}(I^*) = 0$. By Lemma 1, for every two adjacent simple interval I_1, I_2, $|c_r(I_1) - c_r(I_2)| \leq 1$ (since $[p^*, l(I_1)]$ and $[p^*, l(I_2)]$ differ by at most one element from $A \cup B$). Now choose an arbitrary $I \in \mathcal{T}_j$. Starting from I, we visit clockwisely every simple interval until we reach I^*. Since $c_{p^*}(I) = j$ and $c_{p^*}(I^*) = 0$, there exists a simple interval I' for which $c_{p^*}(I') = i$. Hence $\mathcal{T}_i \neq \emptyset$ which contradicts with our assumption. □

Let $\mathcal{P} := \bigcup_{i=1}^{t} \mathcal{T}_i$, $\mathcal{N} := \bigcup_{i=1}^{t} \mathcal{T}_{-i}$, and $\mathcal{Z} := \mathcal{T}_0$. In this way, $\mathcal{P}, \mathcal{N}$, and $\mathcal{Z}$ represent the sets of simple intervals having positive, negative, and zero coefficient values respectively. In a similar way, we define $Z := \bigcup_{I \in \mathcal{Z}} I$, and for each $1 \leq i \leq t$, we define $P_i := \bigcup_{I \in \mathcal{T}_i} I$, and $N_i := \bigcup_{I \in \mathcal{T}_{-i}} I$. Note that these are sets of points while sets like $\mathcal{P}$ and $\mathcal{N}$ defined before are collection of simple intervals. Finally, let $P := \bigcup_{i=1}^{t} P_i$ and $N := \bigcup_{i=1}^{t} N_i$. Clearly, $[\Delta] = P \cup N \cup Z$. Noting that $P_i = \emptyset$ (resp., $N_i = \emptyset$) if and only if $\mathcal{T}_i = \emptyset$ (resp., $\mathcal{T}_{-i} = \emptyset$), and applying Lemma 2, we obtain:

Lemma 3. *If $P_i = \emptyset$ for some $i \geq 0$, then $P_j = \emptyset$ for all $j \geq i$. Similar result holds also for N_i.*

The following lemma entirely determines the coefficient of any simple interval in any cut. The proof is not difficult, but requires a lot of case analysis. For lack of space, we defer it until the full version of the paper.

Lemma 4. *Fix $0 \leq i \leq t$. For any $I \in \mathcal{I}$, we have:*

$$c_p(I) = \begin{cases} c_{p^*}(I) + i & \text{if } p \in P_i \text{ and } I \in \mathcal{N} \cup \mathcal{Z}, \\ |c_{p^*}(I) - i| & \text{if } p \in P_i \text{ and } I \in \mathcal{P}, \\ c_{p^*}(I) + i & \text{if } p \in N_i \text{ and } I \in \mathcal{P} \cup \mathcal{Z}, \\ |c_{p^*}(I) - i| & \text{if } p \in N_i \text{ and } I \in \mathcal{N}. \end{cases}$$

The following corollary is immediate from Lemma 4.

Corollary 1. *$EMD(\mathscr{C}_p) = EMD(\mathscr{C}_{p^*}) = OPT$ for every $p \in Z$.*

We are now ready to prove our main theorem.
Theorem 3 (Restated). Choose a cutting point $p \in [\Delta]$ uniformly at random. Then, for every ε such that $0 < \varepsilon < 1/6$,

$$\boldsymbol{Pr}[EMD(\mathscr{C}_p) \leq (1 + 10\varepsilon)OPT] \geq \varepsilon.$$

Proof. Choose $p \in [\Delta]$ uniformly at random. Then $\mathbf{Pr}[p \in S] = |S|/\Delta$ for any subset $S \subseteq Z$. Thus for every interval I,

$$\mathbf{Pr}[p \in (l(I), r(I)]] = (r(I) - l(I))/\Delta = len(I)/\Delta.$$

If $|Z| \geq \varepsilon\Delta$, then from Corollary 1 we obtain

$$\mathbf{Pr}[EMD(\mathscr{C}_p) = OPT] \geq \mathbf{Pr}[p \in Z] = |Z|/\Delta \geq \varepsilon,$$

and thus the theorem holds. In the remaining of the proof we assume that $|Z| = |P_0 \cup N_0| < \varepsilon\Delta$, which implies that $|\bigcup_{i=1}^{t}(P_i \cup N_i))| = |P \cup N| > (1 - \varepsilon)\Delta > \varepsilon\Delta$. Let k be the smallest nonnegative integer for which $\left|\bigcup_{i=0}^{k}(P_i \cup N_i)\right| \geq \varepsilon\Delta$. Then we have $1 \leq k \leq t$ and

$$\left|\bigcup_{i=0}^{k-1}(P_i \cup N_i)\right| < \varepsilon\Delta. \tag{7}$$

We know that $\mathbf{Pr}[p \in \bigcup_{i=0}^{k}(P_i \cup N_i)] \geq \varepsilon$. Thus, the following claim will conclude the proof of Theorem 3.

Claim. For every $p \in \bigcup_{i=0}^{k}(P_i \cup N_i)$, $EMD(\mathscr{C}_p) \leq (1 + 10\varepsilon)OPT$.

Proof. Let i be any integer such that $0 \leq i \leq k$. Pick an arbitrary cutting point $p \in P_i \cup N_i$. (This can be done since if $P_i \cup N_i = \emptyset$, then by Lemma 3, $P_{i'} \cup N_{i'} = \emptyset$ for all $i' \geq i$, and thus by (7) we have $|P \cup N| = |\bigcup_{j=0}^{i-1}(P_i \cup N_i)| < \varepsilon\Delta$. This gives $\Delta = |P \cup N| + |Z| < 2\varepsilon\Delta < \Delta$, which is a contradiction.)

We only prove the claim for the case $p \in P_i$, since another case $p \in N_i$ is similar. When $i = 0$, the claim follows directly from Corollary 1, so we assume that $i \geq 1$. Due to Lemma 4, we have:

$$\begin{aligned}
EMD(\mathscr{C}_p) &= \sum_{I \in \mathcal{P} \cup \mathcal{N} \cup \mathcal{Z}} c_p(I) \cdot len(I) \\
&= \sum_{I \in \mathcal{N} \cup \mathcal{Z}} (c_{p^*}(I) + i) len(I) + \sum_{I \in \mathcal{P}} |c_{p^*}(I) - i| \cdot len(I) \\
&= \sum_{I \in \mathcal{N} \cup \mathcal{Z}} (c_{p^*}(I) + i) len(I) + \sum_{j=i}^{t} \sum_{I \in \mathcal{P}_j} (j - i) len(I) + \sum_{j=1}^{i-1} \sum_{I \in \mathcal{P}_j} (i - j) len(I).
\end{aligned}$$

For similar reasons, we know that

$$OPT = EMD(\mathscr{C}_{p^*}) = \sum_{I \in \mathcal{I}} c_{p^*}(I) \cdot len(I) = \sum_{I \in \mathcal{N} \cup \mathcal{Z}} c_{p^*}(I) \cdot len(I) + \sum_{j=1}^{t} \sum_{I \in \mathcal{P}_j} j \cdot len(I).$$

Therefore,

$$\begin{aligned}
0 &\leq EMD(\mathscr{C}_p) - OPT \\
&= \sum_{I \in \mathcal{N} \cup \mathcal{Z}} i \cdot len(I) - \sum_{j=i}^{t} \sum_{I \in \mathcal{P}_j} i \cdot len(I) + \sum_{j=1}^{i-1} \sum_{I \in \mathcal{P}_j} (i - 2j) len(I) \\
&\leq i \sum_{I \in \mathcal{N} \cup \mathcal{Z}} len(I) - i \sum_{j=i}^{t} \sum_{I \in \mathcal{P}_j} len(I) + i \sum_{j=1}^{i-1} \sum_{I \in \mathcal{P}_j} len(I).
\end{aligned}$$

By definition we have $\sum_{I \in \mathcal{N}} len(I) = |N|$, $\sum_{I \in \mathcal{Z}} len(I) = |Z| < \varepsilon \varDelta$, and $\sum_{I \in \mathcal{P}_j} len(I) = |P_j|$. Thus,

$$0 \leq EMD(\mathscr{C}_p) - OPT \leq i(|N| + \varepsilon \varDelta - \sum_{j=i}^{t} |P_j| + \sum_{j=1}^{i-1} |P_j|). \tag{8}$$

This indicates that

$$|N| \geq \sum_{j=i}^{t} |P_j| - \sum_{j=1}^{i-1} |P_j| - \varepsilon \varDelta = \sum_{j=1}^{t} |P_j| - 2 \sum_{j=1}^{i-1} |P_j| - \varepsilon \varDelta = |P| - 2 |\bigcup_{j=1}^{i-1} P_j| - \varepsilon \varDelta.$$

Using (7) and the fact that $i \leq k$, we have $|\bigcup_{j=1}^{i-1} P_j| \leq \varepsilon \varDelta$, and hence

$$|N| \geq |P| - 3\varepsilon \varDelta. \tag{9}$$

We show that $N_i \neq \emptyset$. Assume to the contrary that $N_i = \emptyset$, then by Lemma 3 we have $N_{i'} = \emptyset$ for all $i' \geq i$. Thus by (7) it holds that $|N| = |\bigcup_{j=0}^{i-1} N_i| \leq |\bigcup_{j=0}^{k-1} (P_i \cup N_i)| < \varepsilon \varDelta$. Then from (9) we get $|P| \leq |N| + 3\varepsilon \varDelta < 4\varepsilon \varDelta$, and thus $\varDelta = |P| + |N| + |Z| < 4\varepsilon \varDelta + \varepsilon \varDelta + \varepsilon \varDelta = 6\varepsilon \varDelta < \varDelta$, which is a contradiction. Hence our assumption is false, which proves that $N_i \neq \emptyset$. So there exists at least one point $p' \in N_i$. By symmetry, if we use $p' \in N_i$ instead of p and repeat the above steps, we can obtain a counterpart of (9) as follows:

$$|P| \geq |N| - 3\varepsilon \varDelta. \tag{10}$$

Using (10) in (8) yields that

$$\begin{aligned}EMD(\mathscr{C}_p) - OPT &\le i(|N| + \varepsilon\Delta - \sum_{j=i}^{t}|P_j| + \sum_{j=1}^{i-1}|P_j|)\\ &= i(|N| - |P| + 2|\bigcup_{j=1}^{i-1} P_j| + \varepsilon\Delta)\\ &\le 6i\varepsilon\Delta.\end{aligned}$$

Notice that

$$\begin{aligned}OPT &\ge \sum_{j=i}^{t}\sum_{I\in\mathcal{P}_j\cup\mathcal{N}_j} c_{p^*}(I)\cdot len(I) = \sum_{j=i}^{t}\sum_{I\in\mathcal{P}_j\cup\mathcal{N}_j} j\cdot len(I)\\ &\ge i\sum_{I\in\bigcup_{j=i}^{t}(\mathcal{P}_j\cup\mathcal{N}_j)} len(I) = i\cdot|\bigcup_{j=i}^{t}(P_j\cup N_j)|\\ &= i\cdot(\Delta - |\bigcup_{j=0}^{i-1}(P_j\cup N_j)|)\\ &\ge i\cdot(\Delta - \varepsilon\Delta) \quad \text{(using (7) and that } i\le k)\\ &= i(1-\varepsilon)\Delta.\end{aligned}$$

Therefore, as $\varepsilon < 1/6$,

$$EMD(\mathscr{C}_p) \le OPT + 6i\varepsilon\Delta \le OPT + \frac{6\varepsilon}{1-\varepsilon}OPT \le (1+10\varepsilon)OPT.$$

□

References

1. N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999.
2. A. Andoni, K. Do Ba, P. Indyk, and D. P. Woodruff. Efficient sketches for earth-mover distance, with applications. In *Proceedings of the 50th Annual Symposium on Foundations of Computer Science (FOCS)*, 2009.
3. C. A. Cabrelli and U. M. Molter. A linear time algorithm for a matching problem on the circle. *Information Processing Letters*, 66(3):161–164, 1998.
4. K. Grauman and T. Darrell. Fast contour matching using approximate Earth Movers distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.
5. K. Grauman and T. Darrell. Efficient image matching with distributions of local invariant features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

6. D. M. Kane, J. Nelson, and D. P. Woodruff. On the exact space complexity of sketching and streaming small norms. In *Proceedings of the 21st ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2010.
7. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
8. A. McGregor. Open problems in data streams and related topics. In *IITK Workshop on Algorithms For Data Streams*, 2006. Available at `http://www.cse.iitk.ac.in/users/sganguly/workshop.html`.
9. A. McGregor. Open problems in data streams, property testing, and related topics. In *Bernitoro Workshop on Sublinear Algorithms*, 2011.
10. J. I. Munro and M. Paterson. Selection and sorting with limited storage. *Theoretical Computer Science*, 12(3):315–323, 1980.
11. S. Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1(2):117–236, 2005.
12. A. Naor and G. Schechtman. Planar earthmover is not in L_1. *SIAM Journal on Computing*, 37(3):804–826, 2007. Preliminary version in FOCS 2006.
13. J. Rabin, J. Delon, and Y. Gousseau. Circular earth mover's distance for the comparison of local features. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*. IEEE Computer Society, 2008.
14. J. Rabin, J. Delon, and Y. Gousseau. A statistical approach to the matching of local features. *SIAM Journal on Imaging Sciences*, 2(3):931–958, 2009.
15. J. Rabin, J. Delon, and Y. Gousseau. Transportation distances on the circle. *Journal of Mathematical Imaging and Vision*, 41(1–2):147–167, 2011.
16. Y. Rubner, C. Tomassi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Proceedings of the 6th International Conference on Computer Vision (ICCV)*, 1998.
17. 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.
18. R. Venkatesh Babu, P. Pérez, and P. Bouthemy. Robust tracking with motion estimation and local kernel-based color modeling. *Image and Vision Computing*, 25(8):1205–1216, 2007.
19. C. Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Math. Soc., 2003.
20. M. Werman, S. Peleg, R. Melter, and T.Y. Kong. Bipartite graph matching for points on a line or a circle. *Journal of Algorithms*, 7(2):277–284, 1986.