# Better embeddings for planar Earth-Mover Distance over sparse sets

by

Arturs Backurs

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2014





Signature redacted

Author .....
Department of Electrical Engineering and Computer Science
May 13, 2014

Signature redacted

Certified by .....
Piotr Indyk
Professor
Thesis Supervisor

Signature redacted

Accepted by .....
Leslie A. Kolodziejski
Chair, Department Committee on Graduate Students

# Better embeddings for planar Earth-Mover Distance over sparse sets

by

Arturs Backurs

Submitted to the Department of Electrical Engineering and Computer Science
on May 13, 2014, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Computer Science

## Abstract

We consider the problem of constructing low-distortion embeddings of the Planar Earth-Mover Distance ($EMD$) into $\ell_p$ spaces. $EMD$ is a popular measure of dissimilarity between sets of points, e.g., bags of geometric features. We present a collection of embeddings with the property that their distortion and/or host-space dimension are parametrized by the size (or the sparsity) of the embedded sets $s$. Our specific results include:

- An $O(\log s)$-distortion embedding of $EMD$ over $s$-subsets into $\ell_{1-\epsilon}$. This is the first embedding of $EMD$ into a "tractable" $\ell_p$ space whose distortion is a function of the sparsity, not the size of the ambient space;
- An $O(\log n)$-distortion embedding of $EMD$ into $\ell_1$ with dimension $O(s^2 \log^2 n)$, where the embedded sets are subsets of an $n \times n$ grid. For low values of $s$ this significantly improves over the best previous dimension bound of $O(n^2)$ obtained for general sets.

Thesis Supervisor: Piotr Indyk
Title: Professor

# Acknowledgments

I thank my advisor Piotr Indyk for the continuous support.

I would also like to thank the anonymous reviewers for SoCG'14 for their insightful comments about the paper.

Also, I would like to thank Ilya Razenshteyn for the useful feedback on SoCG'14 submission.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

The Earth Mover Distance ($EMD$) between two sets of points in $\mathbb{R}^d$ of the same size (say, $s$) is defined to be the cost of the minimum cost bipartite matching between the two point-sets. For an integer $n$, $[n]$ denotes set $\{0, 1, 2, ..., n-1\}$.

**Definition 1** *Given two pointsets $A, B \subseteq \mathbb{R}^d$ of equal size $|A| = |B| = s$, the Earth Mover Distance between $A$ and $B$ is equal to*

$$EMD(A, B) = \min_{\pi} \sum_{i \in [s]} \|a_i - b_{\pi(i)}\|_1,$$

*where $\pi$ ranges over all permutations of $s$ elements.*

Computing the minimum cost bi-chromatic matching is one of the most fundamental problems in geometric optimization, and there has been an extensive body of work focused on designing efficient algorithms for this problem. More recently, $EMD$ has been extensively studied as a metric for computing the similarity of sets of geometric features. For example, if an image is represented as a set of pixels in a three-dimensional color space (this is known as "bag of pixels" representation), computing $EMD$ between such sets is known to yield an accurate measure of dissimilarity between color characteristics of the images [10]. In an analogous manner, an image can be represented as a set of representative geometric features, such as object contours [4]; and other features [4].

**Definition 2** *Suppose $A$ and $B$ be two weighted pointsets of $\mathbb{R}^d$ with $|A| = |B| = s$ and corresponding non-negative weight functions $m^a$ and $m^b$. We define*

$$EMD(A,B) = \min_{\substack{m_{ij}\geq 0: \\ \forall i, m_i^a = \sum_j m_{ij} \\ \forall j, m_j^b = \sum_i m_{ij}}} \sum_{ij} m_{ij} \|a_i - b_j\|_1.$$

Unfortunately, a major difficulty in employing $EMD$ is the lack of efficient algorithms for dealing with it. For example, the best exact algorithm known for computing planar $EMD$ runs in time $O(s^{3/2+\delta}\log(sn))$ for $\delta > 0$, where all coordinates are positive integers bounded by $n$ [11]. In contrast, computing the $\ell_p$ distance between two vectors can be trivially done in linear time. $\ell_p$ distance between two vectors $x, y \in \mathbb{R}^k$ is defined to be

$$\|x - y\|_p := \left( \sum_{i\in[k]} |x_i - y_i|^p \right)^{1/p}.$$

Furthermore, in many applications, one needs to search for a similar feature set in a large collection of sets. Naively, this requires a number of $EMD$ computations that is linear in the size of the collection. This approach is clearly not scalable to large data sets, which can easily contain millions of feature sets.

To alleviate this issue, there has been a significant interest in developing methods for *geometric* representation of $EMD$. The goal is to design mappings (say, $f$) that map a set of points $A$ into a vector $f(A)$ in a $k$-dimensional space, such that the distance $EMD(A,B)$ between any two point sets is approximated by $\|f(A) - f(B)\|_p$, for some norm $\|\cdot\|_p$. Formally, we want to ensure that for any two sets $A, B$ we have

$$EMD(A,B) \leq \|f(A) - f(B)\|_p \leq C \cdot EMD(A,B) \tag{1.1}$$

for a parameter $C$ called *distortion*. After performing the mapping, the distances can be estimated simply by computing the distances between the vectors. Similarity search can be solved by running the algorithms developed for the $\ell_p$ spaces.

As evident from the table, all of the upper bounds so far incur the distortion that is at least $\log(n)$. It can be also observed that the dimension of the host space

| | Distortion | Dimension | Comments |
|---|---|---|---|
| [5, 6] | $O(d \log n)$ | $O(n^3)$ | Any $s$ |
| [9] | $O(\log n)$ | $O(n^2)$ | Any $s$, $d = 2$ |
| [1] | $O(\log(dn) \log(s))$ | $O(nsd \cdot \log(dsn))$ | |
| [9] | $\Omega(\sqrt{\log n})$ | | $d = 2$, $s = n^2$ |
| [8] | $\Omega(d)$ | | $s = n^d$ |

Table 1.1: Summary of the results on embedding $EMD$ into simpler spaces. The embeddings work for subsets of $[n]^d$, the discrete $d$-dimensional space. We denote the size of the sets by $s$. All of the embeddings are linear and embed into $\ell_1$.

is at least $n$. Moreover, the lower bound implies that at least $\Omega(\sqrt{\log n})$ distortion is necessary for general sparsity $s$. Thus, in order to achieve improved bounds, one needs to restrict the sparsity of the embedded sets. This is a natural restriction, since in many applications the feature sets are indeed sparse.

**Our results** We show a collection of results on embedding $EMD$ over sparse sets (see Table 1). We focus on the case of planar $EMD$, i.e., $d = 2$, although similar results can be obtained for any constant $d$. Concretely, we show that:

- $EMD$ over $s$-subset can be embedded with distortion $O(\log s)$ into $\ell_{1-\epsilon}$ with dimension $n^2$, where $\epsilon = 1/\log s$. This breaks the aforementioned $\log n$ distortion barrier limiting all results so far.

- The dimension of the embedding can be improved further, to $O(s \log s)$, by allowing two relaxations: *snow-flaking*, i.e., embedding into $\ell_1$ raised to the power of $1 + \epsilon$; and *randomization*, i.e., allowing the Equation 1.1 to hold with a constant probability over a random choice of mappings. These relaxations are natural in the context of using the embeddings for the similarity search problems.

- For the dual problem of minimizing the dimension, we show that it is possible to construct an $O(\log n)$-distortion embedding into $\ell_1$ with dimension only $O(s^2 \log^2 n)$. For low values of $s$ this provides an exponential improvement over the earlier bounds. We further extend this result to $s$-sparse measures, with

| | From | To | Distortion | Dimension |
|---|---|---|---|---|
| Th. 6 | $s$-subsets of $[n]^2$ | $\ell_{1-\epsilon}$ | $O(\log s)$ | $O(n^2)$ |
| Cor. 1 | $s$-subsets | $(\ell_1)^{1+\epsilon}$ | $O(\log s)$ | $O(s \log s)$ |
| Th. 7 | $s$-subsets of $[n]^2$ | $\ell_1$ | $O(\log n)$ | $O(s^2 \log^2 n)$ |
| Th. 8 | $s$-sparse measures of $[n]^2$ | $\ell_1$ | $O(\log n)$ | $O(s^5 \log^2 n)$ |
| Cor. 2 | $s$-subsets of $[n]^2$ | $(\ell_1)^{1+\epsilon}$ | $O(\log s)$ | $O(1) \cdot s^2(\log s) \cdot (\log \log s) \cdot \log n$ |

Table 1.2: Results ($\epsilon = 1/\log s$). All of the embeddings are deterministic except the second one.

only a polynomial loss in the dimension, where $s$-sparse measure is one that has support of size at most $s$.

- Finally, we show that the distortion can be further reduced to $O(\log s)$ by allowing snow-flaking.

Since snow-flaking does not affect the applicability of the embeddings to nearest neighbor search, we can combine those embeddings with the approximate near neighbor search algorithms for $\ell_1$ (e.g., see [7]) to obtain $O(\log s)$-distortion algorithms for the $EMD$ metric. That is, given $t$ size-$s$ pointsets $S_1, S_2, S_3, ..., S_t$, approximation parameter $c > 1$ and threshold $r > 0$, we can construct a data structure of size $poly(t, n, d)$ such that, given a query pointset $Q$, we can output a pointset $S_i$ with $EMD(S_i, Q) \leq O(\log s) \cdot r$ in time $o(t)$ provided there exists a pointset $S_j$ with $EMD(S_j, Q) \leq r$.

**Our techniques** Our results fall into two (overlapping) classes. The first class contains embeddings that reduce the distortion, from $O(\log n)$ to $O(\log s)$, while the second class contain dimensionality reduction results, that reduce the dimension from $O(n^2)$ to polynomial in $s$ and $\log n$. Both embeddings use the "standard" $O(\log n)$-distortion embedding of [5, 6] as a starting point. The latter embedding proceeds by constructing a quad-tree of a plane, shifted by a random vector (this step can be derandomized by enumerating all shifts). The levels of the quad-tree correspond to nested grids, and each node in the quad-tree corresponds to a grid cell, as well as a coordinate in the image vector. The value of that coordinate is equal to the number of points falling into the grid cell times its side-length.

In order to reduce the distortion, we first observe that a slight modification of the $O(\log n)$-distortion embedding provides an embedding of $EMD$ over a snowflaked two-dimensional plane metric $(\ell_1^2)^{1-\epsilon}$, with distortion $O(1/\epsilon)$. This is analogous to the proof of the Assouad's theorem [2]. Then we observe that, for sets of size at most $s$ and for $\epsilon = 1/\log s$, the $EMD$ over a snowflaked plane is up to constant factors equivalent to $EMD^{1-\epsilon}$. This leads to a probabilistic $O(\log s)$-distortion embedding of $EMD^{1-\epsilon}$ into $\ell_1$, or equivalently of $EMD$ into $(\ell_1)^{1/(1-\epsilon)}$. To remove snowflaking, we observe that the coordinates in the image of the embedding are integers with absolute value up to $O(s)$, multiplied by fixed weights. This lets us replace $\ell_1$ with $(\ell_{1-\epsilon})^{1-\epsilon}$, or equivalently $(\ell_1)^{1/(1-\epsilon)}$ with $\ell_{1-\epsilon}$.

In order to reduce the dimension, the main step is to construct embeddings that result in image vectors that are sparse. By using the result of [3] we can then reduce the number of entries to be linear in sparsity and logarithmic in the original dimension. The main difficulty in ensuring that the image vectors, i.e., vectors that we obtain by performing the embedding, are sparse comes from the fact that, to make the "standard" embedding deterministic, one needs to enumerate $n$ distinct shifts in order to ensure $O(\log n)$ distortion. In particular, reducing the number of shifts by making them "coarser" would increase the distortion of a short edge, since the probability that both of its endpoints fall to the same node would be no longer proportional to its length. This implies that the dimension is at least $\Omega(n)$. To overcome this issue, we introduce a "fuzzy" variant of the embedding, where a point falling into a cell is weighted based on the distance from the point to the boundary of the cell. Together with some other ideas, this allows us to reduce the number of shifts needed to derandomize the embedding to a value polynomial in the sparsity.

# Chapter 2

# Preliminaries

In this section we introduce the notation and tools used in the rest of the thesis.

**Notation** We will use $A$ and $B$ to denote sets of points in $\ell_1^2$. We could also consider $\ell_2^2$ instead $\ell_1^2$. This changes the distortion by at most a constant factor.

By $u \oplus v$ we denote $(u_1, v_1, u_2, v_2, u_3, v_3, ...)$ for vectors $u$ and $v$. We could define $u \oplus v$ as concatenation of vectors $u$ and $v$, but we choose this definition because we want that it works for the case when vectors are infinite dimensional.

For an integer $n$, by $[n]$ we denote $\{0, 1, 2, 3, ..., n-1\}$.

**Tools** The following lemma can be shown using the probabilistic method.

**Lemma 1** *For any $\delta > 0$ consider a random bipartite graph with the left set $A = [k]$, left degree $d = O(1/\delta)$ and the right set $B = [m]$ for $m = O(k/\delta^2)$. For every left node we choose its $d$ neighbors independently and uniformly at random from the right nodes (some right node could be chosen multiple times for the same left node). Let $N(X)$ be the set of the neighbors of a set $X$. Then:*

$$\Pr[\forall X \subseteq A : |N(X)| \geq (1-\delta)d|X|] \geq 0.99.$$

We will also need a "deterministic version" of the aforementioned lemma.

**Definition 3** *A $(k, \epsilon)$-unbalanced expander is a bipartite simple graph $G = (A, B, E)$*

*with left degree $d$ such that for any $X \subset A$ with $|X| \leq k$, the set of neighbors $N(X)$ of $X$ has size $|N(X)| \geq (1-\epsilon)d|X|$.*

The following deterministic analog of Lemma 1 is "folklore", see e.g., [3]:

**Proposition 1** *For any $n/2 \geq k \geq 1$ and $\epsilon > 0$, there exists a $(k, \epsilon)$-unbalanced expander with left degree $d = O(\log(n/k)/\epsilon)$ and right set size $O(kd/\epsilon) = O(k \log(n/k)/\epsilon^2)$.*

We will use unbalanced expanders to reduce the dimension of sparse vectors. Specifically, we will use the following dimensionality reduction tool:

**Definition 4** *An $m \times n$ matrix $\Phi$ is said to satisfy $RIP1_{k,\delta}$ (restricted isometry property with parameters $k$ and $\delta$) if, for any $k$-sparse vector $x$, we have*

$$\|x\|_1 \leq \|\Phi x\|_1 \leq (1+\delta)\|x\|_1.$$

It was shown in [3] that unbalanced expanders yield RIP1 matrices.

**Theorem 1** *Consider any $m \times n$ matrix $\Phi$ that is the adjacency matrix of an $(k, \epsilon)$-unbalanced expander $G = (A, B, E)$ with left degree $d$, such that $1/\epsilon$, $d$ are smaller than $n$. Then the scaled matrix $\Phi/d$ satisfies the $RIP1_{k,C\epsilon}$ property for some absolute constant $C > 1$.*

# Chapter 3

# Probabilistic embeddings

In this section we show the main ideas behind $O(\log s)$-distortion embeddings, in the simpler probabilistic setting.

## 3.1 Snowflaked version of $EMD$

We start by defining EMD over a snowflaked plane.

**Definition 5** *Suppose $A$ and $B$ are pointsets (from $\ell_1^2$) of equal cardinalities. For $\epsilon > 0$ we define*

$$EMD_\epsilon(A, B) = \min_\pi \sum \|a_i - b_{\pi(i)}\|_1^{1-\epsilon},$$

*where $\pi$ ranges over all permutation of indices $\{1, 2, 3, ..., s\}$.*

We then show an equivalence between EMD over a snowflaked plane and a snowflaked EMD.

**Lemma 2** *Let $A$ and $B$ be pointsets with $|A| = |B| = s$. Then*

$$EMD_\epsilon(A, B) = \Theta\left(EMD^{1-\epsilon}(A, B)\right)$$

*for $\epsilon = 1/\log s$.*

**Proof:** Let $\pi' = \text{argmin}_\pi \sum_{i=1}^{s} \|a_i - b_{\pi(i)}\|_1^{1-\epsilon}$. Then

$$\begin{aligned} EMD_\epsilon(A,B) &= \sum_{i=1}^{s} \|a_i - b_{\pi'(i)}\|_1^{1-\epsilon} \\ &\geq \left( \sum_{i=1}^{s} \|a_i - b_{\pi'(i)}\|_1 \right)^{1-\epsilon} \\ &\geq \left( \min_\pi \sum_{i=1}^{s} \|a_i - b_{\pi(i)}\|_1 \right)^{1-\epsilon} \\ &= EMD^{1-\epsilon}(A,B), \end{aligned}$$

where the first inequality holds because of the subadditivity of $(\cdot)^{1-\epsilon}$.

Let $\pi'' = \text{argmin}_\pi \sum_{i=1}^{s} \|a_i - b_{\pi(i)}\|_1$. Then

$$\begin{aligned} EMD^{1-\epsilon}(A,B) &= \left( sE_{i\in\{1,2,\ldots,s\}} \left[ \|a_i - b_{\pi''(i)}\|_1 \right] \right)^{1-\epsilon} \\ &\geq s^{1-\epsilon} E_{i\in\{1,2,\ldots,s\}} \left[ \|a_i - b_{\pi''(i)}\|_1^{1-\epsilon} \right] \\ &\geq s^{-\epsilon} \min_\pi \sum_{i=1}^{s} \|a_i - b_{\pi(i)}\|_1^{1-\epsilon} \\ &= s^{-\epsilon} EMD_\epsilon(A,B) = \Omega(EMD_\epsilon(A,B)), \end{aligned}$$

where the first inequality follows from the concavity of $(\cdot)^{1-\epsilon}$ and the last equality follows from $s^{-\epsilon} \geq \Omega(1)$ (remember that $\epsilon = 1/\log s$). ∎

## 3.2 Probabilistic embedding of snowflaked $EMD$

In this section we show that the EMD over a snowflaked plane can be embedded into $\ell_1$ with $O(\log s)$ distortion. The embedding guarantee is probabilistic, although it can be made deterministic in a standard way (by enumerating all shifts) at the cost of increasing the dimension and sparsity by a factor of $O(n)$.

**Theorem 2** *There is a probabilistic linear embedding $v$ that maps a finite set of points from $\ell_1^2$ to a vector with the following properties:*

1. $EMD_\epsilon(A,B) \leq O(\|v(A) - v(B)\|_1)$ *with probability* 1;

2. $E[\|v(A) - v(B)\|_1] \leq O(1/\epsilon) EMD_\epsilon(A, B)$.

**Proof:** The embedding is almost identical to the quad-tree embeddings in [5, 6]. The only difference is in adjusting the weights to accommodate snowflaking of the plane.

We start by choosing a random point $x \in \mathbb{R}^2$. For every $t \in \mathbb{Z}$ impose grid on the plane with side length $2^t$ such that $x$ is among the vertices of this grid. We call this grid $G_t$. We count how many points there are in each cell of grid $G_t$ and obtain vector $v_t(A)$ (the vector is infinite dimensional). To obtain an embedding $v(A)$, we concatenate $v_t(A) \cdot (2^t)^{1-\epsilon}$ for all $t \in \mathbb{Z}$.

We start by showing the lower bound, i.e., that the embedding contracts the distance by at most a constant factor. Consider the matching induced by pairing points within the same cells of grid $G_t$ when $t$ ranges from $-\infty$ to $+\infty$. That is, we match as many points as possible between $A$ and $B$ in the grid of $G_t$ so that no matching crosses border of any cell. Then we extend the matching by considering the grid of $G_{t+1}$ by matching non-matched points from $G_t$. The number of points that gets matched in grid $G_{t+1}$ but was not previously matched in grid $G_t$, is

$$\|v_t(A) - v_t(B)\|_1 - \|v_{t+1}(A) - v_{t+1}(B)\|_1.$$

The cost for matching a pair in $G_{t+1}$ is at most $(2^{t+1} \cdot 2)^{1-\epsilon}$. Given that $EMD_\epsilon(A, B)$ is the minimum among all possible matchings, we obtain the following inequality

$$\begin{aligned}
&EMD_\epsilon(A, B) \\
&\leq \sum_{t \in \mathbb{Z}} \frac{1}{2} \left( \|v_t(A) - v_t(B)\|_1 - \|v_{t+1}(A) - v_{t+1}(B)\|_1 \right) 2^{(t+2)(1-\epsilon)} \\
&\leq \sum_{t \in \mathbb{Z}} 2 \cdot \left( \|v_t(A) - v_t(B)\|_1 \right) \left(2^t\right)^{1-\epsilon} = 2\|v(A) - v(B)\|_1.
\end{aligned}$$

We have factor of $\frac{1}{2}$ in the second expression because every *two* points that are not matched in level $t$ but are matched in level $t+1$, contributes at most $(2^{t+1} \cdot 2)^{1-\epsilon}$.

Now it remains to prove the upper bound. First, observe that the following function

$$Z(e,g) = \begin{cases} 1 & \text{if } e > 2 \cdot g, \\ 2e/g & \text{otherwise} \end{cases}$$

upper bounds the probability that a randomly imposed grid of side length $g$ crosses an edge of length $e$. The upper of $2e/g$ for $e \leq 2g$ follows by using union for both dimensions. Let $n_i$ be the number of edges of length at least $2^i$ and less than $2^{i+1}$ in the optimal matching of $EMD_\epsilon(A,B)$. Thus, $EMD_\epsilon(A,B) \geq \sum_{i=-\infty}^{+\infty} n_i \cdot (2^i)^{1-\epsilon}$.

Now the following inequalities hold

$$\begin{aligned}
&E[\|v(A) - v(B)\|_1] \\
&= \sum_{t \in \mathbb{Z}} \left(2^t\right)^{1-\epsilon} \cdot E[\|v_t(A) - v_t(B)\|_1] \\
&\leq \sum_{t \in \mathbb{Z}} \left(2^t\right)^{1-\epsilon} \sum_{i=-\infty}^{+\infty} 2 \cdot n_i \cdot Z(2^{i+1}, 2^t) \\
&= \sum_{i \in \mathbb{Z}} 2 \cdot n_i \left[ \sum_{t=-\infty}^{i} \left(2^t\right)^{1-\epsilon} + \sum_{t=i+1}^{+\infty} \left(2^t\right)^{1-\epsilon} \frac{2 \cdot 2^{i+1}}{2^t} \right] \\
&= \sum_{i \in \mathbb{Z}} 2 \cdot n_i \left[ \left(2^i\right)^{1-\epsilon} \cdot \frac{2^{1-\epsilon}}{2^{1-\epsilon} - 1} + 2 \cdot \left(2^{i+1}\right)^{1-\epsilon} \cdot \frac{1}{1 - 2^{-\epsilon}} \right] \\
&\leq O((1/\epsilon) EMD_\epsilon(A,B)),
\end{aligned}$$

where the first inequality holds because any edge that gets crossed by a grid of side length $2^t$, contributes to $\|v_t(A) - v_t(B)\|_1$ at most 2, the last inequality holds assuming that $\epsilon > 0$ is sufficiently small constant and using inequality $\frac{1}{1-2^{-\epsilon}} = O(1/\epsilon)$. ∎

**Theorem 3** *There is a probabilistic linear embedding that maps each subset $A \subset \ell_1^2$ of cardinality $s$ to a vector $v(A)$ with the following properties.*

1. $EMD(A,B) \leq O(\|v(A) - v(B)\|_1^{\frac{1}{1-\epsilon}})$ *with probability* 1;
2. $(E[\|v(A) - v(B)\|_1])^{\frac{1}{1-\epsilon}} \leq O(1/\epsilon) EMD(A,B)$;
3. $\|v(A) - v(B)\|_1^{\frac{1}{1-\epsilon}} \leq O(1/\epsilon) EMD(A,B)$ *with probability* $\geq 2/3$,

*where* $\epsilon = 1/\log s$.

**Proof:** The first property follows from Theorem 2 and Lemma 2. The second property follows from Lemma 2. The third property follows from the second property by using Markov's inequality and $(1/\epsilon)^{\frac{1}{1-\epsilon}} = \Theta(1/\epsilon)$. ∎

Theorem 3 provides a probabilistic embedding of EMD into a snowflaked $\ell_1$ norm with distortion $O(\log s)$. The sparsity of the images of the embedding is unbounded. However, if the embedded sets live in $[n]^2$, the sparsity can be seen to be $O(s \log n)$. Thus, by using Theorem 1, we can reduce the dimension to $O(s \log^2 n)$. Corollary 1 (below) reduces the dimension further, to $O(s \log s)$, by limiting the number of "important" quad-tree levels to $O(\log s)$, and using a probabilistic version of the dimensionality reduction theorem.

## 3.3 Dimensionality reduction for probabilistic embedding

**Theorem 4** *For any* $C > 1$ *and* $\alpha, \epsilon > 0$ *and any two pointsets* $A$ *and* $B$ *with* $|A| = |B| = s$ *the following holds.*

*Let* $T = EMD_\epsilon(A, B)$ *and* $L = \log(\alpha(T/s)) \cdot \frac{1}{1-\epsilon}$, *and* $U = \log(C^{1-\epsilon}T) \cdot \frac{1}{1-\epsilon}$, *and* $w(A)$ *be a concatenation of vectors* $v_t(A) \cdot (2^t)^{1-\epsilon}$ *for* $t \in \{L, L+1, ..., U-1, U\}$.

*Then* $w$ *satisfies the following properties.*

1. $\sum_{t=-\infty}^{L-1} \|v_t(A) - v_t(B)\|_1 \cdot (2^t)^{1-\epsilon} = O(\alpha T)$;

2. $\Pr[v_t(A) = v_t(B)$ *for all integers* $t \geq U+1] \geq 1 - O(\frac{1}{C})$;

3. $\|w(A) - w(B)\|_1 \geq \Omega(T)$;

4. $E[\|w(A) - w(B)\|_1] \leq O(1/\epsilon) \cdot T$.

**Proof:** We prove the first property. To show that we notice that the following equality holds.

$$\begin{aligned}
&\left[\sum_{t=-\infty}^{L-1} \|v_t(A) - v_t(B)\|_1 \cdot \left(2^t\right)^{1-\epsilon}\right] \left(2^{1-\epsilon} - 1\right) \\
&= \sum_{t=-\infty}^{L} \|v_{t-1}(A) - v_{t-1}(B)\|_1 \cdot \left(2^t\right)^{1-\epsilon} \\
&\quad - \sum_{t=-\infty}^{L-1} \|v_t(A) - v_t(B)\|_1 \cdot \left(2^t\right)^{1-\epsilon} \\
&= \|v_{L-1}(A) - v_{L-1}(B)\|_1 \cdot \left(2^L\right)^{1-\epsilon} \\
&\quad + \sum_{t=-\infty}^{L-1} \left(\|v_{t-1}(A) - v_{t-1}(B)\|_1 - \|v_t(A) - v_t(B)\|_1\right) \cdot \left(2^t\right)^{1-\epsilon} \\
&\leq O(\alpha T) + O(\alpha T) = O(\alpha T),
\end{aligned}$$

where to upper bound the first summand we use that $\|v_{L-1}(A) - v_{L-1}(B)\| \leq 2s$ (we use that $|A| = |B| = s$) and to upper bound the second summand we notice that $\|v_{t-1}(A) - v_{t-1}(B)\|_1 - \|v_t(A) - v_t(B)\|_1$ is equal to number of points that gets matched in grid $G_t$, but was not previously matched in grid $G_{t+1}$, which gives

$$\sum_{t=-\infty}^{L-1} \left(\|v_{t-1}(A) - v_{t-1}(B)\|_1 - \|v_t(A) - v_t(B)\|_1\right) \cdot \left(2^t\right)^{1-\epsilon} \leq s \cdot \left(2^{L-1}\right)^{1-\epsilon} \leq O(\alpha T).$$

To prove the second property it is sufficient to show
$\Pr[v_{U+1}(A) = v_{U+1}(B)] \geq 1 - O(\frac{1}{C})$.

Let $l_i$ be lengths of edges in the optimal matching of $EMD_\epsilon(A, B)$. We have $\sum_{i=1}^{s} l_i^{1-\epsilon} = T$. We can lower bound the probability by the probability that no edge

gets crossed by a grid of side length $2^{U+1}$.

$$\begin{aligned}
&\Pr[v_{U+1}(A) = v_{U+1}(B)] \\
&\geq 1 - \frac{O\left(\sum_{i=1}^{s} l_i\right)}{2^{U+1}} \geq 1 - \frac{O\left(\left(\sum_{i=1}^{s} l_i^{1-\epsilon}\right)^{\frac{1}{1-\epsilon}}\right)}{2^{U+1}} \\
&= 1 - \frac{O(T^{\frac{1}{1-\epsilon}})}{2(C^{1-\epsilon}T)^{\frac{1}{1-\epsilon}}} = 1 - O(\frac{1}{C}),
\end{aligned}$$

where in the second inequality we use subadditivity of $(\cdot)^{1-\epsilon}$.

Notice that Property 1 of Theorem 2 combined with property 1 of Theorem 4 (by choosing $\alpha$ to be sufficiently small constant) implies $\sum_{t=L}^{+\infty} \|v_t(A) - v_t(B)\|_1 \cdot (2^t)^{1-\epsilon} \geq \Omega(T)$. The third property follows by noticing that if $v_U(A) \neq v_U(b)$, then we have $\|w(A) - w(B)\|_1 \geq \left(2^U\right)^{1-\epsilon} \geq \Omega(T)$. If $v_U(A) = v_U(b)$, then $v_u(A) = v_u(B)$ for all $u \geq U$ and we again have the required inequality.

Property 4 follows from Theorem 2. ∎

**Theorem 5** *For any two pointsets $A, B \subseteq [0, 1/2]^2$ with $|A| = |B| = s$ the following holds.*

*Let $v'(A)$ be a concatenation of vectors $v_t(A) \cdot (2^t)^{1-\epsilon}$ for all integer $t \leq Y := \log(O(s)) \cdot \frac{1}{1-\epsilon}$ and $v''(A)$ a vector with $m = O(s \log s)$ entries that we obtain by initially setting all entries to be 0 and then for each entry $e$ of $v'(A)$, increase random $d = O(1)$ (with replacement) entries of $v''(A)$ by $e/d$.*

*Then $v''$ satisfies the following properties. (We apply the same map $v''$ for both $A$ and $B$, i.e., we make the random choices only once.)*

1. *$\|v''(A) - v''(B)\|_1 \geq \Omega(EMD_\epsilon(A, B))$ with probability $\geq 2/3$;*
2. *$E[\|v''(A) - v''(B)\|_1] \leq O(1/\epsilon) EMD_\epsilon(A, B)$.*

*Also, $v''$ is a linear embedding.*

**Proof:** We can write $v'(A)$ as

$$v'(A) = \bigoplus_{t \in \mathbb{Z}: t \leq \log(sC^{1-\epsilon}) \cdot \frac{1}{1-\epsilon}} v_t(A) \cdot (2^t)^{1-\epsilon}.$$

Then we write $v''(A)$ as

$$v''(A) = \sum_{e:\text{ entry of } v'(A)} \frac{(v'(A))_e}{d} \cdot u_e,$$

where $u_e$ denotes a vector that is initially all zeroes and then we increase random $d$ entries (with replacement) by 1. Below we show that it suffices to choose $u_e$ to be a vector with $O(s \log s)$ entries for the theorem to hold.

It is easy to see that $\|v''(A)\|_1$ is finite because we consider levels with $t \leq Y$.

Let $v$ be an embedding from Theorem 2. By setting $C$ to be large enough constant and $\alpha$ to be small enough constant (in the statement of Theorem 4) we get the following. An "interesting action" is happening only in $U - L + 1 = O(\log s)$ levels (we call those levels good) with constant probability arbitrarily close to 1, i.e., only negligible mass of $\|v(A) - v(B)\|_1$ is outside $O(\log s)$ levels with probability $1 - O(1/C)$:

$$\sum_{t:(t<L) \text{ or } (t>U)} \|v_t(A) \cdot (2^t)^{1-\epsilon} - v_t(B) \cdot (2^t)^{1-\epsilon}\|_1 \leq O(\alpha EMD_\epsilon(A, B)). \tag{3.1}$$

Given that there are at most $s$ non-zero entries per level, we get that $v(A) - v(B)$ contains at most $k = O(s \log s)$ non-zero values in the good levels. In the embedding $v''$ we consider *all* good levels because $EMD_\epsilon(A, B) \leq s$ and Theorem 4.

By using Theorem 1 and Lemma 1 (setting $k = O(s \log s)$ - maximal number of non-zero values in good levels, $m = O(k/\delta^2) = O(s \log s)$ and $d = O(1/\delta) = O(1)$), we get that

$$\begin{aligned}
&\|v''(A) - v''(B)\|_1 \\
&\overset{(3.1)}{\geq} (1 - O(\delta))\|v'(A) - v'(B)\|_1 - O(\alpha EMD_\epsilon(A, B)) \\
&\geq (1 - O(\delta))\Omega(EMD_\epsilon(A, B)) - O(\alpha EMD_\epsilon(A, B)) \\
&= \Omega(EMD_\epsilon(A, B))
\end{aligned}$$

with probability $\geq 2/3$ for sufficiently small constants $\alpha > 0$ and $\delta > 0$.

The second property holds because $\|v''(A)-v''(B)\|_1 \leq \|v'(A)-v'(B)\|_1 \leq \|v(A)-v(B)\|_1$ and Theorem 2 (property 2). ■

**Corollary 1** *There exists a probabilistic linear embedding $v$ that maps pointset to a vector with $O(s \log s)$ entries such that for any two pointsets $A, B \subseteq [0, 1/2]^2$ with $|A| = |B| = s$, the following holds.*

1. $\|v(A) - v(B)\|_1^{\frac{1}{1-\epsilon}} \geq \Omega(EMD(A, B))$ *with probability* $\geq 2/3$;
2. $E[\|v(A) - v(B)\|_1] \leq O(\log s) EMD^{1-\epsilon}(A, B)$,

*where* $\epsilon = 1/\log s$.

**Proof:** Follows from Lemma 2 and Theorem 5. ■

# Chapter 4

# Deterministic embeddings

## 4.1 Deterministic $O(\log s)$-distortion embedding

**Theorem 6** *There exists a linear embedding of EMD over s-subsets of* $[n]^2$ *into* $\ell_{1-\frac{1}{\log s}}$ *with distortion* $O(\log s)$ *and dimension* $O(n^2)$.

**Proof:** Consider embedding

$$v = \bigoplus_{t=0}^{\log n} \frac{1}{2^t} \bigoplus_{k=0}^{2^t-1} (2^t)^{1-\epsilon} G_{kt},$$

where $G_{kt}$ denotes an embedding that is obtained by imposing grid with side length $2^t$ with origin at $(k,k)$ and counting number of points in each cell.

An easy modification of the proof of Theorem 2 gives that $v$ is an embedding of $EMD_\epsilon$ into $\ell_1$ with distortion $O(1/\epsilon)$. Furthermore, for $s$-subsets $A$ and $B$, every entry of $G_{kt}(A) - G_{kt}(B)$ is from set $\{-s, -s+1, ..., s\}$. For every entry $e$ from the set, $|e|^{1-\epsilon} = \Theta(|e|)$ when $\epsilon = 1/\log s$. This gives that for

$$v' = \bigoplus_{t=0}^{\log n} \frac{1}{2^{t/(1-\epsilon)}} \bigoplus_{k=0}^{2^t-1} 2^t G_{kt},$$

$\|v(A) - v(B)\|_1 = \Theta(\|v'(A) - v'(B)\|_{1-\epsilon}^{1-\epsilon})$. Therefore, $v'$ embeds $EMD_\epsilon$ into $\ell_{1-\epsilon}^{1-\epsilon}$ with distortion $O(\log s)$. Now, by Theorem 2, we get that $EMD^{1-\epsilon}(A,B)$ embeds into

$\ell_{1-\epsilon}^{1-\epsilon}$ with distortion $O(\log s)$.

It remains to argue that the dimension of the resulting embedding is $O(n^2)$. It can be seen that the dimension of $\bigoplus_{k=0}^{2^t-1}(2^t)^{1-\epsilon}G_{kt}$ is $O(\frac{n^2}{2^t})$. This follows because the dimension of $G_{kt}$ is $O\left(\frac{n^2}{2^{2t}}\right)$. Summing over all $t = 0, 1, 2, ..., \log n$, we obtain the required bound. ■

## 4.2 Deterministic dimensionality reduction

**Theorem 7** *Let $S$ be the set of all pointsets from $[n]^2$ having cardinality $s$. There exists a linear embedding of $EMD$ over $S$ into $\ell_1^d$ with $d = O(s^2 \log^2 n)$ and distortion $O(\log n)$.*

**Proof:** The standard embedding of [5, 6] achieves distortion $O(\log n)$ and dimension $O(n^3)$. In the embedding, $n$ shifts are enumerated to achieve distortion $O(\log n)$. We cannot afford factor of $n$ in the dimension. We show that the dimension can be reduced to that from the statement by sampling shifts less frequently. To achieve the same distortion, we weigh points in each cell (instead of counting) depending on how close they are to the border of the cell.

WLOG, we assume that $n$ is an integer power of 2. $A$ and $B$ will be pointsets of cardinality $s$.

The embedding is concatenation of $1+\log n$ embeddings $u_t$ ($t \in \{0, 1, 2, ..., \log n\}$): $u = \bigoplus_{t=0}^{\log n} u_t$. We set $S_{tk}$ to be transformation that shifts the pointset by vector $\left(2^t \cdot \frac{k}{C_1 C_2 s}, 2^t \cdot \frac{k}{C_1 C_2 s}\right)$. $C_1, C_2 > 0$ are two sufficiently large constants. Embedding $G_t$ is defined as follows. We impose a rectangular grid on the plane of cell size $2^t \times 2^t$ such that one of the vertices of the grid is $(0, 0)$. For every cell $C$ of the grid, there is a corresponding entry in $G_t$ equal to $(G_t)_C = \sum_{a \in C} w_{tC}(a)$, where

$$w_{tC}(a) = \begin{cases} d(a, C) \cdot \frac{C_1 s}{2^t} & \text{if } d(a, C) \leq \frac{2^t}{C_1 s} \\ 1 & \text{otherwise} \end{cases}.$$

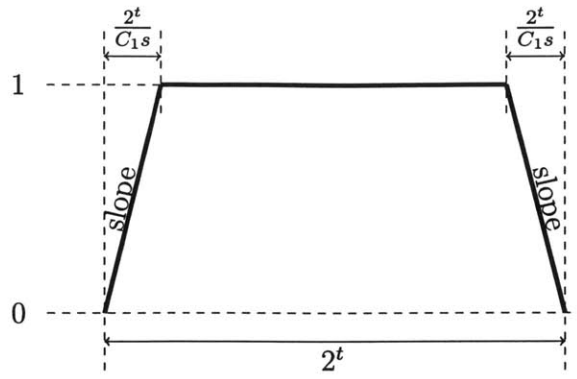


Figure 4-1: Profile view of weight function $w_{tC}$

We define $d(a, C)$ to be the distance for point $a$ to the border of cell $C$. $w_{tC}$ assigns weight 1 for points that are far from the border of cell $C$ and assigns linearly decreasing weight (in terms of the distance to the border of $C$) for points that are close to the border of $C$. The profile view of the weight function is in Figure 4-1.

Now let $u_t$ be

$$u_t = \frac{1}{sC_1C_2} \bigoplus_{k=0}^{sC_1C_2-1} 2^t \cdot w_t^k,$$

where $w_t^k = G_t S_{tk}$ and, i.e., we first shift the pointset, then impose a grid on the plane and sum weighted points in each cell according to $w_{tC}$.

We will show that $\Omega(EMD(A,B)) \leq \|u(A) - u(B)\|_1 \leq O(\log n) EMD(A,B)$.

First, we will show the upper bound. To do that, we show that for all $t$, $K_t := \|u_t(A) - u_t(B)\|_1 \leq O(EMD(A,B))$. Fix some $t$. Consider the matching corresponding to $EMD(A,B)$. Let $e$ be a length of any edge in the matching. Consider two cases. Suppose $e \leq 2^t/(C_1 s)$. Then the contribution of this edge to $K_t$ is at most

$$\begin{aligned} &\Pr_k[\text{a vertex of the edge lands on a slope of } w_t^k] \cdot O(eC_1 s) \\ &= \frac{1}{C_1 s} \cdot O(eC_1 s) = O(e) \end{aligned}$$

for sufficiently large constant $C_2$. Suppose $e > 2^t/(C_1 s)$. Then the contribution of the edge to $K_t$ is at most

$$\Pr_k[\text{the vertices of the edge lands lands in different cells}] \cdot 2^t \leq O(e)$$

for sufficiently large constant $C_2$. This proves the upper bound because the number of levels in $u$ is $O(\log n)$.

Now we prove the lower bound. Consider two pointsets $A$ and $B$ and let $T = EMD(A, B)$. WLOG, we assume that $T$ is an integer power of 2. Let $u'$ denote a concatenation of embeddings $u_t$ for all integer $t \in [0, U := \min(\log n, \log T)]$, i.e., $u' = \bigoplus_{t=0}^{U} u_t$. We will show that $\|u'(A) - u'(B)\|_1 \geq \Omega(EMD(A, B))$. This is sufficient to prove the lower bound (by setting $C_1$ to be a large enough constant) because embedding $u'$ contains a subset of embeddings $u_t$ of $u$.

Let $u''$ be a modification of $u'$, where we count the number of points in cells of $w_t^k$ as opposed to weighting (according to the definition of $u'$), i.e., we assign weight 1 for all points in the cell in $u''$ (use $(G_t'')_C = |\{a | a \in C\}|$ instead of $G_t$).

**Claim 1**

$$\|u'(A) - u'(B)\|_1 \geq \|u''(A) - u''(B)\|_1 - O(T/C_1).$$

**Proof:** By the definition of $u'$, for every $t$, we consider $sC_1C_2$ shifts and every point gets mapped on $C_2$ slopes. Thus, for at most $1/(C_1 s)$ fraction of shifts, a particular point lands on a slope. There are at most $2s$ points in sets $A$ and $B$ combined. Also, every point that lands on a slope, for a particular shift and a particular $u_t$, contributes to the difference $\|u''(A) - u''(B)\|_1 - \|u'(A) - u'(B)\|_1$ at most $2^t/(C_1C_2s)$. We conclude that the total contribution of all points in all $u_t$ $(0 \leq t \leq U)$ is $\sum_{t=0}^{U} 2^t \cdot 2/C_1 = O(T/C_1)$, which is what we wanted. ∎

Let $u'''$ be a modification of $u''$, where, instead of shifting the plane by integer multiples of $(2^t/(C_1C_2s), 2^t/(C_1C_2s))$ and applying the grid embedding for every shift, we shift the plane by $(k, k)$ for all integer $k \in [0, 2^t - 1]$ and weigh each shift so that

the total weight is 1, i.e.,

$$u''' = \bigoplus_{t=0}^{\log n} u_t''', \text{ where } u_t''' = \frac{1}{2^t} \bigoplus_{k=0}^{2^t-1} 2^t \cdot (w_t^k)'''$$

$$\text{and } (w_t^k)''' = G_t''' S_k''',$$

where $S_k'''$ denotes shifting pointset by $(k, k)$.

**Claim 2**

$$\|u''(A) - u''(B)\|_1 \geq \|u'''(A) - u'''(B)\|_1 - O(T/(C_1C_2)).$$

**Proof:** Consider a particular $t$. Notice that the set of shifts in $u''$ is a subset of shifts in $u'''$ and between two neighboring shifts of $u''$ there are $1/(C_1C_2s)$ fraction of shifts of $u'''$. Consider particular neighboring pair of shifts in $u''$. The only event that can contribute to $\|u'''(A) - u'''(B)\|_1 - \|u''(A) - u''(B)\|_1$ is when a point leaves a cell and enters another cell. There are at most 4 such events per point and there are at most $2s$ points for every $t$. Given that every event contributes at most $2^t/(C_1C_2s)$ to $\|u'''(A) - u'''(B)\|_1 - \|u''(A) - u''(B)\|_1$ for $t$, the total contribution is at most $\sum_{t=0}^{U} 8s \cdot \frac{2^t}{C_1C_2s} = O(T/(C_1C_2))$. ∎

The embedding $u'''$ is concatenation of embeddings $u_t'''$ for $t = 0, 1, 2, ..., \log n$. Notice that this embedding can be interpreted in the following way. We impose a randomly located grid with cell sizes $n \times n$ on the plane with integer coordinates. Then refine this grid into cell sizes $n/2 \times n/2$, then into cell sizes $n/4 \times n/4$, ..., then into cell sizes $1 \times 1$ and we count number of points in each cell for every grid and weigh these counts appropriately. Also, we average (concatenate with equal weights that sum upto 1) over all possible emplacements of the largest grid (of size $n \times n$) located at integer positions. Thus, $u'''$ can be interpreted as average over truncated grid embeddings, where by "truncated" we mean that we consider only levels with $t = 0, 1, 2, ..., \log n$ for every grid embedding. Clearly, if we do not truncate the grid embeddings, then $\|u'''(A) - u'''(B)\|_1 \geq \Omega(EMD(A, B))$ holds because the grid

embedding induces a matching. Next we show that his inequality holds even if we consider truncated grid embeddings.

**Claim 3**

$$\|u'''(A) - u'''(B)\|_1 \geq \Omega(EMD(A,B)).$$

**Proof:** The inequality holds because the grid embedding induces a matching and the following observations. If for some shift, the contribution from levels $t > \log T$ is non-zero, then there must be non-zero contribution from level $\log T$ but then the contribution from this level is already $2^{\log T} = \Omega(EMD(A,B))$ because, if there is unmatched point at level $\log T$, then the contribution from level $\log T$ is at least $2 \cdot 2^{\log T}$. Therefore, there is no need to consider levels $t > \log T$. The highest cost for matching two points is $2n$. Thus, we do not need to consider level $t = 1 + \log n$ because two points that are not matched at level $\log n$, induces cost $2^{1+\log n} = \Omega(2n)$.

The total contribution from levels $t < 0$ is equal to contribution from level $t = 0$. Let $c$ be contribution from level $t = 0$. Then the total contribution from levels $t < 0$ is $c/2 + c/4 + c/8 + ... = c$. Therefore, if we do not consider levels $t < 0$, the lower bound gets worse by at most a factor of 2. ∎

By the obtained inequalities for $u, u', u'', u'''$ and $EMD(A,B)$, we obtain the lower bound:

$$\begin{aligned}
&\|u(A) - u(B)\|_1 \\
&\geq \|u'(A) - u'(B)\|_1 \geq \|u''(A) - u''(B)\|_1 - O(T/C_1) \\
&\geq \|u'''(A) - u'''(B)\|_1 - O(T/(C_1C_2)) - O(T/C_1) \\
&\geq \Omega(T) - O(T/(C_1C_2)) - O(T/C_1) \\
&= \Omega(T) = \Omega(EMD(A,B)),
\end{aligned}$$

where the second inequality follows from Claim 1, the third inequality follows from Claim 2 and the last inequality follows from Claim 3.

Now it remains to reduce the dimension of the obtained embedding. Notice that for every $t$, $u_t(A) - u_t(B)$ is $O(s^2)$-sparse vector. It follows that $u(A) - u(B)$ is

$O(s^2 \log n)$-sparse vector because the number of levels is $O(\log n)$. Given that the dimension of $u$ is $O(n^2)$, by using Theorem 1, we obtain an embedding of dimension $O(s^2 \log^2 n)$ and distortion $O(\log n)$. ∎

## 4.3 Embedding of $EMD$ over sparse measures

**Definition 6** *Suppose $A$ and $B$ be two weighted pointsets with $|A| = |B| = s$ and corresponding weight functions $m^a$ and $m^b$. For $\epsilon > 0$ we define*

$$EMD_\epsilon(A,B) = \min_{\substack{m_{ij} \geq 0: \\ \forall i, m_i^a = \sum_j m_{ij} \\ \forall j, m_j^b = \sum_i m_{ij}}} \sum_{ij} m_{ij} \|a_i - b_j\|_1^{1-\epsilon}.$$

**Theorem 8** *Let $S$ be the set of all weighted pointsets from $[n]^2$ with total weight 1 and having cardinality $s$. There exists a linear embedding of $EMD$ over $S$ into $\ell_1^d$ with $d = O(s^5 \log^2 n)$ and distortion $O(\log n)$.*

**Proof:** We use similar embedding as in Theorem 7.

The embedding is concatenation of $1+\log(C_3 s^2(2n))$ embeddings $u_t$ ($t \in \{0, 1, 2, ..., \log(C_3 s^2(2n))\}$):

$$u = \bigoplus_{t=0}^{\log(C_3 s^2(2n))} u_t.$$

We define $u_t$ as

$$u_t = \frac{1}{s^2 C_1 C_2} \bigoplus_{k=0}^{s^2 C_1 C_2 - 1} 2^t \cdot w_t^k,$$

where $w_t^k = G_t S_{tk}$. $C_2$ is sufficiently large constant and $C_1 = C_4 s^2$, where $C_4$ is sufficiently large constant. We set $S_{tk}$ to be transformation that shifts the pointset by vector $\left(2^t \cdot \frac{k}{C_1 C_2 s^2}, 2^t \cdot \frac{k}{C_1 C_2 s^2}\right)$ and $G_t$ is defined as follows. We impose a rectangular grid on the plane of cell size $2^t \times 2^t$ such that one of the vertices of the grid is $(0,0)$. For every cell $C$ of the grid, there is a corresponding entry in $G_t$ equal to

$(G_t)_C = \sum_{a \in C} w_{tC}(a)$, where

$$w_{tC}(a) = \begin{cases} w_a \cdot d(a, C) \cdot \frac{C_1 s^2}{2^t} & \text{if } d(a, C) \leq \frac{2^t}{C_1 s^2} \\ w_a & \text{otherwise} \end{cases}.$$

We define $d(a, C)$ to be the distance for point $a$ to the border of cell $C$ and $w_a$ is the weight of point $a$. Notice that this weight function is almost the same as in Figure 4-1 except that this weight function has steeper slopes.

We will show that $\Omega(EMD(A, B)) \leq \|u(A) - u(B)\|_1 \leq O(\log n) EMD(A, B)$. The upper bound follows analogously as in Theorem 7. It remains to prove the lower bound.

Consider two weighted pointsets $A$ and $B$. Let $l_{ij}$ be the lengths in the optimal matching in $EMD(A, B)$ and $m_{ij}$ be the corresponding weights (as in Definition 6). Consider a certain pair of points at distance $l_{ij}$. Let $u_{ij}$ be a concatenation of embeddings $u_t$ for all integer $t \in [0, U_{ij} := \log(C_3 s^2 l_{ij})]$. Let $u'_{ij}$ (respectively, $u'$) be a modification of $u_{ij}$ (respectively, $u$), where, instead of assigning coefficients to weights in $u_t$, we sum the weights of points in a cell. That is, we use $(G'_t)_C = \sum_{a \in C} w_a$ instead of $G_t$. Also, instead of shifting the plane by integer multiples of some vector $(x, x)$ (for some $x$ as defined above), we shift the plane by integer multiples of $(1, 1)$ and assign equal weighs to shifts such that the total weight is 1. That is,

$$u' = \bigoplus_{t=0}^{\log(C_3 s^2 (2n))} \frac{1}{2^t} \bigoplus_{k=0}^{2^t - 1} 2^t G'_t S_k$$

and

$$u'_{ij} = \bigoplus_{t=0}^{U_{ij}} \frac{1}{2^t} \bigoplus_{k=0}^{2^t - 1} 2^t G'_t S_k,$$

where $S_k$ shifts the plane by vector $(k, k)$. We can interpret the embedding $u'_{ij}$ as the following two step process. We shift the plane by all vectors with equal integer coordinates and then concatenate truncated grid embeddings applied to each shift. We truncate grid embeddings by concatenating levels $t = 0, 1, 2, ..., U_{ij}$ instead for all

$t \in \mathbb{Z}$. When we concatenate the embeddings, we weigh these embeddings with equal weights and sum all entries that correspond to the same cell. We sum all entries corresponding to the same cell because the same cell appears in multiple shifts. One can see that for $1 - O_{C_4}\left(\frac{1}{C_3 s^2}\right)$ fraction of shifts in $u'_{ij}$, the vertices of the edge gets mapped to the same cell. By the union bound, we can show that for $1 - O_{C_4}(1/C_3)$ fraction of shifts of $u'$ (analogous interpretation as for $u'_{ij}$), for all edges the described property holds, i.e., for this fraction of shifts, vertices of every edge gets matched together at level with cell size $C_3 s^2 l_{ij}$. Now, similarly as in the proof of Theorem 7, we can show that

$$\begin{aligned}
&\|u(A) - u(B)\|_1 \\
&\geq (1 - O_{C_4}(1/C_3)) \|u'(A) - u'(B)\|_1 - \sum_{ij} O(2^{U_{ij}}/C_1) m_{ij} \\
&\geq \Omega(\|u'(A) - u'(B)\|_1) - \sum_{ij} O\left(\frac{C_3 s^2 l_{ij}}{C_1}\right) m_{ij} \\
&\geq \Omega(EMD(A,B)) - \sum_{ij} O\left(\frac{C_3 l_{ij}}{C_4}\right) m_{ij} \\
&\geq \Omega(EMD(A,B))
\end{aligned}$$

where we used $C_1 = C_4 s^2$ in the second to last inequality and $EMD(A,B) = \sum_{ij} m_{ij} l_{ij}$ in the last inequality. We get that $\|u(A) - u(B)\|_1 \geq \Omega(EMD(A,B))$ for large enough $C_4$.

We get that the resulting embedding has sparsity $O(s^5 \log n)$. Similarly as in the proof of Theorem 7, we obtain the required upper bound on the dimension. ∎

## 4.4 Embedding of $EMD$ with small distortion and small dimension

**Theorem 9** *Let $S$ be the set of all pointsets from $[n]^2$ having cardinality $s$. For any $0 < \epsilon < 1 - \Omega(1)$ there exists a linear embedding of $EMD_\epsilon$ over $S$ into $\ell_1^d$ with $d = O(s^2(\log s + \frac{1}{\epsilon}\log\frac{1}{\epsilon})\log n)$ and distortion $O(1/\epsilon)$.*

Think $\epsilon = 1/\log s$ (this is the value of $\epsilon$ that we will later use).

**Proof:** We modify the embedding from Theorem 7. We obtain the embedding by concatenating embeddings $w_t^k \cdot (2^t)^{1-\epsilon}$ instead of $w_t^k \cdot 2^t$. Let $u$ be the modified embedding (we do not use dimension reduction step from Theorem 7).

We will show that $\Omega(EMD_\epsilon(A,B)) \leq \|u(A) - u(B)\|_1 \leq O(1/\epsilon)EMD_\epsilon(A,B)$. First, we will construct an embedding such that the dimension is $O(s^2 \log^2 n)$ and then we will modify it again to obtain the claimed dimension.

Now we show the upper bound. Consider the matching corresponding to $EMD_\epsilon(A,B)$. Let $e$ be the length of an edge in the matching. Consider two cases. Suppose $e \leq 2^t/(C_1 s)$. Then the contribution of this edge to $K := \|u(A) - u(B)\|_1$ is at most

$$\Pr_k[\text{a vertex of the edge lands on a slope of } w_t^k] \cdot O(e) \cdot C_1 s \frac{(2^t)^{1-\epsilon}}{2^t}$$

$$\leq \frac{2}{C_1 s} \cdot O(e) \cdot C_1 s (2^t)^{-\epsilon} \leq O(e)(2^t)^{-\epsilon} \tag{4.1}$$

for sufficiently large constant $C_2$. Suppose $e > 2^t/(C_1 s)$. Then the contribution of the edge to $K$ is at most

$$\Pr_k[\text{the vertices of the edge lands lands in different cells}] 2^{t(1-\epsilon)}$$
$$\leq O(\min(e/2^t, 1)) \cdot 2^{t(1-\epsilon)}$$

for sufficiently large constant $C_2$.

Summing over all scales $t$, the contribution of the edge to $K$ is at most a constant multiple of

$$\sum_{t: \substack{t \geq 0, \\ e > 2^t}} (2^t)^{1-\epsilon} + \sum_{t: \substack{t \leq \log n, \\ e \leq 2^t}} e \cdot (2^t)^{-\epsilon} \leq O(e^{1-\epsilon} + e^{1-\epsilon}/\epsilon) = O(1/\epsilon) \cdot e^{1-\epsilon}.$$

By considering all edges, we obtain the upper bound.

Now we prove the lower bound. Consider two pointsets $A$ and $B$ and let $T = EMD_\epsilon(A, B)$. Let $u'$ denote a concatenation of embeddings $u_t$ for all integer

$$t \in [L := \max(0, \frac{1}{1-\epsilon}\log(\alpha T/s)), \qquad U := \min(\log n, \frac{1}{1-\epsilon}\log T)],$$

i.e., $u' = \bigoplus_{t=L}^{U} u_t$. WLOG, we assume that $U$ is an integer power of 2. $\|u'(A) - u'(B)\|_1 \geq (\Omega(1) - O(\alpha) - O(1/C_1))EMD_\epsilon(A, B)$ holds because the contribution of levels corresponding to $t < \frac{1}{1-\epsilon}\log(\alpha T/s)$ is at most $O(\alpha EMD_\epsilon(A, B))$ (Theorem 4) and analogous observations as in Theorem 7, except that in this case, instead of maximal contribution $2^t$ we have maximal contribution $(2^t)^{1-\epsilon}$ (thus, the factor $\frac{1}{1-\epsilon}$ in the expression of $L$). We choose $\alpha$ to sufficiently small and $C_1$ to be sufficiently large constant.

It remains to reduce the dimension of the obtained embedding. Notice that for every $t$, $u_t(A) - u_t(B)$ is $O(s^2)$-sparse vector. It follows that $u(A) - u(B)$ is $O(s^2 \log n)$-sparse vector. Given that the dimension of $u$ is $O(n^2)$, by using Proposition 1 and Theorem 1, we obtain an embedding of dimension $O(s^2 \log^2 n)$ and distortion $O(\log s)$.

Now we reduce the dimension to that from the statement of the statement. The main observation is that an "interesting" action is happening only in $N := O(\log s + \frac{1}{\epsilon}\log\frac{1}{\epsilon})$ *consecutive* scales (we show that below), i.e., the contribution to $\|u(A) - u(B)\|_1$ from other scales is at most $cEMD_\epsilon(A, B)$ for arbitrarily small constant $c$. This allows us to "reuse" entries from multiple scales. We reuse entries from scales modulo $N$, i.e.,

$$u = \bigoplus_{t=0}^{N-1} \sum_{m: m \equiv t \pmod N} \sum_{0 \leq m \leq \log n} u_t,$$

where $u_t = \frac{1}{sC_1C_2}\bigoplus_{k=0}^{sC_1C_2-1}(2^t)^{1-\epsilon}w_t^k$. As a result, no two "interesting" scales collide and, therefore, by triangle inequality, we still have the lower bound. The upper bound follows trivially. As a result, we replace $\log n$ factor with $N$, which is what we wanted.

Now we show that the contribution from all but $N$ scales is at most $cT$ for arbitrarily small constant $c$. We already know that the contribution from the scales with

$t < L$ is at most $O(\alpha T)$. Let

$$U' = \min\left(\log n, \frac{1}{1-\epsilon}\log T + \log(C_1 s) + \frac{1}{\epsilon}\log(1/(\beta\epsilon))\right). \tag{4.2}$$

We will show below that the contribution from scales with $t \geq U'$ is at most $O(\beta T)$. Noticing that $U' - L = O(\log s + \frac{1}{\epsilon}\log\frac{1}{\epsilon})$, we obtain the required bound on the dimension.

Consider an edge in the matching corresponding $EMD_\epsilon(A, B)$. Let $e$ be the length of the edge. Consider scale $t \geq U'$. From (4.2) we obtain that $e \leq \frac{2^t}{C_1 s}$. From (4.1) we get that the contribution of the edge is at most $O(e)(2^t)^{-\epsilon}$. The total contribution of the edge is at most

$$O(e)\sum_{t:t\geq U'}(2^t)^{-\epsilon} \leq O(\beta \cdot e^{1-\epsilon}),$$

which holds for $U' \geq \frac{1}{1-\epsilon}\log T + \frac{1}{\epsilon}\log(1/(\beta\epsilon))$. By summing over all edges in the matching, we obtain the required upper bound on the contribution. ∎

**Corollary 2** *Let $S$ be the set of all pointsets from $[n]^2$ having cardinality $s$. There exists a linear embedding of $EMD^{1-\epsilon}$ over $S$ into $\ell_1^d$ with $d = O(s^2(\log s)(\log\log s)\log n)$ and distortion $O(\log s)$, where $\epsilon = 1/\log s$.*

**Proof:** Set $\epsilon = 1/\log s$ in Theorem 9 and use Lemma 2. ∎

# Bibliography


[1] Alexandr Andoni, Piotr Indyk, and Robert Krauthgamer. Earth mover distance over high-dimensional spaces. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '08, pages 343–352, Philadelphia, PA, USA, 2008. Society for Industrial and Applied Mathematics.

[2] Patrice Assouad. Plongements lipschitziens dans $\mathbb{R}^n$. *Bull. Soc. Math. France,*, 111:429–448, 1983.

[3] R. Berinde, A. C. Gilbert, P. Indyk, H. Karloff, and M. J. Strauss. Combining geometry and combinatorics: A unified approach to sparse signal recovery. *ArXiv e-prints*, April 2008.

[4] Sergio Cabello, Panos Giannopoulos, Christian Knauer, and Günter Rote. Matching point sets with respect to the earth mover's distance. *Comput. Geom. Theory Appl.*, 39(2):118–133, February 2008.

[5] Moses S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing*, STOC '02, pages 380–388, New York, NY, USA, 2002. ACM.

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

[7] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, pages 604–613, New York, NY, USA, 1998. ACM.

[8] Subhash Khot and Assaf Naor. Nonembeddability theorems via fourier analysis. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '05, pages 101–112, Washington, DC, USA, 2005. IEEE Computer Society.

[9] Assaf Naor and Gideon Schechtman. Planar earthmover is not in l1. In *In 47th Symposium on Foundations of Computer Science (FOCS*, page 0509074, 2006.

[10] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth moverâĂŹs distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:2000, 2000.

[11] R. Sharathkumar. A sub-quadratic algorithm for bipartite matching of planar points with bounded integer coordinates. In *Proceedings of the Twenty-ninth Annual Symposium on Computational Geometry*, SoCG '13, pages 9–16, New York, NY, USA, 2013. ACM.