Efficient Sketches for Earth-Mover Distance, with Applications

Alexandr Andoni
MIT
andoni@mit.edu

Khanh Do Ba
MIT
doba@mit.edu

Piotr Indyk
MIT
indyk@mit.edu

David Woodruff
IBM Almaden
dpwoodru@us.ibm.com

Abstract— We provide the first sub-linear sketching algorithm for estimating the planar Earth-Mover Distance with a constant approximation. For sets living in the two-dimensional grid $[\Delta]^2$, we achieve space Δ^ϵ for approximation $O(1/\epsilon)$, for any desired $0 < \epsilon < 1$. Our sketch has immediate applications to the streaming and nearest neighbor search problems.

1. Introduction

For any two multisets A, B of points in $\mathbb{R}^2$, $|A| = |B| = N$, the (planar) *Earth-Mover Distance*[1] between A and B is defined as the minimum cost of a perfect matching with edges between A and B, i.e.,

$$EMD(A,B) = \min_{\pi:A\to B} \sum_{a\in A} \|a - \pi(a)\|$$

where π ranges over all one-to-one mappings. 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 [19], [25], [2], [3], [5], [17], [1], [15].

Recently, there has been a significant interest in developing methods for *geometric* representation of EMD. The goal of this line of research is to provide 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 can be approximated from the vectors $f(A)$ and $f(B)$. To be useful, the space that f maps into must be "simple", e.g., its dimension k must be low, or its distance estimation function should be of simple form. For example, [5], [17] provide a mapping f that works when the sets A, B are subsets of the discrete square grid $[\Delta]^2$, and guarantees that, for some absolute constant $C > 0$, we have

$$\|f(A)-f(B)\|_1 \le EMD(A,B) \le C\log\Delta \cdot \|f(A)-f(B)\|_1.$$

Geometric representations of EMD have found applications in several areas, including:

- *Data streaming computation*: The mapping of EMD into ℓ_1, combined with ℓ_1-distance-preserving mappings into low-dimensions [14], yields an efficient algorithm for estimating the EMD between a set of points given in a streaming fashion [13]. Specifically, the algorithm provides an $O(\log\Delta)$ approximation in one pass over the data, using only $\log^{O(1)}(\Delta N)$ space for sets of size at most N. Obtaining a better EMD estimation algorithm has been an important open problem in the streaming literature [21].
- *Visual search and recognition*: The aforementioned embedding, together with efficient nearest neighbor search methods, has been applied to fast image search in large collections of images [17]. Kernel variants of that embedding, such as *pyramid kernels* [10] and *spatial pyramid kernels* [20], are some of the best known practical methods for image recognition in large data sets [20].

However, representing EMD as vectors in the ℓ_1 space has limitations: it has been shown [23] that any such mapping must incur a distortion of at least $\Omega(\sqrt{\log\Delta})$. Thus, in order to obtain more accurate representations, one must consider mappings into spaces other than ℓ_1.

In this paper, we provide a construction of such mappings. Their key feature is that they map the sets into spaces of dimension that is significantly *sub-linear* in Δ. For a multiset $A \subseteq [\Delta]^2$, let $x(A) \in \mathbb{R}^{\Delta^2}$ be the characteristic vector of A. Our main result is:

Theorem 1.1. *For any $0 < \epsilon < 1$, there is a distribution over linear mappings $F : \mathbb{R}^{\Delta^2} \to \mathbb{R}^{O(\Delta^\epsilon)}$ as well as an* estimator *function E such that for any two multisets $A, B \subseteq [\Delta]^2$ of equal size, we have*

$$EMD(A,B) \le E(F\cdot x(A), F\cdot x(B)) = O(1/\epsilon)\cdot EMD(A,B)$$

with probability $2/3$. Moreover, the entries in the matrix defining F are integers in the range $\{-\Delta^{O(1)}, \dots \Delta^{O(1)}\}$.

The estimation function $E(\cdot,\cdot)$ can be evaluated in time $(\log\Delta)^{O(1)}$. However, $E(\cdot,\cdot)$ is *not* a metric distance function. Instead, it involves operations such as median, and as a result it does not satisfy triangle inequality.

Applications. Theorem 1.1 almost immediately provides an improved algorithm for streaming and nearest neighbor search problems. In the streaming model (cf. [22], [16]), consider the aforementioned problem of computing the EMD between the sets A and B of points given in a stream. It can be seen that, due to the linearity of F, the "sketch" vectors

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

[1]Variants of this notion are also known as the *transportation distance* or *bi-chromatic matching* distance.

$Fx(A)$ and $Fx(B)$ can be maintained under insertions of points to A and B (as well as deletions of points from A and B). Moreover, as per [14], the random bits defining a linear mapping F can be generated using a pseudo-random generator for bounded space [24] that requires generating and storing only $\Delta^\epsilon \log^{O(1)}(\Delta N)$ truly random bits. Finally, for any multi-set B of size at most N, each coordinate of $Fx(B)$ is in the range $\left\{-(\Delta N)^{O(1)}, \ldots (\Delta N)^{O(1)}\right\}$ and can be stored using $O(\log(\Delta N))$ bits. We obtain the following theorem.

Theorem 1.2. *For any $0 < \epsilon < 1$, there is a one-pass streaming algorithm that maintains an $O(1/\epsilon)$-approximation of the value of EMD between point-sets from $[\Delta]^2$ given in a stream of length N, using $\Delta^\epsilon \log^{O(1)}(\Delta N)$ space.*

Another application of Theorem 1.1 is to give an improved data structure for the approximate nearest neighbor problem under EMD. Specifically, consider a set S consisting of s sets $A_i \subseteq [\Delta]^2$, each of size at most N. By increasing the dimension of the mapping F by a factor of $O(\log s)$ we can ensure that, for any fixed set B, one can estimate the distance between B and *all* sets in S up to a factor of $O(1/\epsilon)$ with probability $2/3$. We build a lookup table that, for each value of $Fx(B)$, stores the index i that minimizes the value of the estimated distance $E(Fx(A_i), Fx(B))$. From the properties of the mapping F, we obtain the following theorem.

Theorem 1.3. *For any $0 < \epsilon < 1$, there is a data structure that, given a "query" multi-set B, reports a $O(1/\epsilon)$-approximate nearest neighbor of B in S with probability at least $2/3$. The data structure uses $2^{\Delta^\epsilon \log(s\Delta N)^{O(1)}}$ space and $(\Delta \log(s\Delta N))^{O(1)}$ query time.*

Thus, we obtain a data structure with very fast query time and space sub-exponential in the dimension Δ^2 of the underlying EMD space. This improves over the result of [4], who obtained an algorithm with a similar space bound while having super-constant approximation guarantee with query time polynomial in the number of data points s.

Techniques. Our mapping utilizes two components: one old, and one new. The first component, introduced in [15], provides a decomposition of EMD over $[\Delta]^2$ into a convex combination of closely related metrics, called EEMD, defined over $[\Delta^\epsilon]^2$. Specifically, consider an extension of EMD to any (multi-)sets $A, B \subseteq [\Delta]^2$ (not necessarily of the same size), defined as:

$$EEMD_\Delta(A,B) = \min_{\substack{S \subseteq A, S' \subseteq B \\ |S| = |S'|}} \left[EMD(S,S') + \Delta(|A - S| + |B - S'|)\right]$$

(we often skip the subscript Δ when it is clear from the context). It is known that the EEMD metric can be induced by a norm $\|x\|_{\mathrm{EEMD}}$, such that for any multi-sets A, B we have $EEMD(A,B) = \|x(A) - x(B)\|_{\mathrm{EEMD}}$ (see Section 2 for the definition). The decomposition from [15] can be now stated as follows (after adapting the notation to the setup in this paper):

Fact 1.4 ([15])**.** *For any $0 < \epsilon < 1$, there exists a distribution over n-tuples of linear mappings $\langle F_1, \ldots, F_n \rangle$, for $F_i : \mathbb{R}^{\Delta^2} \to \mathbb{R}^{m^2}$ with $m = \Delta^\epsilon$, such that for any $x \in \mathbb{R}^{\Delta^2}$, we have*

- $\|x\|_{EEMD} \leq \sum_i \|F_i(x)\|_{EEMD}$ *with probability 1, and*
- $\mathbb{E}\left[\sum_i \|F_i(x)\|_{EEMD}\right] \leq O(1/\epsilon) \cdot \|x\|_{EEMD}$.

Furthermore, $n = \Delta^{O(1)}$.

It suffices to estimate the sum of the terms $\|F_i(x)\|_{\mathrm{EEMD}}$ in the decomposition. The second component needed for our result (and the main technical development of this paper) is showing that the sum estimation can be accomplished by using a proper linear mapping. In fact, the method works for estimating the sum $\sum_i \|x_i\|_X$ for a vector $x = (x_1, \ldots, x_n) \in X^n$ for *any* normed space $X = (\mathbb{R}^m, \|\cdot\|_X)$. We denote $\|x\|_{1,X} = \sum_{i \in [n]} \|x_i\|_X$. This component is formalized in the following theorem.

Theorem 1.5 (Linear sketching of a sum of norms)**.** *Fix $n \in \mathbb{N}$, a threshold $M > 0$, and approximation $\gamma > 1$. For $k = (\gamma \log n)^{O(1)}$, there exists a distribution over random linear mappings $\mu : X^n \to X^k$, and a reconstruction algorithm $\mathcal{R}$, such that for any $x \in X^n$ satisfying $M/\gamma \leq \|x\|_{1,X} \leq M$, the algorithm $\mathcal{R}$ produces an $O(1)$-approximation to $\|x\|_{1,X}$ from $\mu(x)$, with high probability.*

Theorem 1.5 immediately implies Theorem 1.1, since we can use the mapping from [5], [17] to obtain an estimation M of $\|x\|_{1,\mathrm{EEMD}}$ with an approximation factor $\gamma = O(\log \Delta)$. For completeness, we include its proof in Section 4.

The main idea behind the construction of the mapping is as follows. First, observe that a natural approach to the sum estimation problem would be to randomly sample a few blocks x_i of the vector x. This does not work, however: the mass of the sum could be concentrated in only a single block, and a random sample would likely miss it. An alternative approach, used in the *off-line* algorithm of [15], is to sample each block x_i with probability approximately proportional to $\|x_i\|_{\mathrm{EEMD}}$. However, this requires existence of a streaming algorithm that supports such sampling. A recent paper of [18] is a step towards achieving such an algorithm. However, it applies to the case [2] where one samples just individual coordinates, while we need to sample and retrieve *blocks*, in order to compute the value of EMD on them directly. Although the two problems are related in principle (having enough samples of block coordinates could

[2]There are other technical obstacles such as that their algorithm samples with probability proportional to $|x_i|^p$ for $p > 2$, while here we would need the sampling probability to be proportional to the norm of x_i, i.e., $p = 1$. However, these issues are not much of a problem.

provide some information about the norm of the block itself), the tasks seem technically different. Indeed, the recovery procedure forms the main technical part of the paper, even though the final algorithm is quite simple.

2. Preliminaries

We start by defining the $\|\cdot\|_{\text{EEMD}}$ norm. For any $x \in \mathbb{Z}^{n^2}$, let $x^+ = (|x| + x)/2$ be the vector containing only the positive entries in x, and let $x^- = x - x^+$. Then define $\|x\|_{\text{EEMD}} = EEMD(x^+, x^-)$, where we identify $x^+ \in \mathbb{N}^{n^2}$ with the multi-set for which x^+ is the indicator vector (and similarly with x^-). The norm naturally extends to entire $x \in \mathbb{R}^{n^2}$ by corresponding weighting; we omit these details as they are not important in this paper. Observe that for any multi-sets A, B we have $EEMD(A, B) = \|x(A) - x(B)\|_{\text{EEMD}}$.

We consider all logs to be in base 2. The notation $\chi[E]$ stands for 1 if event/expression E is true and 0 otherwise.

3. Proof of Theorem 1.5

We first present the construction of the sketching function μ and of the reconstruction algorithm $\mathcal{R}$. The respective algorithms are presented in Figures 1 and 2. We then prove the correctness guarantee, namely that the reconstruction algorithm $\mathcal{R}$ approximates well the norm $\|x\|_{1,X}$.

3.1. Sketch and reconstruction algorithms

We start by giving some intuition behind our constructions.

Fix an input $x \in X^n$. We will refer to x_i's as the *elements* of x. As in [12] and several further papers, the idea is to partition these elements into exponential levels, depending on their X-norm. Specifically, for a level $j \in \mathbb{N}$, we set the threshold $T_j = M/2^j$ and define the level j to be the set

$$S_j = \{ i \in [n] \mid \|x_i\|_X \in (T_j, 2T_j] \}.$$

Let $s_j = |S_j|$ be the size of S_j. We will observe that $\|x\|_{1,X}$ is approximated by $\sum_{j\geq 1} T_j \cdot s_j$. Furthermore, it is sufficient to consider only levels $j \leq \ell := \log(4n\gamma)$.

The main challenge is to estimate each s_j for each $j \in [\ell]$. We will do so for each j separately. We will subsample the elements from $[n]$ such that, with "good probability", we subsample exactly one element from S_j, say $i \in S_j$, and no element from $S_{j'}$ for $j' < j$. We refer to this event as E. In a sense, we "isolate" (at most) one element $i \in S_j$, while the rest of the remaining elements are from "lighter" levels and thus have a much smaller weight than i. Such an isolation allows us to efficiently verify the fact we have succeded to isolate one i.

The probability that we manage to isolate exactly one such $i \in S_j$ (E holds) is in fact roughly proportional to the size of the set S_j. Thus it suffices to just estimate the probability that the event E holds. To ensure the "rough proportionality" we subsample the elements at a rate for

For each $j \in [\ell]$, create $t = 4\gamma\ell^2 \log n$ hash tables, denoted $H^{(j,u)}$ for $u \in [t]$, each with $w = 640\gamma\ell^2 \log^2 n$ cells, and assign to them independent hash functions $h_{j,u} : [n] \to [w]$
For each hash table $H^{(j,u)}$
    Subsample a set $I_{j,u} \subset [n]$ where each $i \in [n]$ is included independently with probability $p_j = 2^{-j}/(40\ell)$
    For each $v \in [w]$
        $H_v^{(j,u)} := \sum_{i\in[n]} \chi[i \in I_{j,u}] \cdot \chi[h_{j,u}(i) = v] \cdot x_i$

Algorithm 1: Construction of the sketch μ.

which E holds with a probability that is inversely poly-logarithmic, $\log^{-\Theta(1)} n$. Thus we repeat the subsampling experiment for $t = (\gamma \log n)^{O(1)}$ times and count in how many experiments the event E holds; this count gives an estimate for s_j (when appropriately scaled).

The following core problem remains: for each subsampling experiment $u \in [t]$, we need to actually verify that E holds in this experiment, i.e., whether exactly one element of S_j is subsampled and no element from $S_{j'}$ for $j' < j$. To do so, we hash the subsampled elements, denoted $I_{j,u}$, into a hash table. Then, E holds roughly when there is exactly one cell that has norm in the right range (roughly $(T_j, 2T_j]$), and all the other cells have small norm. Ideally, if the hash table were huge, then the subsampled elements, $I_{j,u}$, do not collide in the hash table and then the verification procedure is accurate. Since the hash table size is much smaller, of only poly-logarithmic size, this verification procedure may fail. Specifically, the verification procedure fails when either the elements from the "lighter" level-sets $S_{j'}$ for $j' > j$ contribute a lot to one of the cells, or some elements from "heavier" level-sets $S_{j'}$ for $j' < j$ are subsampled and collide. If we set the size w of the hash table sufficiently high, we will ensure that neither of these two bad events happens with a significant probability.[3]

The detailed algorithm for the sketch μ is presented in Figure 1. As defined in the preliminaries, $\chi[E]$ stands for 1 if expression E is true and 0 otherwise. Note that the constructed sketch μ is linear.

Before giving the reconstruction algorithm $\mathcal{R}$, we need the following definition, which describes our procedure of verifying that the event E from the above discussion holds.

Definition 3.1. *For $j \in [\ell]$, $u \in [t]$, call the pair (j, u) an* accepting pair *if the following holds:*

- *there is exactly one position $v \in [w]$ such that $\|H_v^{(j,u)}\|_X \in (0.9 \cdot T_j, 2.1 \cdot T_j]$, and*
- *for all other $v' \in [w]$, $\|H_{v'}^{(j,u)}\|_X \leq 0.9 \cdot T_j$.*

[3]Similar phenomenon has been exploited for the sparse approximation and heavy hitter problems, see [9], [6], [8], [7].

1 For each $j \in [\ell]$, let c_j be the number of accepting pairs (j, u) for $u \in [t]$
2 Return $\mathcal{E} = \sum_{j \in [\ell]} T_j \cdot \frac{c_j}{t} \cdot \frac{1}{p_j}$

Algorithm 2: Reconstruction algorithm $\mathcal{R}$.

The resulting reconstruction algorithm is given in Figure 2.

3.2. Proof of correctness

First we observe that the norm $\|x\|_{1,X}$ is approximated by $\sum_{j \in [\ell]} T_j \cdot |S_j|$ up to a factor of 4. Indeed, $\|x\|_{1,X}$ is 2-approximated by the same sum with unrestricted j, i.e., $\sum_{j \geq 1} T_j \cdot |S_j|$. Moreover, every element $i \in [n]$ from a higher level $j > \ell$ contributes a norm that is at most

$$\|x_i\|_X \leq \frac{M}{2^\ell} = \frac{1}{4n} \cdot \frac{M}{\gamma} \leq \frac{1}{4n} \|x\|_{1,X}.$$

Thus the elements from the ignored levels constitute at most a $1/4$-fraction of $\|x\|_{1,X}$.

We set $s_j = |S_j|$ to be the size of S_j. By notational convention, we also assume that for $j < 1$, we have $S_j = \emptyset$ and $s_j = 0$. Also, we can assume that $\gamma \leq n^c$ for some absolute constant $c > 0$, since, otherwise, the construction with $k = \gamma^{1/c}$ is trivial.

We define $\tilde{s}_j = \frac{c_j}{t} \cdot \frac{1}{p_j}$, which one should think of as our estimate of s_j. Then the reconstruction algorithm returns the estimate $\mathcal{E} = \sum_{j \in [\ell]} T_j \cdot \tilde{s}_j$ of the norm $\|x\|_{1,X}$.

Our main challenge is to prove that $\tilde{s}_j$ is a good estimate of s_j for each $j \in [\ell]$. While we can prove a good upper bound on $\tilde{s}_j$ for all $j \in [\ell]$, we cannot prove a good lower bound on all $\tilde{s}_j$'s. Namely, if s_j is very small, we cannot lower-bound $\tilde{s}_j$ (as we do not have enough subsampling experiments). But in this case, the level j contributes a negligible mass to the norm $\|x\|_{1,X}$, and thus it can simply be ignored.

To formalize the above point, we partition the levels j into two types — important and non-important levels — depending on the number s_j of elements in the corresponding level. Intuitively, the non-important levels are those that contribute a negligible amount of mass to the norm $\|x\|_{1,X}$.

Definition 3.2. *Call level $j \in [\ell]$* important *if $s_j \geq \frac{M/\gamma}{T_j} \cdot \frac{1}{8\ell} = \frac{2^j}{8\gamma\ell}$. Let $\mathcal{J} \subseteq [\ell]$ denote the set of important levels.*

The following two lemmas prove, respectively, lower and upper bound on our estimates $\tilde{s}_j$.

Lemma 3.3. *For every important level $j \in \mathcal{J}$, with high probability,*

$$\tilde{s}_j \geq s_j/8.$$

Lemma 3.4. *For every level $j \in [\ell]$, with high probability,*

$$\tilde{s}_j \leq 2\left(s_{j-1} + s_j + s_{j+1} + \frac{2^j}{8\gamma\ell}\right).$$

First, we show how the two lemmas imply Theorem 1.5.

Proof of Theorem 1.5: We have already observed that $\sum_{j \in [\ell]} T_j \cdot s_j$ approximates $\|x\|_{1,X}$ up to a factor of 4. Thus, by Lemma 3.4, we have

$$\begin{aligned}\mathcal{E} = \sum_{j \in [\ell]} T_j \cdot \tilde{s}_j &\leq O(1) \sum_{j \in [\ell]} T_j \cdot \left(s_{j-1} + s_j + s_{j+1} + \frac{2^j}{8\gamma\ell}\right) \\ &\leq O(1) \sum_{j \in [\ell]} T_j \cdot s_j + O(\ell) \cdot \frac{M}{8\gamma\ell} \leq O(1) \cdot \|x\|_{1,X},\end{aligned}$$

where we have used the fact that $\|x\|_{1,X} \geq M/\gamma$.

On the other hand, we can lower bound $\mathcal{E}$ by dropping all the non-important levels j. By Lemma 3.3, we have

$$\mathcal{E} \geq \sum_{j \in \mathcal{J}} T_j \cdot \tilde{s}_j \geq \Omega(1) \sum_{j \in \mathcal{J}} T_j \cdot s_j.$$

The contribution of the non-important levels is, by the definition of importance,

$$\sum_{j \in [\ell] \setminus \mathcal{J}} T_j \cdot s_j < \ell \cdot \frac{M/\gamma}{8\ell} \leq \frac{1}{8} \|x\|_{1,X}.$$

Thus, we conclude

$$\begin{aligned}\sum_{j \in \mathcal{J}} T_j \cdot s_j &= \sum_{j \in [\ell]} T_j \cdot s_j - \sum_{j \in [\ell] \setminus \mathcal{J}} T_j \cdot s_j \geq \frac{1}{4} \|x\|_{1,X} - \frac{1}{8} \|x\|_{1,X} \\ &= \Omega(1) \cdot \|x\|_{1,X},\end{aligned}$$

which completes the proof of Theorem 1.5. ■

3.2.1. Proofs of Lemmas 3.3 and 3.4: As mentioned before, at a given level j, we are trying to estimate the size s_j of the set S_j. We do so by subsampling the elements t times, each at a rate of roughly $1/|S_j|$, and counting how many times the subsampling produced exactly one element from S_j (and there will be a negligible probability that more than one element is subsampled). The hope is that the pair (j, u) is accepting iff the event E holds, that is, the subsample $I_{j,u}$ contains only one element from S_j and none from $S_{j'}$ for $j' < j$. The main difficulty turns out to be bounding the contribution of the elements from the sets $S_{j'}$ for $j' \geq j+2$: the sets $S_{j'}$ may be much larger than S_j and thus a fraction of them is likely to be present in the subsample. Nonetheless, the elements from these sets $S_{j'}$ are small in norm and thus are distributed nearly uniformly in the hash table $H^{(j,u)}$.

To formalize this intuition, we will prove the *Noise Lemma* that quantifies the "noise" (norm mass) contributed by the elements from the sets $S_{j'}$, for $j' \geq j + 2$, in a hash table $H^{(j,u)}$. This Noise Lemma will be used for both Lemma 3.3 and Lemma 3.4.

The Noise Lemma has two parts. The first part gives a tight bound on the noise in a given cell of the hash table $H^{(j,u)}$, but the probability guarantee is for a given cell only. The second part gives a somewhat weaker bound on the noise, but holds for all the cells of $H^{(j,u)}$ simultaneously.

Lemma 3.5 (Noise Lemma)**.** *Fix some $j \in [\ell]$ and $u \in [t]$. Consider some cell v of the hash table $H^{(j,u)}$, and let $S_{\geq j+2} = \bigcup_{j' \geq j+2} S_{j'}$. Then*

$$\sum_{i \in S_{\geq j+2}} \chi[i \in I_{j,u}] \cdot \chi[h_{j,u}(i) = v] \cdot \|x_i\|_X \leq 0.1 \cdot T_j \quad (1)$$

with probability at least $1 - \frac{1}{2w}$.

Furthermore, with probability at least $1 - \frac{\log^2 n}{w}$, we have

$$\max_{v' \in [w]} \sum_{i \in S_{\geq j+2}} \chi[i \in I_{j,u}] \cdot \chi[h_{j,u}(i) = v'] \cdot \|x_i\|_X \leq 0.6 \cdot T_j. \quad (2)$$

Proof: We begin by proving Eqn. (1). We have by the linearity of expectation that:

$$\mathbb{E}\left[\sum_{i \in S_{\geq j+2}} \chi[i \in I_{j,u}] \cdot \chi[h_{j,u}(i) = v] \cdot \|x_i\|_X\right] \leq \frac{p_j}{w} \cdot \|x\|_{1,X} \leq \frac{M/2^j}{40\ell w} \leq \frac{T_j}{40w}. \quad (3)$$

Using standard Markov's bound, we conclude that Eqn. (1) holds with probability at least $1 - \frac{1}{2w}$, completing the first part of the Noise Lemma.

We now prove the second part, Eqn. (2). Note that we cannot hope to prove that *all* cells will have noise at most $0.1 \cdot T_j$, because even just one element from a set S_{j+2} can contribute as much as $T_j/2$. To prove this part, we partition the elements in $S_{\geq j+2}$ into two types: heavy elements (of mass close to T_j) and light elements (of mass much smaller than T_j). For heavy elements, we will prove that we subsample only a few of them, and thus they are unlikely to collide in the hash table. The light elements are so light that they can be upper-bounded using a tight concentration bound.

Specifically, we define the following sets of light and heavy elements:

$$S_l := \bigcup_{j' \geq j + \log\log n + 1} S_{j'}$$
$$S_h := S_{\geq j+2} \setminus S_l = \bigcup_{j+2 \leq j' < j+\log\log n+1} S_{j'}$$

We first show that the light elements do not contribute more than $(0.1)T_j$ to *any* cell w.h.p. Namely, we will bound the noise in a cell $v' \in [w]$ using a Hoeffding bound, and then use a union bound over all v'. We use the following variant of the Hoeffding inequality, which can be deduced from [11].

Lemma 3.6 (Hoeffding bound)**.** *Let Z_i be n independent random variables such that $Z_i \in [0, B]$, for $B > 0$, and $\mathbb{E}[\sum_i Z_i] = \mu$. Then, for any $a > 0$, we have that*

$$\Pr\left[\sum_i Z_i > a\right] \leq e^{-(a-2\mu)/B}.$$

We use the lemma for variables $Z_i = \chi[i \in I_{j,u}] \cdot \chi[h_{j,u}(i) = v'] \cdot \|x_i\|_X$, where $i \in S_l$. To get a bound on B, we observe that, for $i \in S_\ell$, we have $\|x_i\|_X \leq T_{j+\log\log n} = T_j/2^{\log\log n} = T_j/\log n$. We also have an upper bound of $\mu = \mathbb{E}[\sum_{i \in S_l} Z_i] \leq T_j/(40w)$ (from Eqn. (3)). Thus, applying Lemma 3.6, we obtain

$$\Pr\left[\sum_{i \in S_l} \chi[i \in I_{j,u}] \cdot \chi[h_{j,u}(i) = v'] \cdot \|x_i\|_X > 0.1 \cdot T_j\right] \leq e^{-(0.1-1/(20w))T_j/(T_j/\log n)} < e^{-0.05\log n} = n^{-\Omega(1)}.$$

Taking the union bound over all cells, we obtain the same bound on all cells $v' \in [w]$.

We now analyze the behavior of the heavy elements, i.e., elements from the set S_h. We can bound the expected number of subsampled heavy elements as follows:

$$\mathbb{E}\left[\sum_{i \in S_h} \chi[i \in I_{j,u}]\right] \leq \left(\sum_{j'=j+2}^{j+\log\log n} 2^{j'}\right) \cdot p_j < 2^{j+\log\log n+1} \cdot \frac{2^{-j}}{40\ell} = \frac{\log n}{20\ell} \leq O(1).$$

Applying the Hoeffding bound from above, we obtain

$$\Pr\left[\sum_{i \in S_h} \chi[i \in I_{j,u}] > \log n\right] \leq e^{-\Omega(\log n)} = n^{-\Omega(1)}.$$

Thus, no more than $\log n$ heavy elements are subsampled, w.h.p. We can further bound the probability that any two of them hash into the same cell by $\Pr[\text{there exists a collision of heavy elements}] \leq \binom{\log n}{2}/w \leq \frac{\log^2 n}{2w}$.

To conclude, for every cell v', the light elements can contribute at most $0.1 \cdot T_j$, and the heavy elements can contribute at most $T_j/2$. The lemma then follows. ∎

We are now ready to prove Lemmas 3.3 and 3.4. We will let $A_{j,u}$ represent the event that (j, u) is an accepting pair, as per Definition 3.1.

Proof of Lemma 3.3: Fix an important $j \in \mathcal{J}$, and some $u \in [t]$. Define the following two events:

$E1$: exactly one element of S_j is subsampled in $I_{j,u}$, and

$E2$: no element from $S_{j'}$ is subsampled in $I_{j,u}$, for all $j' < j$ and $j' = j + 1$.

We will prove the following claim.

Claim 3.7. *For fixed $u \in [t]$, if $E1$ and $E2$ hold, then $A_{j,u}$ occurs with probability at least 1/2. Moreover, $E1$ and $E2$ occur simultaneously with probability at least $\frac{1}{2}s_j p_j$.*

Proof of Claim 3.7: To prove the first part, assume $E1$ and $E2$ hold. Let i^* be the element in $I_{j,u} \cap S_j$ (guaranteed to be unique by $E1$), and let v^* be the cell that contains element i^*. First, we note that, using the triangle inequality in X and the Noise Lemma 3.5, we have

$$\|H^{(j,u)}_{v^*}\|_X \geq \|x_{i^*}\|_X - \sum_{i\in I_{j,u}\setminus\{i^*\}} \chi[h_{j,u}(i) = v^*] \cdot \|x_i\|_X$$
$$> T_j - 0.1 \cdot T_j = 0.9 \cdot T_j,$$

and

$$\|H^{(j,u)}_{v^*}\|_X \leq \|x_{i^*}\|_X + \sum_{i\in I_{j,u}\setminus\{i^*\}} \chi[h_{j,u}(i) = v^*] \cdot \|x_i\|_X$$
$$\leq 2.1 \cdot T_j,$$

with probability at least $3/4$. Furthermore, for every other cell $v \neq v^*$, we have that, similarly to above:

$$\max_{v\neq v^*} \|H^{(j,u)}_v\|_X \leq \max_{v\neq v^*} \sum_{i\in I_{j,u}} \chi[h_{j,u}(i) = v] \cdot \|x_i\|_X \leq 0.6 \cdot T_j$$

with probability at least $3/4$. The two bounds hold at the same time with probability at least 1/2, in which case $A_{j,u}$ occurs.

Next we show that $E1$ and $E2$ hold with probability at least $\frac{1}{2}s_jp_j$. We have

$$\Pr[E1] = s_jp_j(1-p_j)^{s_j-1} \geq s_jp_j(1-s_jp_j) \geq \tfrac{2}{3}s_jp_j,$$

where we use the fact that $s_j \leq 2^j = \frac{1}{40\ell p_j}$. To estimate $\Pr[E2]$, we first consider all $j' < j$. Using the union bound, we can bound the probability that anything from $\bigcup_{j'<j} S_{j'}$ is subsampled:

$$\Pr\left[\cup_{j'<j} S_{j'} \bigcap I_{j,u} \neq \emptyset\right] \leq \sum_{j'<j} s_{j'}p_j \leq \sum_{j'<j} 2^{j'} p_j < 2^j p_j$$
$$= \frac{1}{40\ell}.$$

Similarly, we have

$$\Pr[S_{j+1} \cap I_{j,u} \neq \emptyset] \leq s_{j+1}p_j \leq \frac{1}{20\ell}.$$

Thus we obtain that $\Pr[E2] \geq 1 - \frac{1}{10\ell}$.

We note that $E1$ and $E2$ are independent events since they concern different levels. We can conclude that

$$\Pr[E1 \wedge E2] = \Pr[E1] \cdot \Pr[E2] \geq \tfrac{2}{3}s_jp_j \cdot \left(1 - \frac{1}{10\ell}\right) \geq \tfrac{1}{2}s_jp_j,$$

which finishes the proof of Claim 3.7. ■

We now complete the proof of Lemma 3.3. We can lower bound the probability of $A_{j,u}$ as follows:

$$\Pr[A_{j,u}] \geq \Pr[A_{j,u} \wedge E1 \wedge E2]$$
$$= \Pr[A_{j,u} \mid E1 \wedge E2] \cdot \Pr[E1 \wedge E2] \geq \tfrac{1}{4}s_jp_j.$$

Now, we can finally analyze the estimate $\tilde{s}_j$ of the size of the set S_j. Since $\tilde{s}_j = \frac{c_j}{t} \cdot \frac{1}{p_j}$, we will lower bound c_j. Note that

$$\mathbb{E}[c_j] = t\Pr[A_{j,u}] \geq \frac{t}{4}s_jp_j \geq \frac{t}{4} \cdot \frac{2^j}{8\gamma\ell} \cdot \frac{2^{-j}}{40\ell} \geq \Omega(\log n).$$

Thus, a standard application of the Chernoff bound suffices to conclude that $c_j \geq \frac{t}{8}s_jp_j$, w.h.p., and then $\tilde{s}_j = \frac{c_j}{t} \cdot \frac{1}{p_j} \geq \frac{1}{8}s_jp_j \cdot \frac{1}{p_j} = \frac{1}{8}s_j$, also with high probability. This concludes the proof of Lemma 3.3. ■

We now prove the Lemma 3.4 that upper bounds the estimate $\tilde{s}_j$.

Proof of Lemma 3.4: First, fix some $j \in [\ell]$, and consider any particular hash table $H^{(j,u)}$. As before, let $A_{j,u}$ denote the event that (j, u) is an accepting pair, and define the following new event:

E3: at least one element of $S_{j-1} \cup S_j \cup S_{j+1}$ is subsampled.

Claim 3.8. *If $E3$ does not occur, $A_{j,u}$ holds with probability at most $p_j\left(\frac{2^j}{8\gamma\ell}\right)$. Moreover, $E3$ holds with probability at most $p_j(s_{j-1} + s_j + s_{j+1})$.*

Proof: For the first part, we prove that, with probability at least $1 - p_j\left(\frac{2^j}{8\gamma\ell}\right)$, no cell of $H^{(j,u)}$ can have a norm that is in the accepting range of $(0.9 \cdot T_j, 2.1 \cdot T_j]$. A cell v of $H^{(j,u)}$ may have a norm in the accepting range only when either: (1) more than one element from $S_{\leq j-2} = \cup_{j'\leq j-2} S_{j'}$ falls into v, or (2) the noise in v from elements in $S_{\geq j+2} = \cup_{j'\geq j+2} S_{j'}$ exceeds $0.6 \cdot T_j$. In particular, if neither (1) nor (2) hold, then either v contains no element from $S_{\leq j-2}$, in which case $\|H^{(j,u)}_v\|_X \leq 0.6 \cdot T_j \leq 0.9T_j$, or v contains exactly one element from $S_{\leq j-2}$, in which case $\|H^{(j,u)}_v\|_X > 4T_j - 0.6T_j > 2.1T_j$.

Now, the probability that (2) holds for *any* cell v is at most $\frac{\log^2 n}{w}$ by the Noise Lemma 3.5. It remains to bound the probability of (1), that more than one element from $S_{\leq j-2}$ falls into the same cell of the hash table. We note that the expected number of subsampled elements from $S_{\leq j-2}$ is upper bounded by $2^j \cdot p_j \leq O(1)$. Thus, with high probability, only at most $\log n$ of the elements from $S_{\leq j-2}$ may appear in $I_{j,u}$. Furthermore, these $\log n$ elements collide with probability at most $\frac{\log^2 n}{2w}$. It follows that the probability that (1) holds for *any* cell v is at most $\frac{\log^2 n}{w}$.

Thus, we have that

$$\Pr[A_{j,u} \mid \overline{E3}] \leq 2 \cdot \frac{\log^2 n}{w} \leq p_j\left(\frac{2^j}{8\gamma\ell}\right) = \frac{1}{320\gamma\ell^2}.$$

For the second part, we need to bound the probability $\Pr[E3]$. But this follows from a simple union bound over all elements in $S_{j-1} \cup S_j \cup S_{j+1}$. ■

We can now finish the proof of the lemma. From the above claim, we obtain the following bound on the probability of an accepting pair:

$$\Pr[A_{j,u}] \leq \Pr[A_{j,u} \mid \overline{E3}] + \Pr[E3]$$
$$\leq p_j\left(s_{j-1} + s_j + s_{j+1} + \tfrac{2^j}{8\gamma\ell}\right).$$

We can now upper bound the estimate $\tilde{s}_j$:

$$\mathbb{E}[\tilde{s}_j] = \frac{\sum_u \Pr[A_{j,u}]}{t} \cdot \frac{1}{p_j} \leq \left(s_{j-1} + s_j + s_{j+1} + \frac{2^j}{8\gamma\ell}\right).$$

Again, by a Chernoff bound, $\tilde{s}_j \leq 2(s_{j-1}+s_j+s_{j+1}+\frac{2^j}{8\gamma\ell})$ w.h.p. This completes the proof of Lemma 3.4. ■

4. Proof of Theorem 1.1

We now prove our main Theorem 1.1. As mentioned in the introduction, its main ingredient is Theorem 1.5.

Proof of Theorem 1.1: The sketch F consists of two parts. The first part is just a linear map f of planar EMD into ℓ_1 as in [5], [17], that approximates the EMD distance up to $\gamma = O(\log \Delta)$ approximation.

The second part is a collection of $O(\log \Delta)$ sketches ν_i. Each ν_i is a composition of two linear maps: the map $F^{(i)} = \langle F_1^{(i)}, \ldots F_T^{(i)} \rangle$ obtained from an application of Fact 1.4 and a sketch μ_i obtained from an application of the Theorem 1.5. Specifically, for $i \leq \log \Delta$, the sketch μ_i is given by the Theorem 1.5 for $M = 2^i$, $n = T$, and γ as defined above. The final sketch is then the following linear map:

$$F = \langle f, \mu_1 \circ F^{(1)}, \ldots \mu_{\log \Delta} \circ F^{(\log \Delta)} \rangle.$$

The reconstruction algorithm E works in a straightforward manner. Given sketches $Fx(A)$ and $Fx(B)$, compute first a γ approximation to $EMD(A,B)$ using the map f. Then, use the corresponding map $\nu_i = \mu_i \circ F^{(i)}$ to compute the estimate $\sum_j \|F_j^{(i)}(x(A)-x(B))\|_{\mathrm{EEMD}}$. This estimate is a $O(1/\epsilon)$ approximation to $EMD(A,B)$ by Fact 1.4.

This finishes the proof of Theorem 1.1. ■

References

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

[2] P. K. Agarwal, A. Efrat, and M. Sharir, "Vertical decomposition of shallow levels in 3-dimensional arrangements and its applications," *SIAM Journal on Computing*, no. 29, pp. 912–953, 2000, previously appeared in SOCG'95.

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

[4] A. Andoni, P. Indyk, and R. Krauthgamer, "Overcoming the ℓ_1 non-embeddability barrier: Algorithms for product metrics," in *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2009, pp. 865–874.

[5] M. Charikar, "Similarity estimation techniques from rounding," in *Proceedings of the Symposium on Theory of Computing (STOC)*, 2002, pp. 380–388.

[6] M. Charikar, K. Chen, and M. Farach-Colton, "Finding frequent items in data streams," in *Proceedings of International Colloquium on Automata, Languages and Programming (ICALP)*, 2002.

[7] G. Cormode and S. Muthukrishnan, "Improved data stream summaries: The count-min sketch and its applications," *FSTTCS*, 2004.

[8] C. Estan and G. Varghese, "New directions in traffic measurement and accounting: Focusing on the elephants, ignoring the mice," *ACM Transactions on Computer Systems*, 2003.

[9] A. C. Gilbert, S. Guha, P. Indyk, Y. Kotidis, S. Muthukrishnan, and M. J. Strauss, "Fast, small-space algorithms for approximate histogram maintenance," in *ACM Symposium on Theoretical Computer Science*, 2002.

[10] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Beijing, China, October 2005.

[11] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.

[12] P. Indyk and D. Woodruff, "Optimal approximations of the frequency moments of data streams," *Proceedings of the Symposium on Theory of Computing (STOC)*, 2005.

[13] P. Indyk, "Algorithms for dynamic geometric problems over data streams," *Proceedings of the Symposium on Theory of Computing (STOC)*, 2004.

[14] ——, "Stable distributions, pseudorandom generators, embeddings and data stream computation," *J. ACM*, vol. 53, no. 3, pp. 307–323, 2006, previously appeared in FOCS'00.

[15] ——, "A near linear time constant factor approximation for euclidean bichromatic matching (cost)," in *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2007.

[16] ——, "Sketching, streaming and sublinear-space algorithms," 2007, Graduate course notes, available at `http://stellar.mit.edu/S/course/6/fa07/6.895/`.

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

[18] T. Jayram and D. Woodruff, "The data stream space complexity of cascaded norms," 2009, to appear in FOCS'09.

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

[20] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[21] A. McGregor, "Open problems in data streams and related topics," *IITK Workshop on Algorithms For Data Streams*, 2006, available at `http://www.cse.iitk.ac.in/users/sganguly/workshop.html`.

[22] S. Muthukrishnan, "Data streams: Algorithms and applications (invited talk at soda'03)," *Available at http://athos.rutgers.edu/∼muthu/stream-1-1.ps*, 2003.

[23] A. Naor and G. Schechtman, "Planar earthmover is not in L_1," *SIAM Journal on Computing*, vol. 37, no. 3, pp. 804–826, 2007, an extended abstract appeared in FOCS'06.

[24] N. Nisan, "Pseudorandom generators for space-bounded computation," *Proceedings of the Symposium on Theory of Computing (STOC)*, pp. 204–212, 1990.

[25] P. Vaidya, "Geometry helps in matching," *SIAM Journal on Computing*, vol. 18, pp. 1201–1225, 1989.