Rademacher-Sketch: A Dimensionality-Reducing Embedding for Sum-Product Norms, with an Application to Earth-Mover Distance

Elad Verbin[1] and Qin Zhang[2]

[1] Aarhus University
elad.verbin@gmail.com
[2] MADALGO*, Aarhus University
qinzhang@cs.au.dk

Abstract. Consider a *sum-product* normed space, i.e. a space of the form $Y = \ell_1^n \otimes X$, where X is another normed space. Each element in Y consists of a length-n vector of elements in X, and the norm of an element in Y is the sum of the norms of its coordinates. In this paper we show a constant-distortion embedding from the normed space $\ell_1^n \otimes X$ into a lower-dimensional normed space $\ell_1^{n'} \otimes X$, where $n' \ll n$ is some value that depends on the properties of the normed space X (namely, on its *Rademacher dimension*). In particular, composing this embedding with another well-known embedding of Indyk [18], we get an $O(1/\epsilon)$-distortion embedding from the earth-mover metric EMD_Δ on the grid $[\Delta]^2$ to $\ell_1^{\Delta^{O(\epsilon)}} \otimes \mathsf{EEMD}_{\Delta^\epsilon}$ (where EEMD is a norm that generalizes earth-mover distance). This embedding is stronger (and simpler) than the sketching algorithm of Andoni et al [4], which maps EMD_Δ with $O(1/\epsilon)$ approximation into sketches of size $\Delta^{O(\epsilon)}$.

1 Introduction

Sum-product norms A *normed space* $(X, \|\cdot\|_X)$ consists of a linear space X and a norm $\|\cdot\|_X$ (i.e. a positive function from X to the reals, which satisfies the triangle inequality and where for $c \in \mathbb{R}, x \in X$ it holds that $\|c \cdot x\|_X = |c| \cdot \|x\|_X$). A *sum-product normed space* is a normed space of the form $Y = \ell_1^n \otimes X$, where X is another normed space. Each element y in Y consists of a length-n vector $y = (x_1, \ldots, x_n)$ of elements in X, and the norm of y is the sum of the norms of its coordinates, namely, $\|y\|_Y = \sum_{i=1}^n \|x_i\|_X$. Sum-product normed spaces have arisen in the literature on streaming and sketching algorithms. In particular, in 2009 Andoni et. al. [6] used product normed space to overcome the ℓ_1 non-embeddability barrier for the Ulam metric. A year later Andoni and Nguyen [8] used sum-product of Ulam metrics to obtain faster approximation algorithms for computing the Ulam distance between two non-repetitive strings. Sum-product metrics have also been used by Andoni and Onak [9] to compute the edit distance between two strings in near-linear time.

* MADALGO is the Center for Massive Data Algorithmics, a center of the Danish National Research Foundation.

Given two normed spaces Y and Y', an *embedding* (also called *strong embedding*) of Y into Y' is a function $\phi : Y \to Y'$. The *distortion* of ϕ is analogous to the "approximation ratio" achieved by Y' as an approximation of Y. Specifically, the distortion of ϕ is the value

$$\max_{y \in Y} \left(\|y\|_Y \,/\, \|\phi(y)\|_{Y'} \right) \cdot \max_{y \in Y} \left(\|\phi(y)\|_{Y'} \,/\, \|y\|_Y \right) \ .$$

Efficiently-computable embeddings with small distortion have been of much recent interest in theoretical computer science and various branches of mathematics, see, e.g., [25, 19]. In particular, if there is an efficient algorithm for computing the norm in the space Y', then the norm in Y can be computed by first applying the embedding and then performing the computation in Y'; the approximation factor of this algorithm is equal to the distortion. Similar approaches were used when designing sketches and data structures: rather than design a data structure for Y from scratch, simply embed Y in an efficient way into a normed space for which good data structures are already known. Applications of this approach are too numerous to cite, but see for example Indyk's survey [16].

In this paper we show dimensionality-reducing embeddings in sum-product normed spaces: the goal is, given a normed space $Y = \ell_1^n \otimes X$, to find a small-distortion embedding of Y into a smaller-dimensional sum-product normed space $Y' = \ell_1^{n'} \otimes X$. Our embeddings are *generic*, in the sense that their general structures do not depend on the properties of X. This is the first such generic dimension-reduction work that we know of for sum-product spaces. Previous literature has considered dimension-reduction for particular spaces, such as ℓ_1^n or ℓ_2^n. For literature on dimension reduction in ℓ_1^n, see for example the paper of Andoni, Charikar, Neiman and Nguyen [5] as well as the references therein. For dimension reduction in ℓ_2^n, consider the classical Johnson-Lindenstrauss lemma [21].

1.1 Our results

We first define a central concept in this paper: the *Rademacher dimension* of a normed space. As far as we know, this definition was never used or given in the literature; it is somewhat related to the property of being a Rademacher type p metric for $p > 0$ (see e.g. [26], also see the recent paper by Andoni et. al. [7]) but it is not the same.

Definition 1. *A normed space X has* Rademacher dimension α *if for any natural number s, and for any $x_0, x_1, \ldots, x_{s-1} \in X$ with $\|x_i\|_X \leq T$, we have with probability at least $1 - 1/\alpha^c$ (for some universal constant c) that*

$$\left\| \sum_{i \in [s]} \varepsilon_i x_i \right\|_X \leq \alpha \cdot \sqrt{s} \cdot T.$$

Here, $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{s-1}$ are (± 1)-valued random variables such that $\mathbf{Pr}[\varepsilon_i = +1] = \mathbf{Pr}[\varepsilon_i = -1] = 1/2$ for all $i \in [s]$, and the probability is taken over the sample space defined by these variables. If there is no real number α which this holds, then we say that the Rademacher dimension of the space is ∞.

As an illustrative example, it is easy to see that the normed space $(\mathbb{R}^d, \|\cdot\|_1)$ has Rademacher dimension $O(d^2)$. The proof of this fact follows from Hoeffding's inequality, in a similar way as in Lemma 1 below.

For $x = \{x_0, x_1, \ldots, x_{n-1}\} \in \ell_1^n \otimes X$, we denote $\|x\|_{1,X} = \sum_{i\in[n]} \|x_i\|_X$. Our main theorem states that any sum-product normed space $\ell_1^n \otimes X$ can be weakly-embedded with distortion $O(1)$ into $\ell_1^{n'} \otimes X$, where $n' \approx \alpha$ is roughly the Rademacher dimension of X:

Theorem 1. *Let X be a normed space with Rademacher dimension α. Let $\lambda = \max\{\alpha, \log^3 n\}$. Then there exists a distribution over linear mappings $\mu : \ell_1^n \otimes X \to \ell_1^{\lambda^{O(1)}} \otimes X$, such that for any $x \in \ell_1^n \otimes X$ we have*

- $\|\mu(x)\|_{1,X} \geq \Omega(\|x\|_{1,X})$ *with probability* $1 - 1/\lambda^{\Omega(1)}$.
- $\|\mu(x)\|_{1,X} \leq O(\|x\|_{1,X})$ *with probability* 0.99.

Remarks:

1. The embedding is *linear*. This is an important property, since it allows efficient updating of the sketch given updates in a streaming way, as well as computing the associated distance function (the distance function associated with the normed space X is the function $d(x, y) = \|x - y\|_X$).
2. The above embedding is a *weak embedding*, in the sense that for each vector, the norm of its embedded representation is good with constant probability. Thus, for each particular instantiation of the random variable μ, we would expect a constant fraction of the vectors in the source space to embed to vectors that are too large in the target space. This is as opposed to a *strong embedding*, that would be good for all of the vectors simultaneously.
   In another dimension-reduction paper, Indyk [17] showed a weak dimension reduction in ℓ_1, which was sufficient for applications such as norm estimation in data streams and approximate nearest neighbor search. In general, weak embeddings seem applicable for most of the purposes where strong embeddings are used, and they might not encounter the same barriers as strong embeddings: our embedding is in fact a good example of this, as we explain next.
3. The last theorem states that $\ell_1^n \otimes X$ can be weakly-embedded into $\ell_1^{n'} \otimes X$ with constant distortion, where $n' \approx \alpha$. It is natural to ask whether there exists a strong embedding with similar properties. The answer is a resounding "no": Even in the special case when X is simply $\ell_1^1 = \mathbb{R}$, a result by Brinkman and Charikar [10] shows that an n point subset of ℓ_1 cannot be embedded into $\ell_1^{n^{\Omega(1/D^2)}}$ with distortion $o(D)$. Thus, if we require a strong embedding with constant distortion, the dimension can be reduced by no more than constant factors.
4. Also, it is interesting to note that Theorem 1 works when X is a *normed space*, but if X was a *metric space* (i.e. a space where we have a measure for the distance between any two points, but not necessarily a norm for each point) then it is not clear whether any similar result can be obtained. Our embeddings inherently rely on the properties of normed space: in particular, we need the ability to sum elements of the space, which is not available in metric spaces.

What happens when the underlying space X has bad, or even infinite, Rademacher dimension? We can still achieve dimensionality reduction, but this time the more we want to reduce the dimension, the larger the distortion will be. Specifically, to reduce from dimension n to dimension n^ϵ, the distortion will be $O(1/\epsilon)$:

Theorem 2. *For any normed space X and any $\lambda \geq \log^3 n$, there exists a distribution over linear mappings $\mu : \ell_1^n \otimes X \to \ell_1^{\lambda^{O(1)}} \otimes X$, such that for any $x \in \ell_1^n \otimes X$, we have*

- $\|\mu(x)\|_{1,X} \geq \Omega(\|x\|_{1,X})$ *with probability* $1 - 1/\lambda^{\Omega(1)}$.
- $\|\mu(x)\|_{1,X} \leq O(\log_\lambda n \cdot \|x\|_{1,X})$ *with probability* 0.99.

The results of this theorem are easier to achieve, and might be folklore in the field. Theorem 2 can be obtained from a similar embedding as we use for proving Theorem 1 and with a simpler proof, so for most of the paper we concentrate on proving Theorem 1, and where appropriate we discuss Theorem 2.

2 Earth-Mover Distance

2.1 Introduction to Earth-Mover Distance

Earth-mover distance Denote $[n] = \{0, 1, \ldots, n-1\}$. Given two multisets A, B in the grid $[\Delta]^2$ with $|A| = |B| = N$, the *earth-mover distance* is defined as the minimum cost of a perfect matching between points in A and B, where the cost of two matched points $a \in A$ and $b \in B$ is the ℓ_1 distance between them. Namely,

$$\mathsf{EMD}(A, B) = \min_{\pi: A \to B \text{ a bijection}} \sum_{a \in A} \|a - \pi(a)\|_1 \ .$$

Earth-Mover distance (EMD) is a natural metric that measures the difference between two images: If one, for example, thinks of pixels of a certain color laid out in two images, then the distance between the two images can be defined as the minimum amount of work to move one set of pixels to match the other. EMD has been extensively used in image retrieval and experiments show that it outperform many other similarity measures in various aspects [31, 15, 14, 12, 30].

Historically, EMD is a special case of the *Kantorovich metric*, which is proposed by L. V. Kantorovich in an article in 1942 [22]. This metric has numerous applications in probabilistic concurrency, image retrieval, data mining and bioinformatics. One can refer to [13] for a detailed survey. Other equivalent formulations of EMD including Transportation distance, Wasserstein distance and Mallows distance [24].

The general (non-planar) EMD can be solved by the classical *Hungarian method* [23] in time $O(N^3)$. However, this approach is too expensive to scale to perform retrievals from large databases. A number of (approximation) algorithms designed for the planar case are proposed in literature [32, 1, 33, 11, 20, 2, 18]. In particular, Indyk [18] proposed a constant approximation algorithm with running time $O(N \log^{O(1)} N)$, which is almost linear.

Recently, there has been an increasing interest in designing sketching algorithms for EMD. A good sketch can lead to space/time-efficient streaming algorithms and nearest neighbor algorithms [4]. Searching for good sketching algorithms for EMD is considered to be a major open problem in the data stream community [27].

For a multiset A in the grid $[\Delta]^2$, a sketching algorithm defines a mapping f that maps A into a host space (such as the space of short bit-strings $\{0,1\}^S$). The sketching algorithm must satisfy the property that for any two multisets A and B, the earth-mover distance $\mathsf{EMD}(A,B)$ can be approximately reconstructed from the two sketches $f(A)$ and $f(B)$. The sketching algorithm and the reconstruction algorithm should be *space-efficient* (and for practical considerations sometimes also time-efficient) in order to get efficient data structures and streaming algorithms. An embedding can thus be seen as a special type of sketch, where the reconstruction algorithm consists simply of computing the norm of the difference $f(A) - f(B)$ in the host space. Some embeddings of EMD into ℓ_1 space were proposed in [11, 20]. However, in [28] the authors showed that it is impossible to embed EMD_Δ into ℓ_1 with distortion $o(\sqrt{\log \Delta})$. Therefore to get constant-approximation algorithms we need to investigate other, probably more sophisticated host spaces.

Recently, Andoni et. al. [4] obtained a sketch algorithm for planar EMD with $O(1/\epsilon)$ approximation ratio and Δ^ϵ space for any $0 \le \epsilon \le 1$. This is the first sublinear sketching algorithm for EMD achieving constant approximation ratio. Their sketching algorithm is not an embedding since their reconstruction algorithm involves operations such as binary decisions which are not metric operations. It remains an interesting open problem to embed EMD into simple normed spaces or products of simple normed spaces with constant distortion.

2.2 Applying our Results to Earth-Mover Distance

We now introduce the metric EEMD, which is an extension of EMD to any multisets $A, B \subseteq [\Delta]^2$ not necessary having the same size. It is defined as follows:

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

It is easy to see that when $|A| = |B|$, we have $\mathsf{EEMD}(A,B) = \mathsf{EMD}(A,B)$.

EEMD can be further extended to a *norm*: For a multiset $A \subseteq [\Delta]^2$, let $x(A) \in \mathbb{R}^{\Delta^2}$ be the characteristic vector of A. We next define the norm EEMD such that for any multiset $A, B \subseteq [\Delta]^2$, we have $\mathsf{EEMD}(A,B) = \|x(A) - x(B)\|_{\mathsf{EEMD}}$. The norm $\|\cdot\|_{\mathsf{EEMD}}$ is defined as follows: for each $x \in \mathbb{Z}^d$, let x^+ contain only the positive entries in x, that is, $x^+ = (|x| + x)/2$, and let $x^- = x - x^+$. And then we define $\|x\|_{\mathsf{EEMD}} = \mathsf{EEMD}(x^+, x^-)$. One can easily verify that this norm is well-defined. This definition can also be easily extended to $x \in \mathbb{R}^d$ by an appropriate weighting.

Let EEMD_Δ denote the normed space $(\mathbb{R}^{\Delta^2}, \|\cdot\|_{\mathsf{EEMD}})$.

Lemma 1. *EEMD_Δ has Rademacher dimension Δ^4.*

Proof. For any $x_0, \ldots, x_{s-1} \in \mathbb{R}^{\Delta^2}$ with $\|x_i\|_{\mathsf{EEMD}} \le T$ for all $i \in [s]$, let x_i^d ($d \in [\Delta^2]$) be the d-th coordinates of x_i. By the triangle inequality and the definition of

EEMD, $\left\|\sum_{i\in[s]}\varepsilon_i x_i\right\|_{\mathsf{EEMD}} \le \sum_{d\in[\Delta]^2}\left(\Delta\left|\sum_{i\in[s]}\varepsilon_i x_i^d\right|\right)$, where $\varepsilon_0,\ldots,\varepsilon_{s-1}$ are (±1)-valued random variables. Thus we only need to bound $\left|\sum_{i\in[s]}\varepsilon_i x_i^d\right|$ for each $d\in[\Delta^2]$.

Fix a $d\in[\Delta^2]$. Since for each x_i ($i\in[s]$), we have that $\left|x_i^d\right|\le T$. By Hoeffding's inequality we have that $\mathbf{Pr}\left[\left|\sum_{i\in[s]}\varepsilon_i x_i^d\right|\ge\Delta\sqrt{s}T\right] \le 2e^{-\frac{2(\Delta\sqrt{s}T)^2}{s\cdot(2T)^2}} = e^{-\Omega(\Delta^2)}$. Therefore with probability at least $1-\Delta^2\cdot e^{-\Omega(\Delta^2)}\ge 1-1/\Delta^{\Omega(1)}$, we have that $\left\|\sum_{i\in[s]}\varepsilon_i x_i\right\|_{\mathsf{EEMD}}\le\Delta^4\sqrt{s}T$.

The following fact is shown by Indyk [18].

Fact 1 *([18]) For any $\epsilon\in(0,1)$, there exists a distribution over linear mappings $F=\langle F_0,\ldots,F_{n-1}\rangle$ with $F_i:\mathsf{EEMD}_\Delta\to\mathsf{EEMD}_{\Delta^\epsilon}$ for all $i=0,\ldots,n-1$, such that for any $x\in\mathsf{EEMD}_\Delta$ we have*

- $\|x\|_{\mathsf{EEMD}}\le\sum_{i\in[n]}\|F_i(x)\|_{\mathsf{EEMD}}$ *with probability* 1.
- $\sum_{i\in[n]}\|F_i(x)\|_{\mathsf{EEMD}}\le O(1/\epsilon)\,\|x\|_{\mathsf{EEMD}}$ *with probability* 0.95.

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

Combining Theorem 1, Lemma 1 and Fact 1, we have the following.

Theorem 3. *For any $\epsilon\in\left[\frac{\log\log\Delta}{\log\Delta},1\right]$, there exists a distribution over linear mappings $\nu:\mathsf{EEMD}_\Delta\to\ell_1^{\Delta^{O(\epsilon)}}\otimes\mathsf{EEMD}_{\Delta^\epsilon}$, such that for any two $A,B\subseteq[\Delta]^2$ of equal size, we have*

- $\|\nu(x(A)-x(B))\|_{1,\mathsf{EEMD}}\ge\Omega(\mathsf{EMD}(A,B))$ *with probability* $1-1/\Delta^{\Omega(\epsilon)}$.
- $\|\nu(x(A)-x(B))\|_{1,\mathsf{EEMD}}\le O(1/\epsilon\cdot\mathsf{EMD}(A,B))$ *with probability* 0.9.

The embedding given by this theorem can also serve as an alternative to the sketching algorithm of Andoni et al. [4]; it is simpler so its actual performance might be better. Furthermore, there might be additional advantages to having an embedding rather than a sketching algorithm (e.g., if there exists a good nearest neighbor data structure for $\ell_1^{\Delta^{O(\epsilon)}}\otimes\mathsf{EEMD}_{\Delta^\epsilon}$, then we can use it to answer nearest neighbor queries for EMD_Δ).

3 The Embedding

In this section we construct the random linear mapping μ from Theorem 1. The random linear mapping for Theorem 2 is the same, and its analysis is simpler; we shall address the differences in Section 4.3.

Before giving the embedding, we first introduce a few definitions. Let $x=(x_0,\ldots,x_{n-1})\in\ell_1^n\otimes X$ be the vector that we want to embed. The embedding will work in ℓ levels, where $\ell=\lceil\log_\lambda(4\lambda n)\rceil$. Note that $\ell\le\lambda^{1/3}$ since $\lambda\ge\log^3 n$. At each level $k\in[\ell]$ we define a parameter $p_k=\lambda^{-k}$. For each level k we define a subsampled set, a hash function, and a series of (±1)-valued random variables, all of them random and

independent. The subsampled set is a set $I_k \subseteq [n]$ such that each $i \in [n]$ is placed in I_k with probability p_k; the hash function is a random function $h_k : [n] \to [t]$ where $t = \lambda^5$; the (± 1)-valued variables are $\varepsilon_{k,1}, \ldots, \varepsilon_{k,n}$, each of them is $+1$ with probability $1/2$ and -1 with probability $1/2$. All the random choices are independent.

We denote $\chi[E] = 1$ if event E is true and $\chi[E] = 0$ if it is false.

The embedding μ. For each level $k \in [\ell]$ and for each value $v \in [t]$ of the hash function h_k, compute

$$Z_k^v = \textstyle\sum_{i \in [n]} \chi[i \in I_k] \cdot \chi[h_k(i) = v] \cdot \varepsilon_{k,i} \cdot x_i \cdot 1/p_k \ .$$

We see that the embedded vector $\mu(x) \in \ell_1^{t \cdot \ell} \otimes X$ consists of all the values Z_k^v, one after another. These are $t \cdot l = \lambda^{O(1)}$ cells (=coordinates), each of which contains an element from X.

Remarks: The use of ± 1 random variables, also known in this context as Rademacher random variables, is superficially similar to usage in the seminal paper of Alon et. al. [3]. However, these variables are used here for an entirely different purpose. In [3] and other related work, these variables are used to decrease the variance of a random variable that estimates the second frequency moment of a steam of items. In our algorithm they are used for a different purpose: roughly speaking, they are used to isolate a class of items with norms in a certain range from items with much smaller norms by making the variables with smaller norm cancel with one another.

For the purpose of proving Theorem 2, the ± 1 random variables are not needed, and it is enough to define $Z_k^v = \sum_{i \in [n]} \chi[i \in I_k] \cdot \chi[h_k(i) = v] \cdot x_i \cdot 1/p_k$.

Also note that to use the above embedding as a sketching algorithm, it is necessary to remember all the random choices we made. This amount of space is huge: much more than n. However, this is not actually necessary. A standard approach using pseudo-random generators allows to decrease the amount of random bits to $\lambda^{O(1)}$, thus giving the "correct" space complexity. These random bits can be generated by Nisan's pseudo-random generator [29]. See the similar discussions in [17, 4].

4 Analysis

We first introduce a few more definitions. Let $M = \|x\|_{1,X}$. Let $T_j = M/\lambda^j$ $(j = 0, 1, \ldots)$. Let $S_j = \{i \in [n] \mid \|x_i\|_X \in (T_j/\lambda, T_j]\}$ and let $s_j = |S_j|$. We say x_i is in class j if $i \in S_j$.

It is easy to see that we only need to consider classes up to $\ell - 1$ since elements from classes $j \geq \ell$ contribute at most $n \cdot M/\lambda^\ell \cdot \ell \leq M/4$ to all the levels. Therefore for simplicity we assume that all elements belong to classes $\{0, 1, \ldots, \ell - 1\}$.

Let $\beta = 1/100\ell$. We say class $j \in [\ell]$ is *important* if elements from S_j contribute significantly to the sum M, that is, $\sum_{i \in S_j} \|x_i\|_X \geq \beta M$. Thus for an important class j we have $s_j \geq \beta \cdot M/T_j = \beta\lambda^j$. Also note that $s_j \leq M/(T_j/\lambda) = \lambda^{j+1}$ for all $j \in [\ell]$ by definition. Therefore $s_j \in [\beta\lambda^j, \lambda^{j+1}]$ for each important class j. Let J denote the set of all important classes.

During the analysis when we say an event holds with high probability we mean that the probability is at least $1 - 1/\lambda^{\Omega(1)}$.

A few more notations. Before the analysis, we would like to introduce a few more notations to facilitate our exposition. For item class $j \in [\ell]$, sample level $k \in [\ell]$ and cell $v \in [t]$, we define the following random variables.

- Let $S_{j,k}$ be the set of elements in class j that are sampled at sample level k. That is, $S_{j,k} = S_j \cap I_k$. Let $s_{j,k} = |S_{j,k}|$.
- Let $S^v_{j,k}$ be the set of elements in class j that are sampled at level k and hashed to cell v. That is, $S^v_{j,k} = \{i \in S_{j,k} \mid h_k(i) = v\}$. Let $s^v_{j,k} = \left|S^v_{j,k}\right|$.
- For each class j, let $W(S_j) = \sum_{i \in S_j} \|x_i\|_X$. And for each class j and sample level k, let $W(S_{j,k}) = \sum_{i \in S_{j,k}} \|x_i\|_X \cdot 1/p_k$.
- For each class j, sample level k and cell v, let $Z^v_{j,k} = \sum_{i \in S^v_{j,k}} \varepsilon_{k,i} \cdot x_i \cdot 1/p_k$. Note that $\|Z^v_k\|_X \leq \sum_{j \in [\ell]} \left\|Z^v_{j,k}\right\|_X$ by the triangle inequality.
- For each class j and sample level k, let $C_{j,k} = \{v \mid \max\{i \mid i \in S^v_{j,k}\} = j\}$. We also say each cell $v \in C_{j,k}$ a j-dominated cell at level k.
- Let $W(C_{j,k}) = \sum_{v \in C_{j,k}} \|Z^v_k\|_X$. That is, the sum of X-norms of class j elements in those j-dominated cells at level k. Moreover, let $W(C_{j,k}, j') = \sum_{v \in C_{j,k}} \left\|Z^v_{j',k}\right\|_X$ and $W(C_{j,k}, \geq j') = \sum_{j'' \geq j'} W(j, k, j'')$. Note that by the triangle inequality we have $W(C_{j,k}, j) - W(C_{j,k}, \geq j+1) \leq W(C_{j,k}) \leq W(C_{j,k}, j) + W(C_{j,k}, \geq j+1)$.

We need the following tool (c.f. [4]).

Lemma 2. *(A variant of Hoeffding bound) Let $Y_0, Y_1, \ldots, Y_{n-1}$ be n independent random variables such that $Y_i \in [0, T]$ for some $T > 0$. Let $\mu = \mathbf{E}[\sum_i Y_i]$. Then for any $a > 0$, we have $Pr\left[\sum_{i \in [n]} Y_i > a\right] \leq e^{-(a-2\mu)/T}$.*

Now we prove Theorem 1. We accomplish it by two steps.

4.1 No underestimation

In this section we show that $\|\mu(x)\|_{1,X} \geq \Omega(M)$ with probability $1 - 1/\lambda^{\Omega(1)}$. To show this we first prove the following lemma,

Lemma 3. *For each important class $j \in J$, we have $W(C_{j,j-1}) \geq W(S_j)/2$ for all $j \geq 1$ and $W(C_{0,0}) \geq W(S_0)/2$ for $j = 0$ with probability at least $1 - 1/\lambda^{\Omega(1)}$.*

The proof of the lemma is essentially that for each important class $j \in J$, at sample level $\max\{j-1, 0\}$, the scaled contribution of elements in class j is close to $W(S_j)$ and the noise from other classes is small. The intuition of the later is simply that the range of each hash function is large enough such that:

1. There is no collision between elements in S_j and elements in $\bigcup_{j' > j} S_{j'}$ with high probability.
2. The noise from $\bigcup_{j' > j} S_{j'}$ is small since only a small fraction of elements from each $S_{j'}$ $(j' < j)$ will collide with elements in S_j and the X-norm of each item in $S_{j'}$ $(j' < j)$ is much smaller compared with those in S_j.

Proof. (of Lemma 3.) For notational convenience we set $k = j - 1$ in this proof. For each important class $j \geq 1$, we have $\mathbf{E}[|S_{j,k}|] = s_j p_k \in [\beta\lambda, \lambda^2]$. By Chernoff bound we have that with probability at least $1 - e^{-\Omega(\beta\lambda)}$, $\beta\lambda/2 \leq |S_{j,k}| \leq 2\lambda^2$. Similarly, we have that $|S_{j',k}| \leq 2\lambda$ for all class $j' \leq j-1$ with probability $1-\ell \cdot e^{-\Omega(\lambda)}$. Conditioned on these, the probability that any two of $\bigcup_{j' \leq j} S_{j',k}$ hash into the same cell is at most $\binom{2\lambda^2+\ell\cdot 2\lambda}{2}/t \leq O(1/\lambda)$. That is, with probability at least $1 - O(1/\lambda) - \ell \cdot e^{-\Omega(\lambda)} = 1 - 1/\lambda^{\Omega(1)}$, there is no collision between elements in class $0, 1, \ldots, j$ at sample level k. In particular, we have $|C_{j,k}| = |S_{j,k}| \in [\beta\lambda/2, 2\lambda^2]$ and $W(C_{j,k}, j) = W(S_{j,k})$ for each class $j \geq 1$ with high probability.

With similar arguments we can show that $W(C_{0,0}, 0) = W(S_{0,0})$ with high probability if $0 \in J$.

Now we show that $W(S_{j,k}) \geq \frac{2}{3} W(S_j)$ with high probability for each important class $j \geq 1$. We define for each $i \in S_j$ a random variable $Y_i = \chi[i \in S_{j,k}] \cdot \|x_i\|_X / T_j$. Note that $0 \leq Y_i \leq 1$ for all $i \in S_j$ and $W(S_{j,k}) = (\sum_{i \in S_j} Y_i) \cdot T_j \cdot 1/p_k$. We also have that

$$\mu = \mathbf{E}[\textstyle\sum_{i \in S_j} Y_i] = W(S_j) p_k / T_j = \Omega(\beta M / T_j \cdot \lambda^{-(j-1)}) = \Omega(\beta \lambda^j \cdot \lambda^{-(j-1)}) = \Omega(\beta\lambda).$$

By Chernoff bound we have $\mathbf{Pr}\left[\left|\sum_{i \in S_j} Y_i - \mu\right| \geq \mu/3\right] \leq 2e^{-\Omega(\mu)} \leq e^{-\Omega(\beta\lambda)}$.

Therefore with probability at least $1-\ell\cdot e^{-\Omega(\beta\lambda)} \geq 1-1/\lambda^{\Omega(1)}$, we have $W(S_{j,k}) \geq \frac{2}{3} W(S_j)$ for all important classes $j \geq 1$. Consequently, we have $W(C_{j,k}, j) \geq \frac{2}{3} W(S_j)$ for all important $j \geq 1$ with high probability.

Moreover, note that $W(S_{0,0}) = W(S_0)$ is trivial since we pick each item at sample level 0. Therefore it also holds that with high probability, $W(C_{0,0}, 0) = W(S_{0,0}) = W(S_0)$ if $0 \in J$.

Next we bound $W(C_{j,k}, \geq j+1)$ for each important class $j \geq 1$ and $W(C_{0,0}, \geq 1)$ if $0 \in J$. We first bound the former. For each class $j' \geq j+1$, let $w = |C_{j,k}|$, we have $\mathbf{E}\left[\sum_{i \in S_{j',k}} \chi[h_k(i) \in C_{j,k}]\right] = s_{j'} p_k \cdot w/t$. By Lemma 2 we have that $\mathbf{Pr}\left[\sum_{i \in S_{j',k}} \chi[h_k(i) \in C_{j,k}] \geq 2 \cdot s_{j'} p_k \cdot w/t + \lambda\right] \leq e^{-\Omega(\lambda)}$. Summing up for all classes $j' \geq j+1$, with probability at least $1 - \ell \cdot e^{-\Omega(\lambda)}$, we have

$$\begin{aligned} W(C_{j,k}, \geq j+1) &\leq \textstyle\sum_{j' \geq j+1} \sum_{i \in S_{j',k}} \chi[h_k(i) \in C_{j,k}] \cdot T_{j'} \cdot 1/p_k \\ &\leq \textstyle\sum_{j' \geq j+1} (2 \cdot s_{j'} p_k \cdot w/t + \lambda) \cdot T_{j'} \cdot 1/p_k \\ &\leq \ell \cdot (2 \cdot M \cdot w/\lambda^5 + T_j \cdot 1/p_k) \\ &\leq o(\beta M) \leq o(W(S_j)) \end{aligned}$$

The second to the last inequality holds since $w \leq 2\lambda^2$ with probability at least $1 - e^{-\Omega(\lambda)}$. Therefore by a union bound over $j \in [\ell]$ we have that with probability at least $1 - \ell^2 \cdot e^{-\Omega(\lambda)} \geq 1 - 1/\lambda^{\Omega(1)}$, $W(C_{j,k}, \geq j+1) = o(W(S_j))$ for all important $j \geq 1$.

Similarly, we can show that $W(C_{0,0}, \geq 1) = o(W(S_0))$ with high probability if $0 \in J$.

Finally, for each important class $j \geq 1$, we have $W(C_{j,k}) \geq W(C_{j,k}, j) - W(C_{j,k}, \geq j+1)$. Therefore $W(C_{j,k}) \geq \frac{2}{3} W(S_j) - o(W(S_j)) \geq W(S_j)/2$ for all important

$j \geq 1$ with high probability. Similarly we can show that $W(C_{0,0}) \geq W(S_0)/2$ with high probability if $0 \in J$.

Note that Lemma 3 immediately gives the following. With probability at least $1 - 1/\lambda^{\Omega(1)}$, we have

$$\begin{aligned}\|\mu(x)\|_{1,X} &\geq \textstyle\sum_{j\in J: j\geq 1} W(C_{j,j-1}) + \chi[0 \in J] \cdot W(C_{0,0}) \\ &\geq \textstyle\sum_{j\in J} W(S_j)/2 = \left(M - \sum_{j\notin J} W(S_j)\right)/2 \\ &\geq (M - \ell \cdot \beta M)/2 \geq \Omega(M).\end{aligned}$$

4.2 No overestimation

In this section we show that $\|\mu(x)\|_{1,X} \leq O(M)$ with probability 0.99. The general idea is the following:

1. Elements from S_j contribute little to all levels $k < j - 9$. This is because in each of such levels k and each cell $v \in [t]$ at that level, many elements from S_j are sampled and hashed into v and they will cancel with each other heavily due to the random variables $\varepsilon_{k,i} \in \{+1, -1\}$ multiplied.
2. On the other hand, elements from S_j will not be sampled at all levels $k > j + 1$ with high probability according to the sample ratios we choose at each level.

Now we prove the second part of Theorem 1. By the triangle inequality and the fact that $\sum_{v\in[t]} \left\|Z^v_{j,k}\right\|_X \leq W(S_{j,k})$ we have

$$\begin{aligned}\|\mu(x)\|_{1,X} &\leq \sum_{j\in[\ell]} \sum_{k\in[\ell]} \sum_{v\in[t]} \left\|Z^v_{j,k}\right\|_X \\ &\leq \sum_{j\in[\ell]} \sum_{k=j+2}^{\ell-1} W(S_{j,k}) + \sum_{j\in[\ell]} \sum_{k=j-9}^{j+1} W(S_{j,k}) + \sum_{j\in[\ell]} \sum_{k=0}^{j-10} \sum_{v\in[t]} \left\|Z^v_{j,k}\right\|_X .\end{aligned}$$

We bound the three terms of (1) separately. For the first term, the probability that there exists an element in class j that is sampled at level higher than $j + 2$ is at most $\ell \cdot s_j p_{j+2} \leq O(\ell/\lambda)$. Union bound over all class $j \in [\ell]$ we have that with probability at least $1 - O(\ell^2/\lambda) \geq 1 - 1/\lambda^{\Omega(1)}$, the first term is 0.

For the second term, since $\mathbf{E}[W(S_{j,k})] = W(S_j)$ for each $k \in [\ell]$, by the linearity of expectation and Markov inequality we obtain that with probability at least 0.991,

$$\sum_{j\in[\ell]} \sum_{k=j-9}^{j+1} W(S_{j,k}) \leq 2000 \cdot \sum_{j\in[\ell]} W(S_j) = 2000M.$$

Now we try to bound the third term. We know by lemma 2 that $s_{j,k} \leq 2s_j p_k + \lambda$ with probability at least $1 - e^{-\Omega(\lambda)}$. By the assumption X has Rademacher dimension

λ we have that for each j, k such that $j \geq k + 10$,

$$\begin{aligned}
\sum_{v\in[t]} \left\| Z^v_{j,k} \right\|_X &\leq \sum_{v\in[t]} \lambda T_j \sqrt{s^v_{j,k}} \cdot 1/p_k \\
&\leq \lambda T_j \cdot t\sqrt{s_{j,k}/t} \cdot \lambda^k \\
&\leq \lambda M/\lambda^j \cdot \lambda^5 \sqrt{2\lambda^{j+1}\lambda^{-k}/\lambda^5 + \lambda} \cdot \lambda^k \\
&\leq 2M \cdot \lambda^{1+5/2-\frac{j-k-1}{2}} \\
&\leq 2M\lambda^{1+5/2-9/2} = 2M/\lambda.
\end{aligned}$$

Summing over all $j \in [\ell]$ and all $k \in [0, \ldots, j-10]$ we have that the third term is at most $O(2M\ell^2/\lambda) = o(M)$ with probability at least $1 - \ell^2 \cdot e^{-\Omega(\lambda)} \geq 1 - 1/\lambda^{\Omega(1)}$.

To sum up, with probability at least $1 - 1/\lambda^{\Omega(1)} - (1 - 0.991) \geq 0.99$ we have $\|\mu(x)\|_{1,X} \leq 2000M + o(M) = O(M)$.

4.3 Proof for Theorem 2

We can also prove Theorem 2 by two steps. The proof for the first part of the theorem (i.e., no underestimate) is exactly the same as that in Theorem 1 since in that proof we do not use any property of Rademacher dimension. For the second part, just notice that $\|\mu(x)\|_{1,X} \leq \sum_{k\in[\ell]} \sum_{j\in[\ell]} W(S_{j,k})$ where $\ell = O(\log_\lambda n)$, and $\mathbf{E}[W(S_{j,k})] = W(S_j)$ for all $j, k \in [\ell]$. Thus $\mathbf{E}[\|\mu(x)\|_{1,X}] \leq O(\log_\lambda n) \cdot \sum_{j\in[\ell]} W(S_j) = O(\log_\lambda n \cdot M)$. Directly applying Markov inequality gives the result.

References

1. P. K. Agarwal, A. Efrat, and M. Sharir. Vertical decomposition of shallow levels in 3-dimensional arrangements and its applications. In *SoCG*, pages 39–50, 1995.
2. P. K. Agarwal and K. R. Varadarajan. A near-linear constant-factor approximation for euclidean bipartite matching? In *Symposium on Computational Geometry*, pages 247–252, 2004.
3. N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.*, 58:137–147, February 1999.
4. A. Andoni, K. D. Ba, P. Indyk, and D. Woodruff. Efficient sketches for earth-mover distance, with applications. In *FOCS*, 2009.
5. A. Andoni, M. S. Charikar, O. Neiman, and H. L. Nguyen. Near linear lower bound for dimension reduction in ℓ_1. In *IEEE Symposium on Foundations of Computer Science*, 2011.
6. A. Andoni, P. Indyk, and R. Krauthgamer. Overcoming the ℓ_1 non-embeddability barrier: algorithms for product metrics. In *SODA*, pages 865–874, 2009.
7. A. Andoni, R. Krauthgamer, and K. Onak. Streaming algorithms via precision sampling. In *FOCS*, pages 363–372, 2011.
8. A. Andoni and H. L. Nguyen. Near-optimal sublinear time algorithms for Ulam distance. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '10, pages 76–86, 2010.
9. A. Andoni and K. Onak. Approximating edit distance in near-linear time. In *STOC*, pages 199–204, 2009.

10. B. Brinkman and M. Charikar. On the impossibility of dimension reduction in ℓ_1. *J. ACM*, 52:766–788, September 2005.
11. M. S. Charikar. Similarity estimation techniques from rounding algorithms. In *STOC*, pages 380–388, 2002.
12. C. Chefd'hotel and G. Bousquet. Intensity-based image registration using earth mover's distance. In *SPIE*, 2007.
13. Y. Deng and W. Du. The Kantorovich metric in computer science: A brief survey. *Electr. Notes Theor. Comput. Sci.*, 253(3):73–82, 2009.
14. K. Grauman and T. Darrell. Fast contour matching using approximate earth movers distance. In *CVPR*, pages 220–227, 2004.
15. A. S. Holmes, C. J. Rose, and C. J. Taylor. Transforming pixel signatures into an improved metric space. *Image Vision Comput.*, 20(9-10):701–707, 2002.
16. P. Indyk. Algorithmic aspects of geometric embeddings. In *IEEE Symposium on Foundations of Computer Science*, 2001.
17. P. Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *J. ACM*, 53:307–323, May 2006.
18. P. Indyk. A near linear time constant factor approximation for euclidean bichromatic matching (cost). In *SODA*, pages 39–42, 2007.
19. P. Indyk and J. Matousek. Low-distortion embeddings of finite metric spaces. In *in Handbook of Discrete and Computational Geometry*, pages 177–196. CRC Press, 2004.
20. P. Indyk and N. Thaper. Fast color image retrieval via embeddings. In *Workshop on Statistical and Computational Theories of Vision (at ICCV)*, 2003.
21. W. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, 1984.
22. L. V. Kantorovich. On the translocation of masses. *Dokl. Akad. Nauk SSSR*, 37(7-8):227–229, 1942.
23. E. Lawler. *Combinatorial optimization - networks and matroids*. Holt, Rinehart and Winston, New York, 1976.
24. E. Levina and P. J. Bickel. The earth mover's distance is the mallows distance: Some insights from statistics. In *ICCV*, pages 251–256, 2001.
25. N. Linial. Finite metric spaces - combinatorics, geometry and algorithms. In *In Proceedings of the International Congress of Mathematicians III*, pages 573–586, 2002.
26. B. Maurey. Type, cotype and k-convexity. In *HANDBOOK OF THE GEOMETRY OF BANACH SPACES. VOLUME 2*, pages 1299–1332. North-Holland, 2003.
27. A. McGregor. Open problems in data streams, property testing, and related topics. http://www.cs.umass.edu/∼mcgregor/papers/11-openproblems.pdf, 2011.
28. A. Naor and G. Schechtman. Planar earthmover is not in ℓ_1. *SIAM J. Comput.*, 37(3):804–826, 2007.
29. N. Nisan. Pseudorandom generators for space-bounded computations. In *Proceedings of the twenty-second annual ACM symposium on Theory of computing*, STOC '90, pages 204–212, 1990.
30. J. Puzicha, J. M. Buhmann, Y. Rubner, and C. Tomasi. Empirical evaluation of dissimilarity measures for color and texture. In *ICCV*, pages 1165–1173, 1999.
31. Y. Rubner, C. Tomasi, and L. J. Guibas. The earth movers distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:2000, 2000.
32. P. M. Vaidya. Geometry helps in matching. *SIAM J. Comput.*, 18:1201–1225, December 1989.
33. K. R. Varadarajan and P. K. Agarwal. Approximation algorithms for bipartite and non-bipartite matching in the plane. In *SODA*, pages 805–814, 1999.