Earth Mover Distance over High-Dimensional Spaces*

Alexandr Andoni†
MIT
andoni@mit.edu

Piotr Indyk
MIT
indyk@mit.edu

Robert Krauthgamer
IBM Almaden
robi@almaden.ibm.com

October 11, 2007

Abstract

The Earth Mover Distance (EMD) between two equal-size sets of points in $\mathbb{R}^d$ is defined to be the minimum cost of a bipartite matching between the two pointsets. It is a natural metric for comparing sets of features, and as such, it has received significant interest in computer vision. Motivated by recent developments in that area, we address computational problems involving EMD over *high-dimensional* pointsets.

A natural approach is to embed the EMD metric into ℓ_1, and use the algorithms designed for the latter space. However, Khot and Naor [KN06] show that any embedding of EMD over the d-dimensional Hamming cube into ℓ_1 must incur a distortion $\Omega(d)$, thus practically losing all distance information. We circumvent this roadblock by focusing on sets with cardinalities upper-bounded by a parameter s, and achieve a distortion of only $O(\log s \cdot \log d)$. Since in applications the feature sets have bounded size, the resulting distortion is much smaller than the $\Omega(d)$ lower bound. Our approach is quite general and easily extends to EMD over $\mathbb{R}^d$.

We then provide a strong lower bound on the multi-round communication complexity of estimating EMD, which in particular strengthens the known non-embeddability result of [KN06]. Our bound exhibits a smooth tradeoff between approximation and communication, and for example implies that every algorithm that estimates EMD using constant size sketches can only achieve $\Omega(\log s)$ approximation.

1 Introduction

The *Earth Mover Distance (EMD)* between two sets of points in $\mathbb{R}^d$ of equal sizes (say, s) is defined to be the cost of the minimum cost bipartite matching between the two pointsets. It is a natural metric for comparing sets of geometric features of objects. For example, an image can be represented as a set of pixels in a color space; computing EMD between such sets yields an accurate measure of dissimilarity between color characteristics of the images [RTG00]. In an analogous manner, an image can be represented as a set of representative geometric features, such as object contours [GD04] and other features [GD05a].

Starting with the work of [RTG00], the EMD metric has received significant interest in computer vision. Unfortunately, the *computational* properties of EMD are not very good. The best exact algorithm known for computing EMD between two feature sets, even low-dimensional ones, has cubic running time [Law76][1]. Furthermore, in many applications, one needs to select one feature set out of a large collection of sets (e.g., the one closest to a given query set), which 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.

This computational bottleneck motivates the need for faster algorithms dealing with the EMD metric. A particularly versatile approach to this problem is through *metric embeddings*. In this approach, each feature set is mapped to a vector in a normed space (say ℓ_1), such that the distance between sets is well-approximated by the distance between the corresponding vectors (the approximation factor is called the *distortion* of the embedding). After performing the mapping, the distances can be estimated simply by computing the distances between the vectors. Analo-

*A preliminary version of this paper appeared as an ECCC report [?]

†Most of the work was done while the author was visiting IBM Almaden Research Center.

[1] Faster, but still super-quadratic algorithms, are known for the case of $d = 2$. Better running times are achievable by approximation algorithms, cf. [Ind07a] and references therein.

gously, the *nearest neighbor* problem over EMD space can be solved by using efficient algorithms developed for the vector space, such as kd-trees or other methods [KOR98, IM98].

It is shown in [Cha02, IT03] that if the feature sets are subsets of $[\Delta]^d$, then one can obtain an embedding into ℓ_1 with $O(d \log \Delta)$ distortion and furthermore the running time of computing the embedding of a given feature set is near-linear in the set size. Despite non-constant distortion bounds, the embeddings are still quite accurate [IT03, GD04, GD05a, CLL04, CLJ$^+$06, GD06a], especially for low values of d such as 2 or 3. The geometric representation of the EMD space which they provide turns out to be very useful for other reasons as well. For example, it has been used to design *Mercer kernels*, useful for supervised classification of the images [GD05b] and unsupervised learning [GD06b]. Also, the algorithmic techniques have led to efficient algorithms for computing EMD *without* performing the actual embedding. E.g., [Ind07a], building on the work of [AV04], gave an $s \log^{O(1)} s$ time constant factor approximation algorithm for computing EMD between low-dimensional pointsets.

Choice of Norm. Throughout the paper, we assume that distances in $\mathbb{R}^d$ are measured using to the ℓ_1-norm. This norm is often more convenient than the ℓ_2-norm (at least in the context of theoretical analysis of EMD). Our results can generally be adapted to the ℓ_2-norm (using e.g. a constant distortion embedding of ℓ_2^d into $\ell_1^{O(d)}$, see [Ind07b] for a recent account of such embeddings, or using a trivial embedding with distortion $\sqrt{d}$), but we will not discuss such extensions in this version of the paper. It should be noted however that some of the experimental results we mentioned, including [RTG00] and [GD05b], actually use the ℓ_2-norm.

High-Dimension. We focus on the case of high dimension d, which is also important in practice. For instance, in [GD06a] the image recognition rate was studied for dimension d in the range $[8, 128]$ and set-size $s = 256$. The previously mentioned embeddings of [Cha02, IT03, GD05b] suffer (both in theory and in practice [GD06a]) from a high approximation error when the dimension is "too large". Recently, a different method for embedding EMD over high-dimensional spaces was introduced [GD06a]; although achieving much lower empirical error, this method is heuristic and has no guarantees. In fact, it is known [KN06] that *every* embedding of EMD over a d-dimensional Hamming cube into ℓ_1 requires $\Omega(d)$ distortion. However, the lower bound is shown only for rather large subsets of the Hamming space, namely of size exponential in d, leaving open the possibility of designing much more accurate embeddings for sets of more moderate size $s = s(d)$, with distortion bound parametrized by (and growing slowly with) s. For example, an embedding that has low distortion when s is polynomial in d may be useful in practice for sets with a few hundreds or thousands of points.

Results I: Improved Embeddings. We prove that EMD metric over s-subsets of the d-dimensional Hamming cube $\{0,1\}^d$ can be embedded into ℓ_1 with distortion $O(\log s \cdot \log d)$. Since an $\Omega(\log s)$ lower bound on the distortion follows immediately from the aforementioned result of [KN06], our upper bound is within $O(\log d)$ of optimal. The embedding naturally extends to EMD over s-subsets of $[\Delta]^d$, achieving distortion bound $O(\log s \cdot \log(d\Delta))$. Since many metrics (e.g., variants of edit distance [MS00, CM07, OR05, CK06]) can be embedded with bounded distortion into ℓ_1 with small integer coordinates, our result implies that EMD over s-subsets of such metrics can be embedded into ℓ_1 as well. As hinted earlier, the standard applications of these embeddings include a fast estimate of EMD distance and algorithms for nearest neighbor problem over EMD space. Moreover, our embedding yields a low-distortion embedding of a metric called "shift metric", a metric related to the edit distance. These embedding results are presented in Section 3.

Although our embedding into ℓ_1 attains a near-optimal distortion, one might ask whether metric embedding (into ℓ_1 and in general) is the best algorithmic technique for dealing with EMD over high-dimensional spaces. To answer this question, we provide a strong *communication lower bound* that goes beyond the usual embedding lower bound. In particular, it strengthens and immediately implies the known non-embeddability result of [KN06].

Distance Estimation. The problem of estimating EMD in the sketching model is defined as follows: given s-subsets of $\{0,1\}^d$, preprocess each subset separately into a short representation called a *sketch*, so that the EMD distance between two subsets can be approximated only from their sketches. (The preprocessing may use shared randomness, but the approximation should succeed with probability at least $2/3$ over these random coins.) Solving this problem could lead to very fast approximation of EMD distance, since the second step of using the sketches to compute the approximation often runs in time that is linear in the sketch size. For example, it is known that estimating ℓ_1 distance in the sketching model (defined analogously to EMD) can be solved with $1+\epsilon$ approximation (in the sense of a relaxed decision problem) using sketches of size $O(1/\epsilon^2)$, for every $\epsilon > 0$ [KOR98]. Another application of an estimation algorithm that uses small sketches could be as a filtering method for the nearest neighbor problem (i.e.

the EMD distance between a query and each item in a data set could be estimated quickly from their sketches, avoiding many accurate but elaborate computations).

Results II: Communication Lower Bounds. We prove that a sketch of constant size can only yield $\Omega(d)$ approximation; in terms of s, a lower bound of $\Omega(\log s)$ follows immediately. Let us compare this result to non-embeddability into ℓ_1. Observe that a distortion D embedding of the EMD metric into ℓ_1 could be composed with the aforementioned estimate for ℓ_1 distance, yielding an EMD estimation with $2D$ approximation using constant-size sketches. We conclude that $D = \Omega(d)$, recovering the result of [KN06]. But these same arguments give a similar conclusion for embeddings into ℓ_2-squared (negative type) metrics, which strictly includes ℓ_1. We thus see that, in general, lower bounds for the sketching model are much stronger than non-embeddability into ℓ_1.

Our impossibility result for the sketching model is shown in the well-known framework of communication complexity and is actually more general—it pertains to multi-round communication complexity (see Section 2.2 for a review of standard terminology): Alice and Bob are two players that have access to shared randomness; upon receiving s-subsets A and B, respectively, as their inputs, they wish to determine, by exchanging messages, whether $\mathrm{EMD}(A, B) \leq r$ or $\mathrm{EMD} \geq \alpha r$, for some threshold $r > 0$ and approximation ratio $\alpha > 1$. Our main result says that in order to succeed with constant probability, the two parties must exchange a total of $\Omega(\frac{d}{\alpha})$ bits. This bound is optimal for $\alpha = \Omega(d)$, since d approximation is trivial using $O(1)$ bits. This communication lower bound appears in Section 4.

Techniques. A (somewhat simplified) overview of our embedding is as follows. There are two steps. First, we use dimensionality reduction techniques in ℓ_1^d to reduce the dimension of the underlying space to, say, $O(\log(d\Delta))$; this step is implemented by a randomized map that, with high probability, preserves up to a small approximation factor all pairwise distances in a set of $2s$ points. The second step uses an embedding of EMD over low-dimensional spaces [CCG+98], which incurs distortion at most linear in the dimension. In general, dimension reduction in ℓ_1 requires a large distortion [BC03, LN04]; but for our purposes, it essentially suffices to obtain a randomized mapping from $[\Delta]^d$ into a weighted Hamming cube of dimension $O(\log(d\Delta))$ (a subset of $\ell_1^{O(\log(d\Delta))}$) that with probability at least $1 - 1/s^2$ distorts the distance between a pair of points by at most $O(\log s \cdot \log(d\Delta))$. Furthermore, due to our particular weighting of the Hamming cube, we can embed EMD over this weighted cube into ℓ_1 with only constant distortion.

For the lower bound on the total communication between Alice and Bob, we first prove that it suffices to analyze very restricted protocols, where Alice sends only one bit to Bob. These protocols clearly correspond to boolean functions over $\{0, 1\}^d$, and can be analyzed directly using the powerful approach of Fourier analysis, employing in particular (but not only) the technique developed in [KN06]. We show that every such restricted protocol succeeds with a tiny probability, implying the desired lower bound on the communication complexity of general protocols. To the best of our knowledge, the only other metric for which a super-constant non-approximability in communication protocols is known is the ℓ_∞ metric (and the metrics embeddable into it with sufficiently small distortion, such as ℓ_p for $p > 2$) [SS02, BJKS04].

2 Preliminaries

We define the Earthmover distance over a general metric as follows.

Definition 2.1. *Consider a metric space X endowed with distance function d_X. Then, for two multi-sets $A, B \subset X$, of size $s = |A| = |B|$, the* earthmover distance (EMD) *between A and B is defined as*

$$\mathrm{EMD}_X(A, B) = \frac{1}{s} \min_{\phi: A \to B} \sum_{x \in A} d_X(x, \phi(x))$$

where the minimum is taken over all bijections $\phi : A \to B$. The resulting metric is called EMD over X.

When the metric X is ℓ_1, we will omit the subscript and simply write $EMD(A, B) = EMD_{\ell_1}(A, B) = \frac{1}{s} \min_{\phi:A\to B} \sum_{x\in A} \|x - \phi(x)\|_1$.

2.1 Notation.

We use $\lg x$ to denote logarithm in base 2, and denote $[\Delta] = \{1, 2, \ldots \Delta\}$. For a finite alphabet Σ and $x, y \in \Sigma$, we define $H(x, y) = 1$ if $x \neq y$ and $H(x, y) = 0$ if $x = y$. For a distribution μ on (x, y), its marginal distributions are called μ_x, μ_y. Let $\mathrm{Supp}(\mu)$ denote the support of μ. For two vectors $x, y \in \{0, 1\}^d$, let $x + y \in \{0, 1\}^d$ be the component-wise sum modulo 2.

2.2 Communication Complexity of Protocols and Sketching.

Consider a (partial) function $F(x, y) : \mathcal{D} \to \{0, 1\}$, where $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$; we view this function as a communication problem, where Alice gets $x \in \mathcal{X}$, Bob gets $y \in \mathcal{Y}$ and they want to compute $F(x, y)$ (given that $(x, y) \in \mathcal{D}$). To compute $F(x, y)$, Alice and Bob exchange some messages in rounds, where in each round one player sends exactly one message to the other player. At the end, Bob outputs the result. We call communication protocol Π one such procedure by which Alice and Bob compute the messages and their output. The protocol is valid if protocol's output is equal to $F(x, y)$ for all $(x, y) \in \mathcal{D}$.

The protocol is *randomized* if Alice and Bob have access to a shared infinite random string. If the protocol is randomized, then a valid protocol has to output $F(x, y)$ with probability at least $2/3$ for all $(x, y) \in \mathcal{D}$.

The communication complexity of a protocol Π is the maximum total length of the exchanged messages; we denote the communication complexity of Π by $|\Pi|$. Finally, the *communication complexity of a problem* F, denoted $\mathcal{R}(F)$, is the minimum communication complexity of a valid (randomized) protocol for the problem F.

We also define the *sketch size* of a problem $F : \mathcal{D} \to \{0,1\}$, $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$. In a sketching scenario, Alice and Bob independently send some l bits to a common referee, who ultimately decides the output of the function. Formally, sketch size of F is the minimum l such that there exist a distribution Π^S over functions $s_A : \mathcal{X} \to \{0,1\}^l$, $s_B : \mathcal{Y} \to \{0,1\}^l$, and output function $\phi_{ref} : \{0,1\}^l \times \{0,1\}^l \to \{0,1\}$, such that $\phi_{ref}(s_A(x), s_B(y)) = F(x, y)$ for every $(x, y) \in \mathcal{D}$ with probability at least $2/3$ (over the choice of (s_A, s_B, ϕ_{ref}) according to Π^S). We denote by $\mathcal{S}(F)$ the sketch size of the problem F.

It is immediate to check that $\mathcal{S}(F) \geq \mathcal{R}(F)$: since, for a communication protocol, Alice only needs to send l bits representing $s_A(x)$ to Bob, and Bob can locally compute $\phi_{ref}(s_A(x), s_B(y))$. Thus, for proving lower bounds on $\mathcal{S}(F)$, it is sufficient to prove lower bounds on $\mathcal{R}(F)$.

2.3 Fourier Analysis over $\{0,1\}^d$. We review basic properties of Fourier analysis over $\{0,1\}^d$. The set of functions $f : \{0,1\}^n \to \mathbb{R}$ is a vector space of dimension 2^n in which the inner product between two elements f and g is defined as

$$\langle f, g \rangle = \mathbb{E}_x \left[f(x) \cdot g(x) \right] = \frac{1}{2^d} \sum_{x \in \{0,1\}^d} f(x) g(x).$$

For a set $S \subseteq [d]$, define the character $\chi_S : \{0,1\}^d \to \{-1, 1\}$ as $\chi_S(x) = (-1)^{\sum_{i \in S} x_i}$. The set of all characters $\{\chi_S : S \subseteq [d]\}$ forms an orthonormal basis for the vector space. This implies that every function $f : \{0,1\}^d \to \mathbb{R}$ can be expanded uniquely as $f(x) = \sum_{S \subseteq [d]} \hat{f}(S) \chi_S(x)$ where $\hat{f}(S) = \langle f, \chi_S(x) \rangle$ is the Fourier coefficient of f with respect to set S. Moreover, for two functions $f, g : \{0,1\}^d \to \mathbb{R}$, we have that $\langle f, g \rangle = \sum_{S \subseteq [d]} \hat{f}_S \hat{g}_S$.

The notation N_ϵ stands for the random noise vector. Specifically, $N_\epsilon \in \{0,1\}^d$ is a vector of d random independent boolean values, each equal to one with probability $1/2 - \epsilon/2$.

We will make use of the *noise operator* T_ϵ (also called Bonami-Beckner operator), which, for a function $f : \{0,1\}^d \to \mathbb{R}$ is defined as $(T_\epsilon f)(x) = \mathbb{E}_{N_\epsilon} \left[f(x + N_\epsilon) \right]$. A standard fact about this operator states that, for every set $S \subseteq [d]$, $\widehat{(T_\epsilon f)}_S = \hat{f}_S \epsilon^{|S|}$.

3 Upper bound

We will first prove the main theorem of this section. We will then present the algorithmic implications of the theorem. Our main theorem concerns embedding EMD over grids $[\Delta]^d$ into ℓ_1.

THEOREM 3.1. *For any $d, \Delta \geq 1$, one can construct a randomized embedding ψ of EMD over $[\Delta]^d$ into ℓ_1 such that, for every two multi-sets $A, B \subseteq [\Delta]^d$, of size $s = |A| = |B|$, we have that*

1. $\mathbb{E}_\psi \left[\|\psi(A) - \psi(B)\|_1 \right] \leq O(\log(d\Delta)) \cdot \mathrm{EMD}(A, B)$,
2. *With probability at least $1 - 1/s$, we have:*

$$\|\psi(A) - \psi(B)\|_1 \geq \Omega\left(\frac{1}{\log s}\right) \cdot \mathrm{EMD}(A, B).$$

In addition, for a given multi-set A of size s, $\psi(A)$ is computable in $O(sd \log^{O(1)}(sd\Delta))$ time, and $\psi(A)$ is a sparse vector having only $O(s \log(d\Delta))$ non-zero coordinates.

The theorem immediately implies the low distortion embedding of EMD over $[\Delta]^d$ into ℓ_1 (by standard argument of concatenating the probabilistic embeddings).

COROLLARY 3.1. *Let $d, \Delta \geq 1$. Then there exist an embedding ψ of EMD over $[\Delta]^d$ into ℓ_1 for sets of size s that achieves distortion $O(\log s \cdot \log(d\Delta))$.*

The proof of the theorem proceeds in two steps. First we show an embedding of $[\Delta]^d$ into a *weighted Hamming metric* H_w, which is a subset of ℓ_1 of dimension $O(\log(d\Delta)$. This step can be seen as a "dimensionality reduction" in ℓ_1^d, reducing the dimension to $O(\log(d\Delta))$, but incurring a distortion of $O(\log s \cdot \log(d\Delta))$ for a size s pointset. We note that this step is randomized. In the second step, we show that EMD over this new metric H_w is $O(1)$-embeddable into ℓ_1. A trivial implementation of this step would be to use directly one of the standard embeddings of EMD over low-dimensional spaces [CCG+98], but it would incur an additional $\Omega(\log(d\Delta))$ factor distortion. We bypass this additional factor by exploiting the particular weighting of the intermediate space H_w: we embed H_w into a tree metric with distortion 2 and then we can use [Cha02] to embed EMD over trees into ℓ_1 with constant distortion.

Formally, we define a k-dimensional weighted Hamming metric H_w^k as follows. Fix some alphabet Σ. Then the metric H_w^k is over strings in Σ^k, and the distance between $x, y \in \Sigma^k$ is defined as $H_w(x,y) = \sum_{i=1}^k 2^i \cdot H(x_i, y_i)$. We denote the distance function of the metric H_w^k by $H_w(x,y)$.

The following lemma implements the first step of the embedding.

Lemma 3.1. *There exists a probabilistic embedding $\rho : [\Delta]^d \to H_w^k$, for $k = \lg(d\Delta) + 1$ such that for every $x, y \in [\Delta]^d$, it holds that*

1. $\mathbb{E}_\rho \left[H_w(\rho(x), \rho(y))\right] \le 2k \cdot \|x - y\|_1$
2. *For all $\delta > 0$, we have:* $\Pr\left[H_w(\rho(x), \rho(y)) \ge \Omega(\frac{1}{\log 1/\delta}) \cdot \|x - y\|_1\right] \ge 1 - \delta$.

Proof. The embedding ρ is into H_w^k over alphabet $\Sigma = \{0, 1, \ldots \Delta\}^d$. Each coordinate of $\rho(x)$ is a hash of x into Σ. Impose a randomly shifted cubic grid with cell side length 2^t on the space $[\Delta]^d$. Then the point x is hashed into the cell containing x. Formally, for $t \ge 1$, define a hash function on $x = (x_1, x_2, \ldots, x_d)$ to be

$$h_t(x) \triangleq \left(\left\lfloor \frac{x_1 + u_1}{2^t} \right\rfloor, \left\lfloor \frac{x_2 + u_2}{2^t} \right\rfloor, \ldots \left\lfloor \frac{x_d + u_d}{2^t} \right\rfloor \right) \tag{3.1}$$

where each $u_1, u_2, \ldots u_d$ is chosen uniformly at random from $[2^t]$. It is easy to verify that, for every $x, y \in [\Delta]^d$,

$$1 - \frac{\|x - y\|_1}{2^t} \le \Pr_{h_t}[h_t(x) = h_t(y)] \le e^{-\|x-y\|_1/2^t}.$$

Thus, we have that

$$\mathbb{E}_{h_t}\left[H(h_t(x), h_t(y))\right] \le 1 - \left(1 - \frac{\|x - y\|_1}{2^t}\right) = \frac{\|x - y\|_1}{2^t}. \tag{3.2}$$

Also, for $\|x - y\|_1 \ge \Omega(2^t \log 1/\delta)$,

$$\Pr_{h_t}[H(h_t(x), h_t(y)) = 1] \ge 1 - e^{-\|x-y\|_1/2^t} \ge 1 - \delta. \tag{3.3}$$

Finally, we define ρ as follows:

$$\rho \triangleq (id, h_1, h_2, h_3, \ldots, h_{k-1})$$

where id is the identity function $id(x) = x$, and h_t, $t = 1 \ldots k-1$, are chosen randomly as explained above.

Now, to prove property (1) of lemma, note that, by Eqn. (3.2),

$$\mathbb{E}_\rho\left[H_w(\rho(x), \rho(y))\right] = \\ \mathbb{E}_\rho\left[2H(x,y) + \textstyle\sum_{t=1}^{k-1} 2^{t+1} H(h_t(x), h_t(y))\right] \le \\ 2\|x - y\|_1 + (k-1) \cdot 2\|x - y\|_1 \le 2k \cdot \|x - y\|_1.$$

For property (2) of the lemma, note that, by Eqn. (3.3), for $\|x - y\|_1 \ge \Omega(\log 1/\delta)$, with probability at least $1 - \delta$,

$$H_w(\rho(x), \rho(y)) = 2H(x,y) + \textstyle\sum_{t=1}^{k-1} 2^{t+1} H(h_t(x), h_t(y)) \ge \\ \min\left\{d\Delta, \frac{\|x-y\|_1}{O(\log 1/\delta)}\right\} \ge \frac{\|x-y\|_1}{O(\log 1/\delta)}.$$

Also, for $\|x - y\|_1 \le O(\log 1/\delta)$, $\rho(x)$ and $\rho(y)$ differ in the first coordinate and thus $\|\rho(x) - \rho(y)\|_1 \ge 2 \ge \frac{\|x-y\|_1}{O(\log 1/\delta)}$.

The second step of the embedding results from the following lemma.

Lemma 3.2. *For any $k > 0$, EMD over H_w^k is embeddable into ℓ_1 with distortion $O(1)$.*

Proof. Suppose the metric H_w^k is over some alphabet Σ. The embedding results from the following two steps:

1. The metric H_w^k embeds into a tree metric with distortion 2 as follows. For $x = (x_1, \ldots x_k) \in \Sigma^k$, let $R(x) \triangleq (x_k, x_{k-1}, \ldots x_1)$ be the reverse of x. Then take the tree metric to be the trie on all $R(x)$, x in H_w^k, with the corresponding weights: edges at the first level are weighted by 2^k, edges at the second level are weighted by 2^{k-1}, etc. Note that all points in H_w^k correspond to leaves of the tree. We call the resulting tree metric T.
   Now, for any x, y from H_w^k, we have that $T(x,y) = \max_{x_i \ne y_i, i \in [k]} \sum_{j=1}^i 2^j = \max_{x_i \ne y_i, i \in [k]} 2^{i+1} - 1$, which is a 2-approximation to $H_w(x,y) = \sum_{i=1}^k 2^i H(x_i, y_i)$.
2. EMD over a tree metric T is $O(1)$-embeddable into ℓ_1 (cf. [Cha02]).

We are now ready to prove the theorem.

Proof. [Proof of theorem 3.1] We construct a randomized embedding ψ by composing the embeddings from Lemma 3.1, and Lemma 3.2. Let ρ be the embedding from Lemma 3.1, and μ be the embedding from Lemma 3.2. Then define $\psi(A) \triangleq \mu(\rho(A))$, where $\rho(A) \triangleq \{\rho(x) \mid x \in A\}$. We use Lemma 3.1 for $\delta = 1/s^3$. Since there are only s^2 pairs $(x,y) \in A \times B$, by union bound, with probability at least $1 - 1/s$, all pairs $(x,y) \in A \times B$ satisfy $H_w(\rho(x), \rho(y)) \ge \Omega\left(\frac{1}{\log s}\right) \cdot \|x - y\|_1$. Also, $\mathbb{E}_\rho\left[H_w(\rho(x), \rho(y))\right] \le O(\log(d\Delta)) \cdot \|x - y\|_1$ for all $x, y \in A \times B$.

Then, Lemma 3.2 implies that

1. $\mathbb{E}_\rho\left[\|\psi(A) - \psi(B)\|_1\right] \le \mathbb{E}_\rho\left[O(1) \cdot EMD_{H_w^k}(\rho(A), \rho(B))\right] \le O(\log(d\Delta)) \cdot \mathrm{EMD}(A, B)$,

2. With probability at least $1-1/s$,

$$\|\psi(A)-\psi(B)\|_1 \geq \Omega(1)\cdot EMD_{H_w^k}(\rho(A),\rho(B)) \geq \Omega\left(\frac{1}{\log s}\right)\cdot \mathrm{EMD}(A,B).$$

Note that we can compute $\rho(A)$ in $O(sd\log(d\Delta))$ time. Also, the embedding $\psi(a)=\mu(\rho(A))$ can be computed in $O(sd\log^{O(1)}(sd(d\Delta)))$ time. The embedding of [Cha02] of EMD over the tree metric T yields a vector in ℓ_1 that has at most $O(s\log(d\Delta))$ non-zero coordinates. Thus, $\psi(A)$ is computable in $O(sd\log^{O(1)}(sd\Delta))$ time and has $O(s\log(d\Delta))$ non-zero coordinates.

The above theorem has a couple further implications.

Algorithmic applications. Theorem 3.1 implies an efficient algorithm for approximating EMD between two sets $A,B\subseteq[\Delta]^d$ of size s.

Corollary 3.2. *Let $d,\Delta\geq 1$. For any two set $A,B\subseteq[\Delta]^d$ of size s, we can compute $EMD(A,B)$ up to approximation $O(\log s\cdot\log(d\Delta))$ in $O(sd\log^{O(1)}(sd\Delta))$ time.*

Together with the algorithm of, e.g., [Pan06], we obtain an efficient nearest neighbor data structure for approximation $O(\log s\cdot\log(d\Delta))$.

Corollary 3.3. *Let $d,\Delta\geq 1$. Then, for any constant $\epsilon>0$, there exist an $O(\log s\cdot\log(d\Delta))$-approximate nearest neighbor data structure for EMD over $[\Delta]^d$ that achieves $\tilde{O}(ns(d\log\Delta)^{O(1)})$ space and $\tilde{O}(n^\epsilon s(d\log\Delta)^{O(1)})$ query time.*

Embedding of the shift metric. Theorem 3.1 also yields a low-distortion embedding for *the shift metric*, a simplified version of the edit distance. Specifically, for two binary strings x and y, of length d, the shift metric is defined to be the minimum cost of transforming x into y, where the allowed transformation operations are substitutions (of cost 1), and cyclic shifts of the entire string (of cost 0). This metric is of interest because it captures the "core" of the [KR06]'s proof of $\Omega(\log d)$ non-embeddability of edit distance into ℓ_1. In particular, shift metric also has distortion of $\Omega(\log d)$ for embedding into ℓ_1 (via an argument similar to [KR06]). To obtain an upper bound on the distortion, we can construct the set A to contain all the shifts of x and the set B to contain all the shifts of y, and then apply the embedding from Corollary 3.1, obtaining the following corollary.

Corollary 3.4. *Shift metric over $\{0,1\}^d$ has an $O(\log^2 d)$-distortion embedding into ℓ_1.*

4 Lower bound

In this section, we prove a lower bound on the communication complexity on the following problem of estimating EMD over $\{0,1\}^d$ (see Section 2.2 for the relevant definitions): Alice receives a set $A\subset\{0,1\}^d$ and Bob receives a set $B\subset\{0,1\}^d$, where $|A|=|B|$, and they wish to determine whether $\mathrm{EMD}(A,B)>R$ or $\mathrm{EMD}(A,B)\leq R/\alpha$ for some threshold R and approximation ratio $\alpha\geq 1$. This promise problem is described by the partial function

$$F_{R,\alpha}(A,B)=\begin{cases}1 & \text{if } \mathrm{EMD}(A,B)\leq R/\alpha,\\ 0 & \text{if } \mathrm{EMD}(A,B)>R.\end{cases} \tag{4.4}$$

The main result of this section is stated in the following theorem.

Theorem 4.1. *For every dimension $d\geq 1$, and approximation ratio $1\leq\alpha\leq d$, there exists R such that the problem $F_{R,\alpha}$ has communication complexity $\mathcal{R}(F_{R,\alpha})\geq\Omega(\frac{d}{\alpha})$.*

This theorem implies that the sketch size for estimating EMD over $\{0,1\}^d$ is at least $\Omega(\frac{d}{\alpha})$ for any approximation factor $\alpha\leq d$. We note that in this proof, the size of the multisets is $s\leq 2^d$. Hence, in terms of s the lower bound is $\Omega(\frac{\log s}{\alpha})$, and it immediately extends to every larger d (and same s).

Before proving Theorem 4.1 we need to introduce a simpler (more restricted) model of communication, which is easier to analyze. Specifically, we consider protocols where Alice sends exactly one bit to Bob, and Bob decides on the value of the output. For such protocols, we are interested in the correctness probability, and more specifically in the *advantage* of the protocol over a random decision (i.e., difference between protocol's success probability and 1/2). The connection to this restricted model is that an efficient communication protocol for $F_{R,\alpha}$ always implies a 1-bit protocol with a "not too small" advantage (Section 4.1.1). We further show a Fourier-analytic characterization of such 1-bit protocols (Section 4.1.2). We will then show (Section 4.2) an upper bound on the advantage of any 1-bit protocol for our problem $F_{R,\alpha}$. Altogether, this will imply a lower bound on the communication complexity of general communication protocols for $F_{R,\alpha}$, proving Theorem 4.1.

4.1 Restricted Communication Protocols.

In the sequel, we consider two restricted types of communication protocols. In the first one, *one-way protocols*, only Alice can send (once) a message of length l to Bob, and Bob has to decide on the output using this message. The second type of protocols, *LSH protocols*[2], is a further restriction of one-way protocols, where, after receiving Alice's message $f(x)$, Bob computes some function $g(y)$ of his input, and outputs 1 iff $f(x) = g(y)$. As before, we will mostly consider a pair of functions (f, g) that is drawn from some joint distribution (i.e., Alice and Bob use shared randomness).

We proceed to formally define these notions and to establish connections between communication complexity of (unrestricted) protocols and the success probability of 1-bit LSH protocols.

4.1.1 One-way Protocols.

As before, $F(x, y) : \mathcal{D} \to \{0,1\}$ is a communication problem, where $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$.

DEFINITION 4.1. *A* deterministic one-way l-bit protocol *is a pair $(\mathcal{A}, \mathcal{B})$ of functions $\mathcal{A} : \mathcal{X} \to \{0,1\}^l$ and $\mathcal{B} : \mathcal{Y} \times \{0,1\}^l \to \{0,1\}$. For an input distribution μ over $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$, we define the success probability of the protocol $(\mathcal{A}, \mathcal{B})$ on μ to be* $\Pr_\mu [\mathcal{B}(y, \mathcal{A}(x)) = F(x, y)]$.

DEFINITION 4.2. *A* randomized one-way l-bit δ-error protocol *is a distribution Π over one-way deterministic l-bit protocols, such that for every input $(x, y) \in \mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$, with probability at least $1 - \delta$ the chosen protocol gives a correct answer on the input, namely,* $\Pr_\Pi [\mathcal{B}(y, \mathcal{A}(x)) = F(x, y)] \geq 1 - \delta$.

To measure the efficiency of a one-way randomized l-bit protocol, we use the success probability of the protocol, and, more specifically, the advantage of the protocol over a random guess. We'll use the following notation: $\mathrm{AdvOW}_F(l)$ is the difference between the maximum success probability of one-way randomized l-bit protocols and 1/2; i.e., $\mathrm{AdvOW}_F(l) \geq 1/2 - \delta$ iff there exists some one-way randomized l-bit δ-error protocol for the problem F. For a distribution μ over $(x, y) \in (\mathcal{X}, \mathcal{Y})$, we also define $\mathrm{AdvOW}_F(l, \mu)$ to be the difference between the maximum success probability of an l-bit deterministic protocol and 1/2.

By Yao's minimax principle [Yao83], $\mathrm{AdvOW}_F(l) = \min_\mu \mathrm{AdvOW}_F(l, \mu)$.

We now prove that if the communication complexity of problem F is $\mathcal{R}(F) \leq l$, then there exists a one-way 1-bit protocol for F with advantage $\mathrm{AdvOW}_F(1) \geq 2^{-O(l)}$. To this purpose, we prove first that if there exist a *one-way* l-bit protocol for F, then $\mathrm{AdvOW}_F(1) \geq 2^{-O(l)}$.

[2]The name comes from protocol's resemblance to the *Locality Sensitive Hashing*, a framework used for solving approximate nearest neighbor problem in high dimensions [IM98]. In particular, for the considered distance estimation problem in a metric, the LSH protocol is exactly equivalent to some LSH family of hash functions in the same metric.

LEMMA 4.1. *For any $l > 1$, if $\mathrm{AdvOW}_F(l) \geq \epsilon > 0$, then $\mathrm{AdvOW}_F(1) \geq 2^{-l}\epsilon$.*

Proof. We construct a new randomized protocol with success probability $\geq 1/2 + 2^{-l}\epsilon$. Fix $(x, y) \in \mathcal{D}$. Let $\delta = 1/2 - \epsilon$, and let Π_0 be a one-way randomized l-bit δ-error protocol.

In the new protocol, first pick $(\mathcal{A}_0, \mathcal{B}_0)$ according to the distribution Π_0. Alice and Bob jointly pick a random $s^* \in \{0,1\}^l$. Let $s = \mathcal{A}_0(x)$ be the string that Alice would send in the original protocol. If $s = s^*$, Alice sends 1, otherwise Alice sends 0. If Bob receives 1, Bob will run the original protocol on string s^*, i.e., Bob outputs $\mathcal{B}_0(y, s^*)$; otherwise Bob outputs a random output 0 or 1.

Next, we compute the probability of success of the new protocol. Let $\mu(s)$ be the distribution of $s = \mathcal{A}_0(x)$. Also, let $\eta(s) = \Pr_{\Pi_0} [\mathcal{B}_0(y, s) = F(x, y) \mid \mathcal{A}_0(x) = s]$. Note that the success probability of the old protocol Π_0 is $\sum_{s \in \{0,1\}^l} \mu(s)\eta(s) \geq 1 - \delta$.

The success probability of the new protocol is

$$\sum_{s^* \in \{0,1\}^l} 2^{-l} \left(\mu(s^*)\eta(s^*) + (1 - \mu(s^*)) \cdot \tfrac{1}{2}\right) = \frac{1}{2^l} \sum_s \mu(s)\eta(s) + \frac{1}{2}(1 - 2^{-l}) \geq \frac{1}{2^l}(1 - \delta) + \frac{1}{2}(1 - \frac{1}{2^l}) = \frac{1}{2} + \epsilon/2^l$$

LEMMA 4.2. *If $\mathcal{R}(F) \leq l$, then* $\mathrm{AdvOW}_F(1) \geq 2^{-4l}/6$.

The idea of the proof is to eliminate round by round inductively. For each round, reduce the number of communicated bits to one using the previous lemma. We defer the full proof to Appendix A.

4.1.2 LSH Protocols.

We further restrict our attention to a particular type of one-way protocols, which we call *LSH protocols*. Here, Bob simply computes a function $\mathcal{H}^B(y) : \mathcal{Y} \to \{0,1\}^l$, compares it to the value $\mathcal{H}^A(x)$ he receives from Alice, and outputs 1 iff $\mathcal{H}^A(x) = \mathcal{H}^B(y)$. Thus, in terms of a one-way protocol, Alice's function is $\mathcal{A} = \mathcal{H}^A$, and Bob's function is $\mathcal{B}(y, z) = \chi[z = \mathcal{H}^B(y)]$, where $\chi[\alpha = \beta]$ is 1 if $\alpha = \beta$ and 0 otherwise.

DEFINITION 4.3. *A* deterministic LSH l-bit protocol *is defined to be a pair of functions $(\mathcal{H}^A, \mathcal{H}^B)$, where $\mathcal{H}^A : \mathcal{X} \to \{0,1\}^l$, $\mathcal{H}^B : \mathcal{Y} \to \{0,1\}^l$. For a given input distribution μ, define the success probability of $(\mathcal{H}^A, \mathcal{H}^B)$ on μ to be* $\Pr_\mu \left[F(x, y) = \chi[\mathcal{H}^A(x) = \mathcal{H}^B(y)]\right]$.

As before, we define a randomized LSH l-bit δ-error protocol to be a distribution over deterministic protocols, where for every input (x, y) the probability of success is $\geq 1 - \delta$. For a problem F and input distribution μ, we also define $\mathrm{AdvLSH}_F(l, \mu)$ ($\mathrm{AdvLSH}_F(l)$) to be the analogue of $\mathrm{AdvOW}_F(l, \mu)$

($\mathrm{AdvOW}_F(l)$) for deterministic (randomized) LSH l-bit protocols. Again, Yao's minimax principle implies $\mathrm{AdvLSH}_F(l) = \min_\mu \mathrm{AdvLSH}_F(l, \mu)$.

Note that $\mathrm{AdvLSH}_F(l, \mu) \le \mathrm{AdvOW}_F(l, \mu)$ for any problem F and distribution μ since LSH protocols are a restriction of one-way protocols. However, when $l = 1$, the protocol is so restricted that we can also show that $\mathrm{AdvLSH}_F(1, \mu) = \mathrm{AdvOW}_F(1, \mu)$ for a specific type of μ's. We defer the proof of the following lemma to Appendix B.

LEMMA 4.3. *Let μ be a distribution on (x, y) such that for every $y_0 \in \mathrm{Supp}(\mu_y)$, $\Pr_\mu[F(x, y) = 0 \mid y = y_0] = 1/2$. Then $\mathrm{AdvLSH}_F(1, \mu) = \mathrm{AdvOW}_F(1, \mu)$.*

We can further give the following Fourier-analytic characterization of the success probability of LSH 1-bit protocols on input distributions μ of a certain structure. Recall that N_ϵ is a vector of d random independent boolean values, each equal to one with probability $1/2 - \epsilon/2$. The proof is deferred to Appendix C.

LEMMA 4.4. *Fix $f, g : \{0,1\}^d \to \{0,1\}$. Let μ_0 be the distribution where $x, y \in \{0,1\}^d$ are random and independent. Let μ_1 be the distribution where $x \in \{0,1\}^d$ is random and $y = x + N_\epsilon$. Then*

$$\frac{\Pr_{\mu_0}[f(x) = g(y)] + \Pr_{\mu_1}[f(x) \neq g(y)]}{2} = \frac{1}{2} - \sum_{S \neq \emptyset} \hat{f}_S \hat{g}_S \epsilon^{|S|}$$

4.2 Analysis of 1-bit Protocols. We are now ready to complete the proof of Theorem 4.1. In fact, we shall give a direct analysis of the advantage of 1-bit protocols, as stated in the following proposition.

PROPOSITION 4.1. *For every dimension $d \ge 1$, and approximation ratio $1 \le \alpha \le d$, there exists R, such that every randomized 1-bit protocol for $F_{R,\alpha}$ has advantage at most $2^{-\Omega(d/\alpha)}$.*

By combining this proposition together with Lemma 4.2, we immediately conclude that $\mathcal{R}(F_{R,\alpha}) = \Omega(d/\alpha)$, proving Theorem 4.1. It thus remains to prove the proposition. Our proof builds on the techniques and insights of the non-embeddability result of [KN06]; in particular, we use their code-based distribution of "hard" inputs, as well as its Fourier-analytic properties.

Proof. [Proof of Proposition 4.1] We assume that $1 < \alpha \le d/200$, since for $\alpha > d/200$ the conclusion is trivial. Fix $R = d/100$, and define $r = R/2\alpha$. Fix $C \subset \{0,1\}^d$ to be a linear code with dimension $\ge d/4$ and weight $\ge cd$, where c is a constant; for existence of such code, see e.g. [KN06, Corollary 3.5]. For $x \in \{0,1\}^d$, we denote by Gx the set $\{x + a\}_{a \in C^\perp}$ (formally, G is the set of isometries $f_g(x) = x + g$, $g \in C^\perp$, and Gx is the orbit of x induced by the group action G). In the sequel, we will only consider as inputs (for Alice and Bob) sets of the form $X = Gx$ for some $x \in \{0,1\}^d$. Notice these sets all have size $s = |C^\perp| \le 2^{3d/4}$. Furthermore, for all $X = Gx, Y = Gy$, and $y' \in Y$, we have $\mathrm{EMD}(X, Y) = \min_{x' \in X} \|x' - y'\|_1$ (see e.g. [KN06, Lemma 3.1]).

Recall that N_ϵ is a vector of d random independent boolean values, each equal to one with probability $1/2 - \epsilon/2$. Let η_0 be the uniform distribution over pairs $(x, y) \in \{0,1\}^d \times \{0,1\}^d$, and let η_1 be a distribution over pairs $(x, y) \in \{0,1\}^d \times \{0,1\}^d$ where x is random and $y = x + N_\epsilon$ for $\epsilon = 1 - 2r/d$. For $i \in \{0,1\}$, define μ_i as the distribution of (Gx, Gy) where (x, y) are picked from η_i. Since not all the pairs (x, y) in the support of μ_i satisfy $F(x, y) = i$, we define μ'_i to be the distribution μ_i conditioned on the fact that $F_{R,\alpha}(X, Y) = i$.

We next show that for all $i \in \{0,1\}$ the statistical distance between μ_i and μ'_i is $2^{-\Omega(r)}$. For (x, y) drawn from μ_1, with probability at least $1 - 2^{-\Omega(r)}$, we have $\|x - y\|_1 \le 2r$, implying that $\mathrm{EMD}(X, Y) \le 2r = R/\alpha$, or $F_{R,\alpha}(X, Y) = 1$. Similarly, for (x, y) drawn from μ_0, we can prove that for every $x' \in Gx$, the probability that $\|y - x'\|_1 > d/100$ is at least $1 - 2^{-\Omega(d)}$. Indeed, for every $x' \in Gx$, the number of points at distance $\le d/100$ from x' is at most by $\binom{d}{d/100} \le 2^{d/10}$. Thus, with probability at least $1 - s \cdot 2^{d/10}/2^d \ge 1 - 2^{\Omega(r)}$, for all $x' \in Gx$ we have $\|y - x'\|_1 > d/100$, implying that $F(x, y) = 0$.

Now let $(\mathcal{A}, \mathcal{B})$ be a deterministic one-way 1-bit protocol with maximal success probability on $\mu' = \frac{\mu'_0 + \mu'_1}{2}$. Let δ be such that this success probability is $1 - \delta$. Applying Lemma 4.3 to μ', we conclude there exists some LSH protocol $(\widetilde{\mathcal{H}^A}, \widetilde{\mathcal{H}^B})$ with success probability $1 - \delta$ on μ'. Note that μ' satisfies the precondition of the Lemma 4.3, because (i) both μ'_0, μ'_1 have marginals that are uniform (by symmetry), and (ii) after conditioning μ' on $y = y_0$, the probability that $F(x, y) = 0$ is exactly $1/2$.

Since $(\widetilde{\mathcal{H}^A}, \widetilde{\mathcal{H}^B})$ is a deterministic δ-error LSH protocol on distribution μ', and because μ_i and μ'_i are statistically close for all $i \in \{0,1\}$, we have

$$\begin{aligned} & \frac{\Pr_{\mu_0}[\widetilde{\mathcal{H}^A}(X) = \widetilde{\mathcal{H}^B}(Y)] + \Pr_{\mu_1}[\widetilde{\mathcal{H}^A}(X) \neq \widetilde{\mathcal{H}^B}(Y)]}{2} \\ \le\; & \frac{\Pr_{\mu'_0}[\widetilde{\mathcal{H}^A}(X) = \widetilde{\mathcal{H}^B}(Y)] + \Pr_{\mu'_1}[\widetilde{\mathcal{H}^A}(X) \neq \widetilde{\mathcal{H}^B}(Y)]}{2} + 2^{-\Omega(r)} \\ \le\; & \delta + 2^{-\Omega(r)} \end{aligned} \tag{4.5}$$

We define $\mathcal{H}^A, \mathcal{H}^B : \{0,1\}^d \to \{0,1\}$ as the extensions of $\widetilde{\mathcal{H}^A}$ and $\widetilde{\mathcal{H}^B}$ in the natural way: $\mathcal{H}^A(x) \triangleq \widetilde{\mathcal{H}^A}(Gx)$ and $\mathcal{H}^B(x) \triangleq \widetilde{\mathcal{H}^B}(Gx)$. By definition, for all $a \in C^\perp$ we have $\mathcal{H}^A(x) = \mathcal{H}^A(x+a)$ and $\mathcal{H}^B(x) = \mathcal{H}^B(x+a)$. Furthermore, for all $i \in \{0,1\}$, $\Pr_{\eta_i}\left[\mathcal{H}^A(x) = \mathcal{H}^B(y)\right] = \Pr_{\mu_i}\left[\widetilde{\mathcal{H}^A}(Gx) = \widetilde{\mathcal{H}^B}(Gy)\right]$, and thus we have from Eqn. (4.5) that

$$\frac{\Pr_{\eta_0}[\mathcal{H}^A(x) = \mathcal{H}^B(y)] + \Pr_{\eta_1}[\mathcal{H}^A(x) \neq \mathcal{H}^B(y)]}{2} \leq \delta + 2^{-\Omega(r)} \tag{4.6}$$

By the Fourier characterization of Lemma 4.4,

$$\frac{\Pr_{\eta_0}[\mathcal{H}^A(x) = \mathcal{H}^B(y)] + \Pr_{\eta_1}[\mathcal{H}^A(x) \neq \mathcal{H}^B(y)]}{2} = \frac{1}{2} - \sum_{S \neq \emptyset} \widehat{\mathcal{H}^A_S}\widehat{\mathcal{H}^B_S}\epsilon^{|S|} \geq \frac{1}{2} - \sum_{S \neq \emptyset} \hat{h}_S^2 \epsilon^{|S|}, \tag{4.7}$$

where $h : \{0,1\}^d \to \{0,1\}$ is equal to either $\mathcal{H}^A$ or $\mathcal{H}^B$ (the last inequality is by Cauchy-Schwarz). Since $h(x) = h(x+a)$ for all $a \in C^\perp$, we have from [KN06, Lemma 3.3] that $\hat{h}_S = 0$ for all S with $0 < |S| < w(C)$, where $w(C) \geq cr$ is the weight of the code C. Now since $\sum_S \hat{h}_S^2 \leq 1$,

$$\sum_{S \neq \emptyset} \hat{h}_S^2 \epsilon^{|S|} = \sum_{|S| \geq cd} \hat{h}_S^2 \epsilon^{|S|} \leq \epsilon^{cd} = (1-2r/d)^{cd} \leq e^{-2cr}. \tag{4.8}$$

Putting together Equations (4.6), (4.7),and (4.8), we get $\frac{1}{2} - e^{-2cr} \leq \delta + 2^{-\Omega(r)}$. We conclude that $\mathrm{AdvOW}_{F_{R,\alpha}}(1) \leq \mathrm{AdvOW}_{F_{R,\alpha}}(1, \mu') = 1/2 - \delta \leq e^{-2cr} + 2^{-\Omega(r)} = 2^{-\Omega(d/\alpha)}$, proving the proposition.

References

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

[BC03] Bo Brinkman and Moses Charikar. On the impossibility of dimension reduction in l1. *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, 2003.

[BJKS04] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *J. Comput. Syst. Sci.*, 68(4):702–732, 2004.

[CCG+98] M. Charikar, C. Chekuri, A. Goel, S. Guha, and S. Plotkin. Approximating a finite metric by a small number of tree metrics. *Proceedings of the Symposium on Foundations of Computer Science*, 1998.

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

[CK06] Moses Charikar and Robert Krauthgamer. Embedding the ulam metric into ℓ_1. *Theory of Computing*, 2(11):207–224, 2006.

[CLJ+06] M. Charikar, C. Lv, W. Josephson, Z. Wang, and K. Li. Ferret: A toolkit for content-based similarity search. In *Proceedings of ACM SIGOS EuroSys Conference*, 2006. To appear.

[CLL04] M. Charikar, Q. Lv, and K. Li. Image similarity search with compact data structures. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*, 2004.

[CM07] Graham Cormode and S. Muthukrishnan. The string edit distance matching problem with moves. *ACM Trans. Algorithms*, 3(1):2, 2007.

[GD04] K. Grauman and T. Darrell. Fast contour matching using approximate Earth Mover's Distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington DC, June 2004.

[GD05a] K. Grauman and T. Darrell. Efficient image matching with distributions of local invariant features. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, June 2005.

[GD05b] 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.

[GD06a] K. Grauman and T. Darrell. Approximate correspondences in high dimensions. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2006.

[GD06b] K. Grauman and T. Darrell. Unsupervised learning of categories from sets of partially matching image features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, New York City, NY, June 2006.

[IM98] P. Indyk and R. Motwani. Approximate nearest neighbor: towards removing the curse of dimensionality. *Proceedings of the Symposium on Theory of Computing*, pages 604–613, 1998.

[Ind07a] P. Indyk. A near linear time constant factor approximation for euclidean bichromatic matching (cost). In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2007. To appear.

[Ind07b] Piotr Indyk. Uncertainty principles, extractors, and explicit embeddings of L2 into L1. In *Proceedings of the Symposium on Theory of Computing*, 2007. To appear.

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

[KN06] S. Khot and A. Naor. Nonembeddability theorems via fourier analysis. *Mathematische Annalen*, 334(4):821–852, 2006.

[KOR98] E. Kushilevitz, R. Ostrovsky, and Y. Rabani. Efficient search for approximate nearest neighbor in high dimensional spaces. *Proceedings of the Symposium on Theory of Computing*, pages 614–623, 1998.

[KR06] Robert Krauthgamer and Yuval Rabani. Improved lower bounds for embeddings into l_1. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithm*, pages 1010–1017, 2006.

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

[LN04] J. Lee and A. Naor. Embedding the diamond graph in l_p and dimension reduction in l_1. *Geometric and Functional Analysis (GAFA)*, 14(4):745–747, 2004.

[MS00] S. Muthukrishnan and C. Sahinalp. Approximate nearest neighbors and sequence comparison with block operations. *Proceedings of the Symposium on Theory of Computing*, pages 416–424, 2000.

[OR05] R. Ostrovsky and Y. Rabani. Low distortion embedding for edit distance. In *Proceedings of the Symposium on Theory of Computing*, pages 218–224, 2005.

[Pan06] R. Panigrahy. Entropy-based nearest neighbor algorithm in high dimensions. *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2006.

[RTG00] Y. Rubner, C. Tomassi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[SS02] Michael Saks and Xiaodong Sun. Space lower bounds for distance approximation in the data stream model. In *Proceedings of the Symposium on Theory of Computing*, pages 360–369, 2002.

[Yao83] A. C-C. Yao. Lower bounds by probabilistic arguments. In *Proceedings of the Symposium on Foundations of Computer Science*, pages 420–428, 1983.

A Proof of Lemma 4.2

Proof. If $\mathcal{R}(F) \leq l$, then there must exist some protocol Π with success probability $\geq 2/3$ that is t-round protocol, where $t \leq 2l+1$ is odd, and Π is an *alternating protocol*, defined as follows. An *alternating protocol* is a protocol where all messages are of length 1 and Alice sends messages in the odd rounds and Bob in the even rounds.

We prove by induction that for every even $i \in \{0, 2, \dots t-1\}$, there exists a $(t-i)$-rounds alternating protocol Π_i with success probability at least $\frac{1}{2} + \frac{1}{6} \cdot 2^{-2i}$. The induction is on i.

For the base case notice that $\Pi_0 = \Pi$. Assume the statement for $i-2$: Π_{i-2} is a $(t-i+2)$-rounds alternating protocol with success probability at least $\frac{1}{2} + \frac{1}{6}2^{-2(i-2)}$. Construct the protocol Π_i as follows. In the last two rounds from Π_{i-2}, Bob sends a bit to Alice, and then Alice sends a bit to Bob. Note that, these two rounds could be replaced by a round where Alice just sends two bits: one for each possible bit coming from Bob. Thus, there exists a $(t-i+1)$-rounds protocol Π' which has first $t-i$ rounds exactly the same as in Π_{i-2}, and in the $(t-i+1)^{th}$ round, Alice sends two bits. Combining round $t-i$ and round $t-i+1$, both with messages from Alice to Bob, we get a $(t-i)$-rounds protocol Π'' with same success probability as Π_{i-2} and with first $t-i-1$ rounds having messages of length 1, and the $(t-i)^{th}$ message of length 3. Apply Lemma 4.1 to round $t-i$ of Π'' to obtain a $(t-i)$-rounds protocol Π_i with message lengths 1 and success probability at least $\frac{1}{2} + \frac{1}{6}2^{-2(i-2)} \cdot 2^{-3} > \frac{1}{2} + \frac{1}{6}2^{-3i}$. Note that Π_i is a $(t-i)$-rounds alternating protocol.

In the end, Π_{t-1} is exactly a one-way 1-bit protocol. Since Π_{t-1} has success probability at least $\frac{1}{2} + \frac{1}{6}2^{-2t}$, we conclude that $\mathrm{AdvOW}_F(1) \geq \frac{1}{6}2^{-2t} \geq 2^{-4l}/6$.

B Proof of Lemma 4.3

Proof. Let $\epsilon = \mathrm{AdvOW}_F(1, \mu)$, $\delta = 1/2 - \epsilon$, and let $(\mathcal{A}, \mathcal{B})$ be the deterministic protocol realizing the success probability $1 - \delta$.

We argue that there exists a one-way protocol $(\mathcal{A}, \mathcal{B}')$ with success probability $\geq 1 - \delta$ on μ, where $\mathcal{B}'(y, 0) \neq \mathcal{B}'(y, 1)$ for every $y \in Supp(\mu_y)$. We construct $\mathcal{B}'$ by taking $\mathcal{B}$ and modifying the function on all y's for which $\mathcal{B}(y, 0) = \mathcal{B}(y, 1)$. Let y_0 be such that $\mathcal{B}(y_0, 0) = \mathcal{B}(y_0, 1)$. Let $A_0 = \mathcal{A}^{-1}(0) \subseteq \mathcal{X}$ and $A_1 = \mathcal{A}^{-1}(1) \subseteq \mathcal{X}$. Next, let $a = \Pr_\mu[\mathcal{A}(x) = F(x, y_0) \mid y = y_0]$ and $b = 1 - a = \Pr_\mu[\mathcal{A}(x) \neq F(x, y_0) \mid y = y_0]$. If $a \leq 1/2$ then set $\mathcal{B}'(y_0, 0) = 0$ and $\mathcal{B}'(y_0, 1) = 1$, and, if $b < 1/2$, then $\mathcal{B}'(y_0, 0) = 1$ and $\mathcal{B}'(y_0, 1) = 0$.

Such modification of the protocol for an y_0 increases the error by

$$\begin{aligned} & \Pr_\mu[y = y_0] \cdot (\Pr[\mathcal{B}'(y_0, \mathcal{A}(x)) = F(x, y_0) \mid y = y_0] \\ & - \Pr[\mathcal{B}(y_0, 0) = F(x, y_0) \mid y = y_0]) \\ = \; & \Pr_\mu[y = y_0] \cdot \left(\min\{a, b\} - \tfrac{1}{2}\right) \\ \leq \; & 0 \end{aligned}$$

After doing as above for all y_0 where $\mathcal{B}(y_0, 0) = \mathcal{B}(y_0, 1)$, take the LSH protocol to be $\mathcal{H}^A(x) \triangleq \mathcal{A}(x)$ and $\mathcal{H}^B(y) \triangleq \mathcal{B}'(y, 1)$. Since this LSH protocol has success probability at least $1 - \delta$, we conclude that $\mathrm{AdvLSH}_F(1, \mu) \geq \mathrm{AdvOW}_F(1, \mu)$.

C Proof of Lemma 4.4

Proof. Let $p_f \triangleq \mathbb{E}_x[f(x)] = \hat{f}_\emptyset$ and $p_g \triangleq \mathbb{E}_y[g(y)] = \hat{g}_\emptyset$. If we set $\bar{f}(x) \triangleq 1 - f(x)$, then note that $\hat{\bar{f}}_\emptyset = 1 - p_f$ and $\hat{\bar{f}}_S = -\hat{f}_S$ for $S \neq \emptyset$ (analogously for g).

We can then write that

$$\Pr_{\mu_0}[f(x) = g(y)] = p_f \cdot p_g + (1 - p_f) \cdot (1 - p_g)$$

and

$$\Pr_{\mu_1}[f(x) \neq g(y)] = p_f \Pr_{\mu_1}[g(y) = 0 \mid f(x) = 1] + (1-p_f) \Pr_{\mu_1}[g_y = 1 \mid f(x) = 0].$$

If we set $a \triangleq \Pr_{\mu_1}[g(y) = 0 \mid f(x) = 1]$, then we have that

$$a = \Pr_{\mu_1}[\bar{g}(y) = 1 \mid f(x) = 1] = 2^d \cdot \langle \frac{1}{|f^{-1}(1)|} T_\epsilon f, \bar{g} \rangle = \frac{1}{p_f} \sum_S \hat{f}_S \hat{\bar{g}}_S \epsilon^{|S|} = \frac{1}{p_f} \left(p_f(1-g_f) - \sum_{S \neq \emptyset} \hat{f}_S \hat{g}_S \epsilon^{|S|} \right).$$

Similarly, define b as

$$b \triangleq \Pr_{\mu_1}[g(y) = 1 \mid f(x) = 0] = \frac{1}{1-p_f} \left((1-p_f)g_f - \sum_{S \neq \emptyset} \hat{f}_S \hat{g}_S \epsilon^{|S|} \right).$$

Thus,

$$\begin{aligned}
&\Pr_{\mu_0}[f(x) = g(y)] + \Pr_{\mu_1}[f(x) \neq g(y)] = \\
&= \quad p_f g_f + (1-p_f)(1-g_f) + p_f a + (1-p_f) b \\
&= \quad p_f p_g + (1-p_f)(1-p_g) + p_f(1-g_f) + (1-p_f) g_f - 2 \sum_{S \neq \emptyset} \hat{f}_S \hat{g}_S \epsilon^{|S|} \\
&= \quad 1 - 2 \sum_{S \neq \emptyset} \hat{f}_S \hat{g}_S \epsilon^{|S|}.
\end{aligned}$$