Algorithms for Massive Data
Final Project

Yufei Guo* Walter McKelvie† Manuel Paez‡

December 21, 2023

*yg2892@columbia.edu. yg2892
†w.mckelvie@columbia.edu
‡map2332@columbia.edu. map2332

Contents

1 Introduction

This paper will describe the research for sketching and embedding Earth-Mover's Distance (EMD). For this, we will introduce the methods, procedures, and algorithms from different papers surrounding this question. Firstly, we will describe the methods laid out in [IT03] and [Ind07], which sketch EMD into the l_1 norm. Lastly, we will finish off by describing [ABIW09], which modifies the methods from the previous two papers. Afterward, we will describe the work our group did for sketching Chamfer Distance (CH), a distance described as the relax-Earth Mover's Distance. We will describe our motivation for sketching Chamfer Distance by describing the algorithm for [BIJ+23], a recent paper that described a near-linear time approximation algorithm for Chamfer Distance. We will show our group's attempt to adapt several of the methods used for sketching EMD for Chamfer Distance and how these techniques fail for Chamfer Distance, and in advance provide a similar algorithm to calculate Chamfer Distance. We will also discuss communication complexity lower bounds for Chamfer Distance assuming equal cardinality between two sets.

2 Embedding and Sketching for EMD

We will define Earth-Mover Distance as follows:

Definition 2.1 (Earth-Mover Distance)**.** *Given two (multi)sets A, B in a metric space, and $|A| = |B|$. $EMD(A, B)$ is the cost of the minimum matching between them.*

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

where π is the bijective mapping.

Earth Mover's Distance is a metric. This is important as the algorithms and methods for sketching and embedding EMD that we will be describing exploit the properties of a metric.

2.1 Embed EMD into ℓ_1

In this section, we will describe the methods used in [IT03] and [Ind07] for embedding EMD. We start by describing a random mapping f such that when the multisets A, B are subsets on the grid $[\Delta]^d$, with high probability we have

$$\|f(A) - f(B)\|_1 \leq EMD(A, B) \leq O(\log \Delta)\|f(A) - f(B)\|_1$$

As π is a bijective mapping, it requires the subsets to be equal to each other. To cover the case where A, B are different sizes, an extension of Earth-Mover's Distance is also given (over $[\Delta]^2$).

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

i.e., we need to pay a cost of Δ for each unmatched point. EEMD is also shown to satisfy the metric and norm properties, however, we will not show the proof here. We should also note that EEMD will be used in the algorithms and methods we will later discuss.

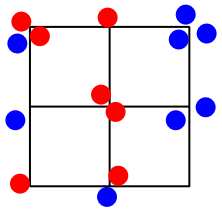

Figure 1: Points on small grids

2.1.1 Algorithm

We start by describing a mapping $\psi : S \mapsto n^{[\Delta]^2}$ such that f counts the number of points on each location on the $[\Delta]^2$ grid. We can describe f as the characteristic function. e.g., in 1, if red points represent set A and blue points represent set B, we have

$$f(A) = (2, 1, 0, 0, 2, 0, 1, 1, 0)$$
$$f(B) = (1, 0, 3, 1, 0, 2, 0, 1, 0)$$

On a small grid of constant side lengths, if two points do not coincide, they have at most constant distance. Then, when representing the sets using characteristic function f, these two points must be in different coordinates and thus the characteristic function of the two sets differ by 2, and gives the following fact.

Fact 2.2. *If grid size are constant (e.g., 2),* $\|f(A) - f(B)\|_1 = O(1) \cdot EEMD(A, B)$

With this fact, we can develop a higher-level visualization of the grid argument: we can recursively divide the large grid of $[\Delta]^2$ into levels of smaller grids (of grids) such that we can represent each level as many constant-size small grids.

Algorithm 1: Algorithm to embed EMD into ℓ_1 [IT03]

Input: set $A \subset [\Delta]^d$

for $t = 1, \cdots, O(\log \Delta)$ **do**

  Create a grid G^t of side length 2^t, randomly shift it by $[0, \Delta]$, count the number of points in each cells. Call this vector v_A^t.

end

After obtaining $v_A^1, \cdots, v_A^{\log \Delta}$, we concatenate them together to get $(2 \cdot v_A^1, 4 \cdot v_A^2, \cdots, 2^t v_A^t)$. We will call this $\psi(A)$ and output it.

The obtained concatenated vector $\psi(A)$ in the algorithm is the embedding into ℓ_1. We can use $\psi(A)$ and $\psi(B)$ to approximate $EMD(A, B)$

2.1.2 Notations from [Ind07]

We will now define the notation that will be used for our analysis of the algorithm. Let A_c be the set of points that fall into the same grid cell c after applying a grid G on set A. For a grid with side length m, we have at most $\left\lceil \frac{\Delta}{m} \right\rceil + 1$ such cells. Let $G^i(A)$ be the (multi)set after applying the grid G^i. i.e., we merge each cell.

2.1.3 Analysis

We first discuss the intuition behind this algorithm and we will prove that its distortion is at most $O(\log \Delta)$. For this, we will ignore the random shift that is part of the algorithm, although it will be useful later. If we partition the grid $[\Delta]^2$ into several parts, then a bipartite matching of the sets becomes the summation of

1. bipartite matching within each grid part
2. bipartite matching across different grid parts, due to unequal number of points in each part.

To achieve this, we recursively partition each grid part, where each part is of constant sizes. We introduce EEMD a modified version of EMD, to take into account the unequal number of points in one part. By 2.2 we know $\left\| v_A^{\log \Delta - 1} - v_A^{\log \Delta - 1} \right\|_1$ gives a constant approximation of EEMD between different cells. Formally, we have the following

Claim 2.3. $EMD(A, B) \leq O(1)\|\psi(A) - \psi(B)\|_1$

Proof.

$$
\begin{aligned}
& EMD(A, B) && (1)\\
&= EEMD_\Delta(A, B) \quad \text{Since } |A| = |B| && (2)\\
&\leq \sum_c EEMD_{\frac{\Delta}{2}}(A_c, B_c) + \frac{\Delta}{2} EEMD_2(G^1(A), G^1(B)) && (3)\\
&\leq \sum_c \left(\sum_d EEMD_{\frac{\Delta}{4}}(A_{cd}, B_{cd}) + \frac{\Delta}{4} EEMD_4(G^2(A_{cd}), G^2(B_{cd})) \right) + \frac{\Delta}{2} EEMD_2(G^1(A), G^1(B)) && (4)\\
&\leq O(1)\|\psi(A) - \psi(B)\|_1 && (5)
\end{aligned}
$$

The proof is simple since we can give a bipartite matching induced by line 3, which is composed of two parts:

1. For matchings within each grid cells, they satisfy the properties of a matching.
2. For matching over several cells, they cover the unmatched pairs separated by cells. Since $EEMD$ already has a cost for unmatched points, the bipartite matching induced will be at least the actual matching distance between them.

With this, we can always construct an actual bipartite matching from the induced matching. We can recursively process each subgrid to obtain the actual matching to get line 4, which allows us to complete the proof. □

Proving the lower bound for the embedding does not require randomness over the grids, however, we will need randomly shifted grids to obtain an upper bound for EMD embedding into the grid. With this, we have the following lemmas.

Lemma 2.4.

$$\mathbb{E}\left[\sum_c EEMD_m(A_c, B_c) \right] \leq 2 \cdot EEMD_\Delta(A, B)$$

Lemma 2.5.

$$\mathbb{E}\left[\frac{\Delta}{k} EEMD_k(G(A), G(B))\right] \leq EEMD_\Delta(A, B)$$

Proof. We will construct this proof such that we prove both lemma 2.4 and lemma 2.5.
Since $EEMD$ is a metric, WLOG, assume that $|A| \leq |B|$. Now we fix some certain matching M from A to $S' \subset B$. Then

$$EEMD_\Delta(A, B) = EMD_\Delta(A, S) + \Delta\left(|B| - |S|\right)$$

The extra cost that we could pay from this estimation arises from separating close points. After applying a grid, close points will separate into different grid parts, making the induced matching more expensive. Another aspect to the extra cost is if an edge of length k in M is not cut, it will add k to the estimated cost. If it is cut, it will add $2 \cdot m$ unmatched cost to the final result. An edge with length k has at most $\frac{k}{m}$ probability of getting cut, and thus in expectation, it will add k to the total cost. With this, we can obtain the inequality

$$\mathbb{E}\left[\sum_c EEMD_m(A_c, B_c)\right] \leq 2 \cdot EEMD_\Delta(A, B)$$

□

By the equation above, we can apply lemma 2.5 $\log(\Delta)$ times to obtain the following claim below, which is based on the randomness of the grid.

Claim 2.6.

$$\mathbb{E}[\|\psi(A) - \psi(B)\|_1] \leq O(\log \Delta) EMD(A, B)$$

2.2 Efficient Sketches for EMD

In this section, we will survey the methods, procedures, and algorithms for the paper [ABIW09]. Similar to the papers on sketching and embedding EMD described previously, we will be working on the $[\Delta]^2$ grid. The previous papers (with the main results described above), obtained mappings f into l_1 where given the sets $A, B \subseteq [\Delta]^2$ and for some $C > 0$, we have

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

However, these papers primarily represented EMD as vectors into l_1 obtains a distortion of at least $\Omega(\sqrt{\log \Delta})$. The goal of this paper is to find an estimate for $EMD(A, B)$ from maps $F(A), F(B)$ using an efficient algorithm E. This paper diverges from the methods of the previous papers by providing a construction of mappings into spaces other than l_1. Particularly, this paper shows one can map sets into spaces with dimensions sublinear in Δ. For this paper, we will use this definition of EMD. which is only a slight modification of the one used previously:

Definition 2.7. *(EMD definition for [ABIW09])*
Given multi-sets A, B of points in $[\Delta]^2, |A| = |B| = N$

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

The formal outline of the goal is as follows:

Theorem 2.8. *For any $\epsilon \in (0,1)$, there exists a distribution over linear mappings $F : \mathbb{R}^{\delta^2} \to \mathbb{R}^{\delta^\epsilon}$ such that for multisets $A, B \subset [\Delta]^2$ of equal size, one can produce a $O(1/\epsilon)$-approximation to $EMD(A,B)$ from $F(A), F(B)$ using an estimator function E. Particularly, we have*

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

with probability 2/3.

Since this paper uses the EEMD norm defined in the previous papers, we should also note this notable fact from [Ind07] for the decomposition EEMD norm:

Fact 2.9. *For any $0 < \epsilon < 1$, there exists a distribution over n-tuples of linear mappings $\langle F_1, ..., F_n \rangle$ for $F_1 : \mathbb{R}^{\Delta^2} \to \mathbb{R}^{m^2}$ with $m = \Delta^\epsilon$ such that for any $x \in \mathbb{R}^{\Delta^2}$ we have*

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

2.2.1 Main Method: Summation of Norms

The main method introduced by this paper is to create a linear sketch of the summation of norms. Unlike the previous papers, which create a decomposition of a modified EMD (known as EEMD) into a weighted sum of small modified EMD on the $[\Delta]^2$ grid, this paper seeks to create an $O(1)$-approximation to $\|x\|_{1,X}$. For this, denote $\|x\|_{1,X} = \sum_i \|x\|_X$. We will be using the $EEMD$ norm as this modified norm. The summation of the norm is as follows:

Theorem 2.10. *Given $n \in \mathbb{N}, M > 0$ and $\lambda > 1$, there exists a distribution over random linear mappings*

$$\mu : X^n \to X^{(\lambda \log n)^{O(1)}}$$

and a reconstruction algorithm $\mathcal{R}$ such that $\forall x \in X^n$ satisfying

$$M/\lambda \leq \|x\|_{1,X} \leq M$$

the algorithm $\mathcal{R}$ produces an $O(1)$ approximation to $\|x\|_{1,X}$ from $\mu(x)$ with high probability.

This theorem is important; it gives us a mapping to obtain an estimation M of $\|x\|_{1,EEMD}$ with the approximation factor $\lambda = O(\log \Delta)$. With this, the main idea for constructing the mapping of the summation of norms is as follows. [ABIW09] does an exponential-level partition of the elements of a metric space. Assuming that $\|M\|_{1,X} \leq C$, they define the level $k \in \mathbb{N}$ to be

$$S_k = \{i \in [n] s.t. \|M_i\|_X \in (T_k, 2T_k]\}, T_k = C/2^k$$

With this, they sufficiently estimate $|S_k|$ at each level k. To obtain the estimation, at each level k, they create t hash tables. For each hash table, the subsample from $[n]$ with probability p_k with each cell maintaining the sum of x_i's that hash to it. For the recovery algorithm, they count the number of accepting hash tables, which is denoted as c_k. With this, they return the quantity $\sum_k T_k \cdot (c_k/t) \cdot (1/p_k)$

2.2.2 Algorithm

We will combine the several methods to build the algorithm for sketching EMD. The sketch F, obtained from which from Theorem 2.10 and Fact 2.9, is as follows:

1. Create a linear map f of planar EMD into l_1 that approximates the EMD distance up to $\lambda = O(\log \Delta)$
2. Create a collection of $O(\log \Delta)$ sketches of the linear maps $v_i = \mu_i \circ F^{(i)}$

With these points, we will obtain $F = \langle f, \mu_1 \circ F^{(1)}, ...\mu_{\log \Delta} \circ F^{(log\Delta)} \rangle$. We can use F to construct the algorithm for sketching EMD. Given sketches $Fx(A)$ and $Fx(B)$, compute the $\lambda = O(\log \Delta)$ approximation using the linear map f. Then, use the map $v_i = \mu_i \; F^{(i)}$ to compute the estimate

$$\sum_{j=} \|F_j^{(i)}(x(A) - x(B))\|_{EEMD}$$

which is an $O(1/\epsilon)$ approximation to $EMD(A, B)$ by Fact 2.9. With this, [ABIW09] provides a construction for mappings that are sublinear in Δ.

3 Approximating Chamfer Distance

In this section, we will introduce Chamfer Distance, a distance that has been described as a relaxation of Earth-Movers Distance. Chamfer Distance is a popular measure of dissimilarity between point clouds, used for computer vision and machine learning problems. We will describe a breakthrough on approximating Chamfer Distance. Afterward, we will describe our group's work on sketching Chamfer Distance and the communication complexity lower bounds of Chamfer Distance.

Definition 3.1 (Chamfer Distance)**.** *For two multi-sets A and B such that $|A|, |B| \leq n, A, B \subseteq \mathbb{R}^d$. The Chamfer Distance from A to B is defined as*

$$CH(A, B) \coloneqq \sum_{a \in A} \min_{b \in B} d_X(a, b)$$

where d_X is the metric, e.g. the l_1 or l_2 norm.

3.1 Estimating Chamfer Distance via sampling

In a recent breakthrough, [BIJ+23] presents the first $(1 \pm \epsilon)$-approximation algorithm for Chamfer Distance that runs in $O(nd \log(n)/\epsilon^2$ where the underlying metric is defined by for the l_1 and l_2 norm. This is formally described in the following theorem:

Theorem 3.2. *(Estimating Chamfer Distance in Nearly Linear Time) Given as input two datasets $A, B \subset \mathbb{R}^d$ such that $|A|, |B| \leq n$ and a accuracy parameter $0 < \epsilon < 1$. An Algorithm to estimate Chamfer Distance runs in time $O(nd \log(n)/\epsilon^2)$ and outputs an estimator η such that given the underlying metric is ℓ_1 or ℓ_2 and with probability at least 99/100, we have*

$$(1 - \epsilon)CH(A, B) \leq \eta \leq (1 + \epsilon)CH(A, B)$$

The paper describes a main algorithm and a subroutine algorithm for estimating the Chamfer Distance. In their approach, they use importance sampling instead of uniform sampling. Uniform sampling for Chamfer Distance would be described as the following: sample an $a \in A$ uniformly at

random and explicitly compute $\min_{b\in B} \|a-b\|_1$. There is a main problem to this approach: if a small fraction of elements affects $CH(A,B)$ significantly, then $s = \Omega(n)$ samples to compute $CH(A,B)$ with 99% probability of correctness. As each sample requires a linear-time scan to find the nearest neighbor, this results in a quadratic run-time. Given the possibility that the distributions of the distances from points in A to their nearest neighbors in B can be skewed, this would not be the recommended approach.

Instead of the uniform sampling approach, [BIJ+23] uses importance sampling, which is described as the following: sample $a \in A$ with probability proportional to $CH(A,B)$. With importance sampling, one can create an estimator for $CH(A,B)$ with estimates of the distribution D_a over elements $a \in A$ such that

$$\min_{b\in B} \|a-b\|_1 \leq D_a$$

Algorithm 2: Nearest-Neighbor Subroutine(A,B) [BIJ+23]

Input: Given two subsets $A, B \subset \mathbb{R}^d$ of size at most n where all non-zero distances between any point in A and any point in B is between 1 and $poly(n/\epsilon)$

We instantiate $L = O(\log(n/\epsilon)$ and for $i \in \{0,...,L\}$, we let $r_i = 2^i$
for $i \in \{0,...,L\}$ **do**
  sample a hash function $h_i : X \to U$ from $h_i \sim \mathcal{H}(r_i)$
end
for $a \in A$ **do**
  find the smallest $i \in \{0,...,L\}$ for which there exists a point $b \in B$ with $h_i(a) = h_i(b)$ and set $D_a = \|a-b\|_1$
end
return *A list of numbers* $\{D_a\}_{a\in A}$ *where* $D_a \geq \min_{b\in B} \|a-b\|_1$

Algorithm 3: Estimating Chamfer Distance [BIJ+23]

Input: Two subsets $A, B \subset \mathbb{R}^d$ of size at most n and $T \in \mathbb{N}$

Execute a Nearest-Neighbor Subroutine Algorithm to obtain $\{D_a\}_{a\in A}$ which satisfy $\min_{b\in B} \|a-b\|_1 \leq D_a$. Let $D = \sum_{a\in A} D_a$ Construct a probability distribution $\mathcal{D}$ which satisfies that for every $a \in A$
$Pr_{x\sim\mathcal{D}}[x = a] = \frac{D_a}{D}$ **for** $l \in [T]$ **do**
  sample $x_l \sim \mathcal{D}$ and spend $O(|B|, d)$ time to compute $\eta_l = \frac{D}{D_{x_l}} \cdot \min_{b\in B} \|x_l - b\|_1$
end
return $\eta = \sum_{t=1}^{T} \eta_l$

With this, [BIJ+23] uses the subroutine algorithm to obtain a family of distributions $\{D_a\}_{a\in A}$ over the elements in A. The main algorithm then uses the family of distributions to create a reasonable estimator for the $CH(A,B)$. The subroutine algorithm, Nearest-Neighbor Subroutine(A,B), and the main algorithm, Estimating-Chamfer(A,B), are outlined above.

3.2 Impossibility to embed Chamfer into ℓ_1

The remainder of the paper will discuss work done to possibly sketch Chamfer Distance. As noted before, the algorithms used to sketch Earth Mover's Distance relied on the fact that Earth Mover Distance satisfies the metric and norm properties. Unlike EMD, however, Chamfer Distance is not a metric as it does not satisfy the following metric properties:

- Symmetry: $CH(A,B) \neq CH(B,A)$
- Triangle Inequality: For $A,B,C \subseteq \mathbb{R}, CH(A,B) \nleq CH(A,C) + CH(C,B)$

With this fact, using the same methods and algorithms for sketching EMD for Chamfer Distance might not be possible without significant modification. One such method is embedding into a different metric space. Although EMD can be embedded into another metric space e.g. ℓ_1, ℓ_2, it is impossible to embed Chamfer Distance into ℓ_1 with reasonable distortion. This means the method and algorithm we surveyed in section 2.1 and [IT03] to embed EMD into ℓ_1 cannot be done for Chamfer Distance with a distortion such that sketching via embedding is unreasonable. We will construct our proof by considering the Symmetric Chamfer Distance, which is defined in the following theorem

Theorem 3.3. *Let the metric space be $(X, d_X(\cdot,\cdot))$, where X is of dimension d, and $d_X(\cdot,\cdot)$ is a metric. And suppose we have $A,B,C \subseteq X$ of size at most s. For any d, to construct a ψ that $\|\psi(A) - \psi(B)\|_1$ is an approximation of $C(A,B)$ where $C(A,B) = CH(A,B) + CH(A,B)$ must have a distortion $\Omega\left(s^{O(1)}\right)$*

Proof. By triangle inequality of ℓ_1 norm, we must have

$$\|\psi(A) - \psi(C)\|_1 \leq \|\psi(A) - \psi(B)\|_1 + \|\psi(B) - \psi(C)\|_1$$

Consider A, C being two small balls $B_{r_1}(a), B_{r_2}(c)$, and construct $B = \{a, c\}$. Now $CH(A,B) \leq s \cdot r_1$ and $CH(C,B) \leq s \cdot r_2$. Then,

$$C(A,B) \leq s \cdot r_1 + d_X(a,c), C(B,C) \leq s \cdot r_2 + d_X(a,c)$$

$$CH(A,C) \leq s \cdot (d_X(a,c) - r_1 - r_2)$$

For $\|\psi(X) - \psi(Y)\|_1$ to be a α multiplicative approximation of $C(X,Y)$, we need

$$2\alpha s\left(d_X(a,c) - r_1 - r_2\right) \leq sr_1 + d_X(a,c) + sr_2 + d_X(a,c)$$

Which gives $\alpha = \frac{1}{s^{\Theta(1)}}$ □

3.3 Modification of embedding

As mentioned above, it is impossible to embed (symmetric) Chamfer distance into ℓ_1 with proper distortion. However, we can give a slight modification to the algorithm described in [IT03] and use it to approximate Chamfer distance with the same distortion. For this, we will work in $[\Delta]^2$, a discrete planar grid, and similar to EMD, we can transform points in continuous space $\mathbb{R}^2$ into this grid with small distortions [Ind07]. Similar to the $EEMD$, which extends EMD to two sets of different size, we consider an extension ECH of the normal Chamfer Distance to cover some edge cases where the target (multi)set is an empty set.

$$ECH_\Delta(A,B) = CH(A,B) + \Delta \cdot |A| \cdot \chi\left[|B| = 0\right]$$

Where $\chi\left[|B| = 0\right]$ indicates the event if set B is empty or not, i.e., if we wish to map a set A to an empty set B, we must spend extra cost to move all the points in A to another grid.
We use the same f as mentioned earlier in the EMD embedding algorithm. Additionally, we create g from f, such that

$$g(A)_i = 0 \Leftrightarrow f(A)_i = 0$$

i.e., make characteristic function binary. Instead of directly calculating ℓ_1 norm, we calculate

$$\langle f(A), (g(A) - g(B))\rangle = |A| - \langle f(A), g(B)\rangle$$

Fact 3.4. *If grid size are constant (e.g., 2)*

$$\langle f(A), (g(A) - g(B))\rangle = |A| - \langle f(A), g(B)\rangle = O(1) \cdot ECH(A, B)$$

Proof. If grid sizes are constant, the Chamfer Distance from A to B is the number of points in A that does not coincide with any points in B, multiplied by a constant distance. If B is empty, the extra cost of this matching is simply $O(|A|)$. □

This approximation scheme allows us to give an algorithm to estimate Chamfer Distance. By similar reasoning done in 2.1.3, we have the following:

$$CH_\Delta(A, B) = ECH_\Delta(A, B) \tag{1}$$

$$\leq \sum_i \frac{\Delta}{2^i} ECH(G^i(A), G^i(B)) \tag{2}$$

$$= \sum_i \frac{\Delta}{2^i} \left(|A| - \langle v_A^i, u_B^i\rangle\right) \tag{3}$$

where v_A^i is the number of points in each cell after applying grid G^i and u_B^i is the binary vector indicating if there's a point in each grid cell (i.e., $g(G^i(B))$). With this, we can prove the following bound:

$$\mathbb{E}[\|\psi(A) - \psi(B)\|_1] \leq O(\log \Delta) EMD(A, B)$$

Proof. Using the same justification in the previous proof for EMD, we can find a matching from A to B and fix that matching such that an edge of length k in the matching will be cut with probability at most $\frac{k}{m}$ (Assume grids have side length m). With this, either the cut point has no point in B, which by definition of ECH will incur a m cost, or it will be matched to some points in its cell with a cost at most m. With this, we obtain the inequality $\mathbb{E}[\|\psi(A) - \psi(B)\|_1] \leq O(\log \Delta) EMD(A, B)$, which is similar to the inequality proven in the EMD case. □

Although this algorithm shows that Chamfer Distance and Earth-Mover's Distance share the same distortion, we should note that as soon as all points in A have some point in B in their grid cell, all of the later (grids of larger side lengths) terms will be 0. With this, we will not apply the above lemma $\log \Delta$ times if the Chamfer Distance is far from the Earth-Mover's Distance. Additionally, storing v_A^i and u_B^i has smaller bit complexity.

3.4 Communication Complexity of Chamfer Distance: Impossibility

In this section, we will show the communication complexity lower bounds of Chamfer Distance. Specifically, we will define the Indexing and Disjoint Problem and show the possibility of a sublinear sketch of Chamfer Distance for both problems.

Definition 3.5. *Indexing Problem*
Alice has $x \in \{0, 1\}^n$ and Bob has $i \in [n]$. Alice sends Bob k bits (one-way) and Bob has to output x_i.

Definition 3.6. *Disjoint (DISJ) Problem*
Alice has $A \subseteq [n]$, Bob has $B \subseteq [n]$. They can talk to each other and want to decide if $A \cap B =$

With the communication complexity problems now defined, we can show the following theorems:

Theorem 3.7. *Indexing requires $k = \Omega(n)$ bits*

Theorem 3.8. *Disjoint requires $\Omega(n)$ communication complexity.*

We will now show the possibility of a sublinear sketch for Chamfer Distance through both arguments.

Theorem 3.9. *Let X be a metric space that contains an infinite number of elements. Then there is no sublinear sketch for for Chamfer distance on X.*

Proof. Because Chamfer distance is not commutative, we must separately argue that both the first and second arguments do not have sublinear sketches. First, suppose that we could sublinearly sketch the first argument; i.e., that there exists a sketch function g, and a function F, such that $F(g(S_1), S_2) = (1 \pm \epsilon) \cdot \mathrm{CH}(S_1, S_2)$ and $g(S_1)$ has sublinear size in $|S_1|$. Label a countably infinite subset of X by the natural numbers. From the above two functions, we will show that a sublinear algorithm for DISJ is possible. The reduction is as follows: Alice has $A \subseteq [n]$, Bob has $B \subseteq [n]$. Then,

- Alice sends $g(A)$ to Bob
- Bob uses $g(A)$ determine whether $CH(A, B^C) = 0$, outputting 1 if so.

This has sublinear communication complexity because $|A| \leq n$. Correctness comes from $CH(A, B^C) = 0 \iff A \subseteq B^C \iff A \cap B = \emptyset$. Hence, we get a contradiction. Now suppose that we could sublinearly sketch the second argument; i.e., that there exists a sketch function g, and a function F, such that $F(S_1, g(S_2)) = (1 \pm \epsilon) \cdot \mathrm{CH}(S_1, S_2)$ and $g(S_2)$ has sublinear size in $|S_2|$. We will give a reduction to the indexing problem. We could solve the indexing problem in sublinear communication as follows:

- Given $x \in \{0, 1\}^n$, Alice takes $A = \{j : x_j = 1\}$, sends Bob $g(A)$.
- Bob takes $B = \{i\}$, uses $g(A)$ to estimate $CH(B, A)$. If estimate is zero, output 1; otherwise output 0.

Note that $|A| \leq n$. If Bob gets a $(1 \pm \epsilon)$-approximation of real Chamfer distance, he will get zero if and only if the real distance if 0. But the real distance is zero $\iff i \in A \iff x_i = 1$. So Bob has indexed with sublinear communication complexity, a contradiction. □

Since the above reductions depend on very different cardinalities of S_1 and S_2, it is natural to wonder whether a sketch is possible given the constraint $|S_1| = |S_2|$. We give a weaker result below which encompasses many common metric spaces (in particular $\mathbb{R}^n$), although notably not lattices.

Corollary 3.10. *Let X be a a metric space. Suppose that X has an accumulation point p, and that there exists a countably infinite subset $Y \subseteq X$ (where and an $\alpha > 0$ so that $d(p, Y) > \alpha$ and any two distinct points $x, y \in Y$ satisfy $d(x, y) > \alpha$. Then there is no sublinear sketch for equal-cardinality Chamfer distance on X.*

Proof. Identify Y with the natural numbers. Reductions proceed similarly to those above, with minor changes. The idea in both cases is to cluster the remaining points in a δ-neighborhood around p. If we choose δ sufficiently small relative to α and ϵ, approximation will still allow us to determine whether the "true" value is at least α. □

4 Discussions

We have described several algorithms for sketching and embedding Earth Mover's Distance. Speificially, we have shown how EMD was shown to embed into l_1 using the grid model from [IT03] and [Ind07] and described an efficient way to efficently sketch EMD from [ABIW09]. We have shown that the Chamfer Distance algorithm suggested in 3.3 is intrinsically as hard as the algorithm for Earth-Mover's Distance, since they both use mapping $f(A)$, and it is possible to generate certain sets such that the Earth-mover Distance and Chamfer Distance between them are the same. In some cases, calculating Chamfer Distance could be much easier than calculating Earth-Mover's Distance, but the algorithm does not provide this improvement. One possible direction of research is to design an algorithm that will solve the problem with less time complexity when Chamfer Distance and Earth-Mover Distance differ by a lot. We also detailed some outlines for the communication complexity of Chamfer Distance for certain cases.

References

[ABIW09] Alexandr Andoni, Khanh Do Ba, Piotr Indyk, and David Woodruff. Efficient sketches for earth-mover distance, with applications. In *2009 50th Annual IEEE Symposium on Foundations of Computer Science*, pages 324–330, 2009.

[BIJ+23] Ainesh Bakshi, Piotr Indyk, Rajesh Jayaram, Sandeep Silwal, and Erik Waingarten. A near-linear time algorithm for the chamfer distance, 2023.

[Ind07] Piotr Indyk. A near linear time constant factor approximation for euclidean bichromatic matching (cost). In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, page 39–42, USA, 2007. Society for Industrial and Applied Mathematics.

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