Sketching Earth-Mover Distance on Graph Metrics

Andrew McGregor* Daniel Stubbs†

Abstract

We develop linear sketches for estimating the Earth-Mover distance between two point sets, i.e., the cost of the minimum weight matching between the points according to some metric. While Euclidean distance and Edit distance are natural measures for vectors and strings respectively, Earth-Mover distance is a well-studied measure that is natural in the context of visual or metric data. Our work considers the case where the points are located at the nodes of an implicit graph and define the distance between two points as the length of the shortest path between these points. We first improve and significantly simplify an existing result by Brody et al. [4] for the case where the graph is a cycle. We then generalize our results to arbitrary graph metrics. Our approach is to recast the problem of estimating Earth-Mover distance in terms of an ℓ_1 regression problem. The resulting linear sketches also yield space-efficient data stream algorithms in the usual way.

1 Introduction

Given two multi-sets $A, B \subseteq \mathcal{X}$ where $|A| = |B| = k$ and a metric d on $\mathcal{X}$, the *Earth-Mover Distance* (EMD) between A and B is defined as the minimum cost of a matching between A and B, i.e.,

$$\mathrm{EMD}_d(A, B) = \min_{\pi: A \to B} \sum_{a \in A} d(a, \pi(a))$$

where π ranges over all bijective mappings between A and B. Earth-Mover distance is a natural and well-studied notion of the difference between two point sets. It was initially proposed in the context of image retrieval and has been shown to correspond closely to the perceptual difference between two images [14]. While Euclidean distance and Edit distance are natural measures of dissimilarity for vectors and strings respectively, EMD is perhaps the most natural measure for metric and visual data.

Linear sketching is a popular technique for processing large data sets. See Cormode et al. [7] for a survey. The basic idea is to take random linear projections of the data set and then post-process these projections in order to evaluate properties of the original data. The main parameters of the sketch are the size, or dimension, of the projection and the time required to perform the post-processing. Important applications of sketching include processing data streams or distributed data. A significant fraction of the work on linear sketches has focused on the problem of distance estimation and estimating the Earth-Mover distance is a long-standing open question [9, 12] that

*University of Massachusetts Amherst. Email: mcgregor@cs.umass.edu. Supported by NSF CAREER Award CCF-0953754.

†University of Massachusetts Amherst. Email: dstubbs@student.umass.edu. Supported by NSF CAREER Award CCF-0953754 REU Supplement.

has remained open (in the case where the point sets lie on a $\Delta \times \Delta$ grid) despite a substantial body of work dedicated to the problem [3, 4, 8, 10, 17]. The best known results achieve a logarithmic approximation with sketches of poly-logarithmic size and an $O(1/\epsilon)$ approximation with sketches of Δ^ϵ size. The most relevant work to this paper is a recent paper by Brody et al. [4] in which they consider the more restricted case where $\mathcal{X}$ corresponds to the nodes of a cycle and d is the shortest-path metric on this cycle (see also Cabrelli and Molter [5] for an optimal solution in the offline, non-streaming case). In this case they show that $(1+\epsilon)$-approximation is possible with sketches of poly-logarithmic size.

1.1 Our Techniques and Results

In this paper, we consider d to be the shortest-path metric in an arbitrary graph $G = (V, E)$ on $n = |V|$ nodes with $m = |E|$ edges. Note that the graph structure is assumed to be known in advance[1] and the input is point sets A and B of size k. Our results are as follows.

1. *Cycles:* $O(\epsilon^{-2} \operatorname{polylog} nk)$-size sketches for approximating $\mathrm{EMD}(A, B)$ up to a $(1+\epsilon)$ factor with high probability. This improves over the existing sketch of Brody et al. which used sketches of length $O(\epsilon^{-3} \operatorname{polylog} nk)$. We then show how to ensure that post-processing the sketch can be performed in $O(\operatorname{polylog} n)$ time. Our analysis also has the advantage of being significantly simpler. See Section 3.

2. *Trees:* $O(\epsilon^{-2} \operatorname{polylog} nk)$-size sketches for approximating $\mathrm{EMD}(A, B)$ up to a $(1+\epsilon)$ factor with high probability. By combining recent results on range-summable random variables by Tirthapura and Woodruff [16] with a natural path-decomposition we show how such a sketch can be applied in the data-stream setting with $O(\operatorname{polylog} n)$ update time whereas even in the cycle case, the existing sketch has $\Omega(n)$ update time. See Section 4.

3. *Arbitrary Graphs:* $O(\epsilon^{-2} \cdot t \cdot \operatorname{polylog} nk)$-size sketches for approximating $\mathrm{EMD}(A, B)$ up to a $(1+\epsilon)$ factor with high probability where $t = m - n + 1$ is the number of edges that need to be removed from G such that the resulting graph is acyclic. This generalizes our result on cycles in which $t = 1$. While our results hold for arbitrary t, our results are most interesting in the case where there are relatively few cycles and hence t is moderate in size. See Section 5.

Technical Approach. The general approach is follows. We define vectors $x, y \in \mathbb{R}^{|E|}$ corresponding to the two multi-sets A and B. We then relate $\mathrm{EMD}(A, B)$ to an ℓ_1-regression problem involving x, y, and a set of vectors defined by the structure of the underlying graph. To achieve our results, we first sketch the vectors, i.e., construct random projections of these vectors, and then perform the ℓ_1-regression on the sketched vectors rather than manipulating the original vectors explicitly.

2 Preliminaries

Notation. We use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. We say an algorithm is an (ϵ, δ)-approximation for a quantity Q if the value returned $\tilde{Q}$ satisfies $\mathbb{P}\left[|Q - \tilde{Q}| < \epsilon Q\right] \geq 1 - \delta$. Given a tree $T = (V, E)$ and $u, v \in V$ we define,

$$P_T(u, v) = \{e \in E : \ e \text{ on the path between nodes } u \text{ and } v\} .$$

[1] This is in contrast to recent work in graph sketching [1, 2] where the goal is to sketch the actual graph.

We denote the ℓ_1-norm of a vector x by $\|x\|_1 = \sum_i |x_i|$.

Sketches for ℓ_1-norm estimation. ℓ_1-norm estimation is one of the canonical sketching and data stream problems. We will make extensive use of the following result due to Kane et al. [11].

Theorem 1 (ℓ_1 Sketching [11])**.** *There exists a distribution ν over linear maps from $\mathbb{R}^n \to \mathbb{R}^q$ where $q = O(\epsilon^{-2} \log n \log \delta^{-1})$ and a "post-processing" function $f : \mathbb{R}^q \to \mathbb{R}$ such that for any $x \in \mathbb{R}^n$ with polynomially-bounded entries,*

$$\Pr_{M \sim \nu}[|\|x\|_1 - f(Mx)| \leq \epsilon \|x\|_1] \geq 1 - \delta \ .$$

Note that it immediately follows by rescaling δ and applying the union bound that by increasing q to $O(\epsilon^{-2} \log n \log(t\delta^{-1}))$ we ensure that for any t vectors $\mathcal{X} = \{x_1, \ldots, x_t\}$,

$$\Pr_{M \sim \nu}[\forall x \in \mathcal{X} \ ; \ |\|x\|_1 - f(Mx)| \leq \epsilon \|x\|_1] \geq 1 - \delta \ .$$

In particular, if $\mathcal{X}$ consists of all linear combinations of some set $\{y_1, \ldots, y_r\}$ where the linear coefficients are from the set $\{-k, -k+1, \ldots, k-1, k\}$ then $t = (2k+1)^r$ and we can estimate the ℓ_1 norm of any vector $x \in \{x_1, \ldots, x_t\}$ from $O(r\epsilon^{-2} \log n \log(k\delta^{-1}))$-dimensional sketches $My_1, My_2, \ldots, My_r$ since

$$M(\sum_{i \in [r]} \lambda_i y_i) = \sum_{i \in [r]} \lambda_i M y_i \ .$$

One-Dimensional EMD. We next describe an important folklore result for sketching earth-mover distance in one dimension. For the sake of future sections, it will be helpful to describe this result in terms of graph distances when the graph is a path. Let $G = (V, E)$ be a path on n nodes, i.e., $V = \{1, 2, \ldots, n\}$ and edges $E = \{e_1, e_2, \ldots, e_{n-1}\}$ where $e_i = \{i, i+1\}$. Suppose $A, B \subset V$ and define the distance between $i \in A$ and $j \in B$ to be shortest path distance $d(i, j) = |i - j|$.

We can relate $\mathrm{EMD}(A, B)$ to a norm estimation problem as follows. Define the vectors $x, y \in \mathbb{R}^{n-1}$ where:

$$\forall i \in [n-1] \ ; \quad x_i = |\{a \in A : i \geq a\}| \quad \text{and} \quad y_i = |\{b \in B : i \geq b\}| \ .$$

Then the following theorem establishes that $\mathrm{EMD}(A, B)$ equals $\|x - y\|_1$.

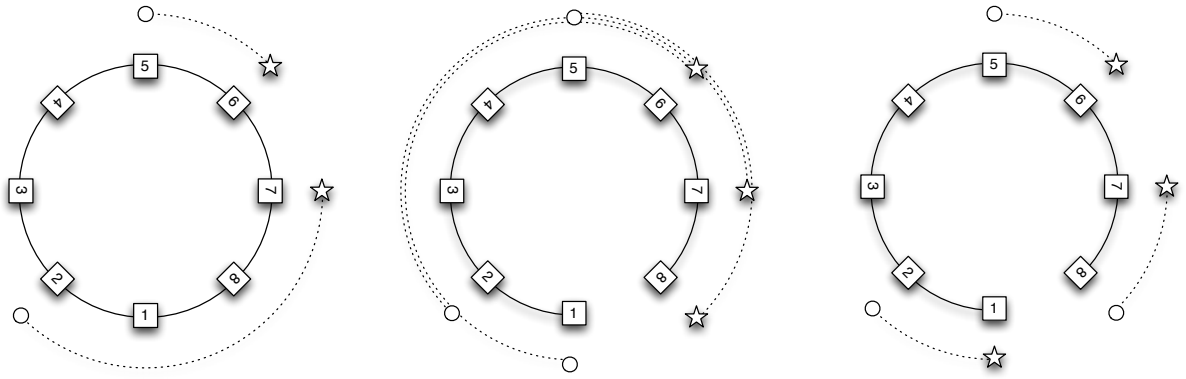
Theorem 2 (Folklore)**.** $\mathrm{EMD}(A, B) = \|x - y\|_1$.

We will actually prove a more general result in Lemma 7 from which the above theorem follows immediately. For intuition, suppose $A = \{i\}$ and $B = \{j\}$ and $i < j < n$. Then, $x = (0, \ldots, 0, 1, \ldots, 1)$ where the first "1" is in the i-th position and $y = (0, \ldots, 0, 1, \ldots, 1)$ where the first "1" is in the j-th position. Therefore $\|x\|_1$ and $\|y\|_1$ correspond to the distances that would be covered moving points i and j to node n. However, $y - x = (0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0)$ where $(y - x)_k = 1$ iff $i \leq k < j$ and so $\|y - x\|_1 = |j - i|$. Essentially, the effect of moving both points i and j to n cancels out along edges on which both points are being moved. The following example illustrates that the theorem applies in a less trivial case.

Example 1. *Suppose $A = \{2, 3, 10\}$ and $B = \{3, 4, 8\}$ and note that $\mathrm{EMD}(A, B) = 4$. Then*

$$x = (0, 1, 2, 2, 2, 2, 2, 2, 2) \quad \text{and} \quad y = (0, 0, 1, 2, 2, 2, 2, 3, 3)$$

and $\|x - y\|_1 = 4$ as required.



(a) Original Cycle Instance where $\mathrm{EMD}_d(A, B) = 4$.

(b) Linear Instance with $\lambda = 1$ where $1 + \mathrm{EMD}_{d'}(A + C_\lambda, B + C_{-\lambda}) = 14$.

(c) Linear Instance with $\lambda = -1$ where $1 + \mathrm{EMD}_{d'}(A + C_\lambda, B + C_{-\lambda}) = 4$.

Figure 1: Reducing Cyclic EMD to Linear EMD. Points in A are denoted by circles and points in B are denoted by stars.

3 Cycles

Consider a cycle on n nodes $\{1, 2, \ldots, n\}$ and edges $e_1, e_2, \ldots, e_n$ where $e_i = \{i, i+1\}$ for $i \in [n-1]$ and $e_n = \{n, 1\}$. The basic idea for solving EMD on the cycle is to reduce it to the one-dimensional, or path metric, case by simply ignoring the last edge e_n. This has the effect of changing the distance between nodes i and j from

$$d(i, j) = \min(|i - j|, |i - n| + 1 + |1 - j|, |i - 1| + 1 + |n - j|)$$

to a new distance

$$d'(i, j) = |i - j| \ .$$

Depending on the point sets, A and B, this can change the earth-mover distance significantly since two points that were previously close may now be far apart. For example, if $A = \{n\}$ and $B = \{1\}$ then the earth-mover distance increases from $\mathrm{EMD}_d(A, B) = 1$ to $\mathrm{EMD}_{d'}(A, B) = n - 1$.

To rectify this issue, we will effectively make a series of guesses $\{-k, -k+1, \ldots, k-1, k\}$ for how many pairs of points will be paired using the edge e_n.

Lemma 3. *For $\lambda \in \{-k, -k+1, \ldots, k-1, k\}$, let C_λ be the multi-set consisting of λ copies of "1" if $\lambda > 0$ and $|\lambda|$ copies of "n" if $\lambda < 0$. Then,*

$$\mathrm{EMD}_d(A, B) \leq |\lambda| + \mathrm{EMD}_{d'}(A + C_\lambda, B + C_{-\lambda})$$

with equality for some $\lambda \in \{-k, -k+1, \ldots, k-1, k\}$.

Proof. Consider a bijection π between $A + C_\lambda$ and $B + C_{-\lambda}$. We will show that π induces a bijection σ between A and B such that

$$\sum_{a \in A} d(a, \sigma(a)) \leq |\lambda| + \sum_{a \in A + C_\lambda} d'(a, \pi(a)) \ , \tag{1}$$

and this establishes the first part of the lemma.

It will be convenient to enumerate the elements of $C_\lambda = \{c_1, c_2, \ldots, c_\lambda\}$ and $C_{-\lambda} = \{d_1, d_2, \ldots, d_\lambda\}$ such that we may assume that $\pi(c_i) = d_j$ implies $i = j$. We then define σ as follows. If $\pi(a) \in B$ for $a \in A$ then define $\sigma(a) = \pi(a)$ and hence

$$d(a, \sigma(a)) = d'(a, \pi(a)) \ . \tag{2}$$

If $\pi(a) = d_i$ for some $a \in A$ and $d_i \in C_{-\lambda}$ then define $\sigma(a) = \pi(c_i)$. Hence,

$$d(a, \pi(a)) \leq d'(a, d_i) + 1 + d'(c_i, \pi(c_i)) \ .$$

Note that there are at most $|\lambda|$ elements $a \in A$ such that $\pi(a) \in C_{-\lambda}$ and together with Eq. 2 this establishes Eq. 1.

To prove that there exists λ such that $\mathrm{EMD}_d(A, B) = |\lambda| + \mathrm{EMD}_{d'}(A + C_\lambda, B + C_{-\lambda})$ consider the bijection $\sigma = \mathrm{argmin}_\sigma \sum_{a \in A} d(a, \sigma(a))$. Suppose there are λ_1 elements $a \in A$ such that the shortest path from a to $\sigma(a)$ visits n then 1. Similarly, suppose there are λ_2 elements $a \in A$ such that the shortest path from a to $\sigma(a)$ visits 1 then n. Note that at most one of λ_1 and λ_2 is non-zero since σ is the minimal cost bijection. Then setting $\lambda = \lambda_1 - \lambda_2$ ensures $\mathrm{EMD}_d(a, \sigma(a)) = |\lambda| + \mathrm{EMD}_{d'}(A + C_\lambda, B + C_{-\lambda})$ as required. □

3.1 Sketch Details

To construct the sketch we first define the vectors $x, y \in \mathbb{R}^n$ where for $i \in [n-1]$

$$x_i = |\{a \in A : i \geq a\}| \quad \text{and} \quad y_i = |\{b \in B : i \geq b\}| \ .$$

and $x_n = y_n = 0$. Define $z = x - y$ and let $c = (1, 1, \ldots, 1, 1) \in \mathbb{R}^n$.

Lemma 4. $\min_{-k \leq \lambda \leq k} \|z - \lambda c\|_1 = \mathrm{EMD}(A, B)$.

Proof. Let $z_{[n-1]}$ and $c_{[n-1]}$ be the vectors corresponding to the first $n-1$ elements of z and c respectively and note that

$$\|z - \lambda c\|_1 = |\lambda| + \|z_{[n-1]} - \lambda c_{[n-1]}\|_1 \ .$$

The proof then follows from Theorem 2 and Lemma 3. □

We define the function $f(\lambda) = \|z - \lambda c\|_1$. From the above lemma, it suffices to find $\min_\lambda f(\lambda)$. From Theorem 1 (and the surrounding discussion), it is possible to compute estimates $\{\tilde{f}_\lambda\}_{\lambda \in \{-k, \ldots, k\}}$ from a $O(\epsilon^{-2} \log n \log(k\delta^{-1}))$-dimensional sketch of z such that

$$\mathbb{P}\left[\forall \lambda \in \{-k, \ldots, k\} \ : \ |\tilde{f}_\lambda - f(\lambda)| \leq \epsilon f(\lambda)\right] \geq 1 - \delta \ .$$

Hence, if we return $\min \tilde{f}_\lambda$ then we have an (ϵ, δ)-approximation for $\mathrm{EMD}(A, B)$. However, rather than evaluating every $\tilde{f}_\lambda$ to find the minimum, in the next section we show that it is possible to find $\min_{\lambda \in \{-k, \ldots, k\}} \tilde{f}_\lambda$ while only evaluating $O(\log k)$ of the terms.

3.2 Improved Post-Processing

The main observation is that since $f(\lambda) = \sum_i |z_i - \lambda c_i|$ is a sum of convex functions, $f(\lambda)$ itself is convex and can therefore be minimized by using something like a binary search.

Lemma 5. *$f(\lambda) = \|z - \lambda c\|_1$ is convex.*

However, although $f(\lambda)$ is convex, the errors in our estimates $\tilde{f}_\lambda$ of $f(\lambda)$ may violate the convexity property. To accommodate this we perform a quaternary search that includes tolerances for these errors. See Algorithm 1.

Algorithm 1 Approximate Quaternary Search

$(l, u) \leftarrow (-k, k)$
while $l \neq u$ **do**
  $(a, b, c) \leftarrow \left(\left\lfloor \frac{3l+u}{4} \right\rfloor, \left\lfloor \frac{2l+2u}{4} \right\rfloor, \left\lfloor \frac{l+3u}{4} \right\rfloor\right)$
  if $\max(\tilde{f}_a, \tilde{f}_b, \tilde{f}_c)/\min(\tilde{f}_a, \tilde{f}_b, \tilde{f}_c) < \frac{1+\epsilon}{1-\epsilon}$ or $\tilde{f}_b = \max(\tilde{f}_a, \tilde{f}_b, \tilde{f}_c)$ **then**
    return $\tilde{f}_b$
  else

$$(l, u) \leftarrow \begin{cases} (a, u) & \text{if } \tilde{f}_a = \max(\tilde{f}_a, \tilde{f}_b, \tilde{f}_c) \\ (l, c) & \text{if } \tilde{f}_c = \max(\tilde{f}_a, \tilde{f}_b, \tilde{f}_c) \end{cases}$$

  end if
end while
return $\tilde{f}_l$

Lemma 6. *Algorithm 1 returns a value that is within a factor $1 \pm O(\epsilon)$ of $\min_\lambda f(\lambda)$.*

Proof. Let $\lambda^* = \text{argmin}_{\lambda \in \{-k, \ldots, k\}} f(\lambda)$. We first prove the invariant that l and u always satisfy $l \leq \lambda^* \leq u$. Note that it is true initially since $l = -k$ and $u = k$. Suppose it is true at a given iteration, then (by symmetry) it suffices to show that if $\max(\tilde{f}_a, \tilde{f}_b, \tilde{f}_c)/\min(\tilde{f}_a, \tilde{f}_b, \tilde{f}_c) \geq \frac{1+\epsilon}{1-\epsilon}$ and $\tilde{f}_a = \max(\tilde{f}_a, \tilde{f}_b, \tilde{f}_c)$ then $a \leq \lambda^*$. Then,

$$\frac{f(a)}{\min(f(b), f(c))} \geq \frac{\tilde{f}_a/(1+\epsilon)}{\min(\tilde{f}_b/(1-\epsilon), \tilde{f}_c/(1-\epsilon))} = \frac{1-\epsilon}{1+\epsilon} \cdot \frac{\max(\tilde{f}_a, \tilde{f}_b, \tilde{f}_c)}{\min(\tilde{f}_a, \tilde{f}_b, \tilde{f}_c)} \geq 1 \ .$$

and hence $f(a) \geq \min(f(b), f(c))$. By the convexity of f we deduce that $a \leq \lambda^*$ as required.

It remains to show that when the algorithm terminates, the return value is sufficiently accurate.

Case 1: If $l = u$ then $\tilde{f}_l = (1 \pm \epsilon) f(l) = (1 \pm \epsilon) f(\lambda^*)$.

Case 2: Suppose that $\max(\tilde{f}_a, \tilde{f}_b, \tilde{f}_c)/\min(\tilde{f}_a, \tilde{f}_b, \tilde{f}_c) < \frac{1+\epsilon}{1-\epsilon}$ and therefore

$$\frac{\max(f(a), f(b), f(c))}{\min(f(a), f(b), f(c))} < \left(\frac{1+\epsilon}{1-\epsilon}\right)^2 \ .$$

By symmetry, assume that $\lambda^* \leq b$. Then, by the convexity of f we have:

$$f(\lambda^*) \geq f(b) - 1/2 \cdot \frac{f(c) - f(b)}{1/4} = f(b)(3 - 2f(c)/f(b)) \geq (1 - O(\epsilon))\tilde{f}_b \ .$$

Case 3: Suppose that $\tilde{f}_b = \max(\tilde{f}_a, \tilde{f}_b, \tilde{f}_c)$, and assume by symmetry $\lambda^* \leq b$. Then

$$(1+\epsilon)^2 f(b) \geq (1+\epsilon)\tilde{f}_b \geq (1+\epsilon)\tilde{f}_c \geq f(c)$$

which gives us that the difference between $f(c)$ and $f(b)$ is at most $(2\epsilon+\epsilon^2)f(b)$. By convexity, the difference between $f(b)$ and $f(\lambda^*)$ is at most twice this, since λ^* is at most twice as far from b as c is, so $f(\lambda^*) \geq f(b) - 2(2\epsilon+\epsilon^2)f(b) = (1-O(\epsilon))f(b)$.

□

4 Trees

In this section, we generalize the one-dimensional case discussed in Section 2 to trees. Let $T = (V, E)$ be a tree on n nodes. Suppose $A, B \subseteq V$ where for $a \in A, b \in B$, $d(a, b)$ is the length of the unique path between a and b.

To relate EMD_d with the tree metric to ℓ_1 norms we first pick an arbitrary root r of T. Now define the vectors $x, y \in \mathbb{R}^E$ where

$$x_e = |\{a \in A : e \in P_T(a, r)\}| \quad \text{and} \quad y_e = |\{b \in B : e \in P_T(b, r)\}| \ .$$

and define $z = x - y$. Recall that $P_T(u, v)$ is the set of edges on the unique path in T between u and v. The following lemma generalizes Theorem 2 (the "root" in the path case was implicitly chosen to be node n) and will play an important role in the next section.

Lemma 7. $\|z\|_1 = \text{EMD}_d(A, B)$.

Proof. For each edge $e = (u, v) \in T$ where u is a child of v, define the value

$$w_e = \big||A \cap V_u| - |B \cap V_u|\big|$$

where V_u is the set of nodes of the subtree rooted at u. Then $\text{EMD}_d(A, B) = \sum_{e \in E} w_e$ since in the optimal bijection, either all points in $A \cap V_u$ will be mapped to elements in $B \cap V_u$ or vice versa and hence the edge e appears in exactly $||A \cap V_u| - |B \cap V_u||$ of the shortest paths between matched points. But $w_e = |x_e - y_e|$ since $e \in P_T(v, r)$ iff $v \in V_u$. Hence, $\text{EMD}(A, B) = \sum_{e \in E} w_e = \sum_{e \in E} |x_e - y_e| = \|z\|_1$. as required. □

Therefore, appealing to the ℓ_1 sketch result in Theorem 1, it immediately follows that there is an $O(\epsilon^{-2} \log n \log \delta^{-1})$-dimensional sketch that returns an (ϵ, δ) approximation for $\text{EMD}_d(A, B)$ when d is a tree metric.

4.1 Improved Update Time

A naive implementation of the above algorithm requires $\Omega(n)$ update time since every update requires updating as many as $n - 1$ entries of the vector. However, this can be reduced to $O(\text{polylog}\, n)$ time using the range-efficient ℓ_1 sketching algorithm of Tirthapura and Woodruff [16]. This allows contiguous segments of the vector z to be updated in $O(\text{polylog}\, n)$ time rather than $O(w \,\text{polylog}\, n)$ time where w is the length of the segment. Hence, if we can ensure that any update of z involves updating $O(\log n)$ contiguous segments we enable any update to be performed in $O(\text{polylog}\, n)$ time. To do this, we will use the following path decomposition of the tree.

Lemma 8. *For any tree $T = (V, E)$ on n nodes with ℓ leaves and root r, it is possible to decompose E into ℓ paths $P_1, \ldots, P_\ell$ such that for any $u \in V$, $P_T(u, r)$ intersects at most $O(\log \ell)$ paths.*

Proof. We define the segments $P_1, \ldots, P_\ell$ as follows. Start a segment for each leaf consisting of the edge incident on it, and associate a value of 1 with the segment. Extend these segments in the direction of the root until each reaches a node of degree ≥ 3. At each such node, we continue the segment with the highest value (ties broken arbitrarily) but add the sum of the values of the concluded segments to the value of the continued segment. Note that this value is now at least twice the value of any of the segments that were concluded. We continue in this manner until we reach the root. In the end, each edge will belong to exactly one segment. Note that the path from an arbitrary node $u \in V$ to the root can intersect with at most $\log \ell$ of the resulting segments because the value of successive intersecting segments at least doubles and the maximum value is ℓ. □

Then, if we let the first $|P_1|$ elements of z correspond to P_1, the next $|P_2|$ elements correspond to P_2, etc. we ensure that when we add (or subtract) 1 to each entry corresponding to $P_T(u, r)$ for some u, this involves only $O(\log n)$ updates of contiguous intervals.

5 Arbitrary Graphs

In this final section, we generalize all our previous results and design a sketch for earth-mover distance for arbitrary graph metrics. Let $G = (V, E)$ be a graph on n nodes. Define a metric d where for $a, b \in V$, $d(a, b)$ is the length of the shortest path between a and b in G.

The approach to estimating $\mathrm{EMD}_d(A, B)$ is to reduce to the tree-metric case solved in the previous section. This naturally extends the approach in Section 3 where we reduced the cycle case to the path-metric case. Specifically, let $T = (V, E_T)$ be an arbitrary spanning tree and let $F = E \setminus E_T$. For example, see Figure 2 where $E_T = \{e_1, e_2, e_3, e_4\}$ and $F = \{f_1, f_2\}$. The tree T defines a metric d' where for $a, b \in V$, $d'(a, b)$ is the length of the shortest path between a and b in T.

The next lemma shows that we can express $\mathrm{EMD}_d(A, B)$ in terms of $\mathrm{EMD}_{d'}(A', B')$ where $A \subseteq A'$ and $B \subseteq B'$. The lemma is a generalization of Lemma 3.

Lemma 9. *For $f = (u, v) \in F$ and $\lambda \in \{-k, -k+1, \ldots, k-1, k\}$, let C^f_λ be the multi-set consisting of λ copies of "u" if $\lambda > 0$ and $|\lambda|$ copies of "v" if $\lambda < 0$. Then,*

$$\mathrm{EMD}_d(A, B) \leq \sum_{f \in F} |\lambda_f| + \mathrm{EMD}_{d'}(A + \sum_{f \in F} C^f_{\lambda_f}, B + \sum_{f \in F} C^f_{-\lambda_f}) \tag{3}$$

with equality for some set of coefficients λ_f.

Proof. Consider a bijection π between $A' = A + \sum_{f \in F} C^f_{\lambda_f}$ and $B' = B + \sum_{f \in F} C^f_{-\lambda_f}$. We will show that π induces a bijection σ between A and B such that

$$\sum_{a \in A} d(a, \sigma(a)) \leq \sum_{f \in F} |\lambda_f| + \sum_{a \in A'} d'(a, \pi(a)) \ , \tag{4}$$

and this will establish the first part of the lemma.

It will be convenient to enumerate the elements of $C^f_{\lambda^f}$ and $C^f_{-\lambda^f}$:

$$C^f_{\lambda^f} = \{c^f_1, c^f_2, \ldots\} \quad \text{and} \quad C^f_{-\lambda^f} = \{d^f_1, d^f_2, \ldots\}$$

such that we may assume that $\pi(c_i^f) = d_j^f$ implies $i = j$. Then, for each $a \in A$, define the sequence:

$$s_a = (a, d_{i_1}^{f_1}, c_{i_1}^{f_1}, d_{i_2}^{f_2} \ldots, c_{i_{k-1}}^{f_{k-1}}, d_{i_k}^{f_k}, c_{i_k}^{f_k}, b)$$

where each successive element is uniquely defined by π and the indexing of the elements in each $C_{\lambda f}^f$ and $C_{-\lambda f}^f$:

$$d_{i_1}^{f_1} = \pi(a) \ , \quad d_{i_2}^{f_2} = \pi(c_{i_1}^{f_1}) \ , \quad d_{i_3}^{f_3} = \pi(c_{i_2}^{f_2}) \ , \quad \ldots \ , \quad d_{i_k}^{f_k} = \pi(c_{i_{k-1}}^{f_{k-1}}) \ , \quad \text{and} \quad b = \pi(c_{i_k}^{f_k}) \ .$$

Given s_a, define $\sigma(a) = b$, i.e., we match a with the last element of the sequence. Note that

$$d(a, \pi(a)) \leq d'(a, d_{i_1}^{f_1}) + 1 + d'(c_{i_1}^{f_1}, d_{i_2}^{f_2}) + 1 + \ldots + 1 + d'(c_{i_k}^{f_k}, b) \ .$$

Summing over all $a \in A$, gives Eq. 4 since each pair (d_i^f, c_i^f) appears in at most one sequence because π is a bijection.

To prove that there exists a set of coefficients such that Eq. 3 is tight, consider the bijection $\sigma = \text{argmin}_\sigma \sum_{a \in A} d(a, \sigma(a))$. Then, for each $f = (u, v)$, let

$$\begin{aligned} \lambda_f \quad = \quad & |\{a \in A : \ u \text{ appears before } v \text{ on path between } a \text{ and } \sigma(a)\}| \\ & -|\{a \in A : \ u \text{ appears before } v \text{ on path between } \sigma(a) \text{ and } a\}| \ . \end{aligned}$$

Then $\text{EMD}_{d'}(A + C_{\lambda_f}^f, B + C_{-\lambda_f}^f) \leq \text{EMD}_d(A, B) - \sum_{f \in F} |\lambda_f|$ since with the addition of the $C_{\lambda_f}^f$ and $C_{-\lambda_f}^f$ sets we can consider the matching between $A + C_{\lambda_f}^f$ and $B + C_{-\lambda_f}^f$ induced by removing all edges $f \in F$. □

5.1 Sketch Details

For a graph $G = (V, E)$, let $T = (V, E_T)$ be an arbitrary spanning tree with root r. Define the vectors $x, y \in \mathbb{R}^E$ and $z = x - y$ where

$$x_e = \begin{cases} |\{a \in A : e \in P_T(a, r)\}| & \text{if } e \in E_T \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad y_e = \begin{cases} |\{b \in B : e \in P_T(b, r)\}| & \text{if } e \in E_T \\ 0 & \text{otherwise} \end{cases} \ .$$

For each $f = (u, v) \in F$, we define a vector c^f where

$$c_e^f = \begin{cases} 1 & \text{if } e \in P_T(u, r) \setminus P_T(v, r) \\ -1 & \text{if } e \in P_T(v, r) \setminus P_T(u, r) \\ 1 & \text{if } e = f \\ 0 & \text{otherwise} \end{cases}$$

The intuition behind the definition of c^f is that if z corresponds to point sets A and B, then $z + \lambda_f c^f$ corresponds to point sets $A + C_{\lambda_f}^f$ and $A + C_{-\lambda_f}^f$.

Example 2. *Consider the instance in Figure 2. In this case*

$$x = (1, 1, 0, 1, 0, 0) \ , \quad y = (0, 1, 1, 0, 0, 0) \ , \quad z = (1, 0, -1, 1, 0, 0)$$

$$c^{f_1} = (1, 1, 0, -1, 1, 0) \quad \textit{and} \quad c^{f_2} = (0, 0, 1, -1, 0, 1) \ .$$

Note that $\|z - 0c^{f_1} + 1c^{f_2}\|_1 = \|(1, 0, 0, 0, 0, 1)\|_1 = \text{EMD}_d(A, B)$ *and for arbitrary* λ_1, λ_2 *we have*

$$\|z - \lambda_1 c^{f_1} - \lambda_2 c^{f_2}\|_1 \geq \text{EMD}(A, B) \ .$$

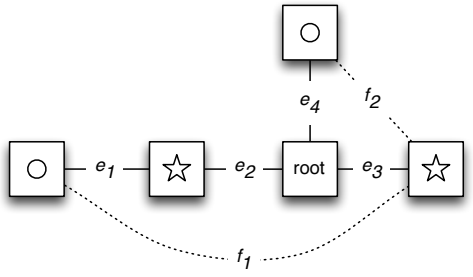


Figure 2: An instance of earth-mover distance on an arbitrary graph metric. See the text in Example 2. Points in A are denoted by circles and points in B are denoted by stars.

Lemma 10. $\min_{-k\leq\lambda_1,\ldots,\lambda_t\leq k} \|z - \sum_{f\in F} \lambda_f c^f\|_1 = \mathrm{EMD}_d(A,B)$.

Proof. Let $z_{[n-1]}$ and $c^f_{[n-1]}$ be the vectors corresponding to the first $n-1$ elements of z and c^f for each f. Note that

$$\|z - \sum_{f\in F} \lambda_f c^f\|_1 = \sum_{f\in F} |\lambda_f| + \|z_{[n-1]} - \sum_{f\in F} \lambda_f c^f_{[n-1]}\|_1 \ .$$

The proof then follows from Lemma 9 and Theorem 2. □

Extending the idea in Section 3, we now define the function $f(\lambda_1,\ldots,\lambda_t) = \|z - \sum_{f\in F} \lambda_f c^f\|_1$. From the above lemma, it suffices to estimate $\min_{-k\leq\lambda_1,\ldots,\lambda_t\leq k} f(\lambda_1,\ldots,\lambda_t)$. From Theorem 1 (and the surrounding discussion), it is possible to compute estimates $\{\tilde{f}_{\lambda_1,\ldots,\lambda_t}\}_{-k\leq\lambda_1,\ldots,\lambda_t\leq k}$ from a $O(t\epsilon^{-2}\log n\log(k\delta^{-1}))$-dimensional sketch of z such that with probability at least $1-\delta$, for all $-k\leq\lambda_1,\ldots,\lambda_t\leq k$

$$|f(\lambda_1,\ldots,\lambda_t) - \tilde{f}_{\lambda_1,\ldots,\lambda_t}| \leq \epsilon f(\lambda_1,\ldots,\lambda_t) \ .$$

Hence, if we return the minimum estimate then we have an (ϵ,δ) approximation for $\mathrm{EMD}(A,B)$. However, as in the cycle case, rather than evaluating every $\tilde{f}_{\lambda_1,\ldots,\lambda_t}$ to find the minimum, it is possible to find the minimum more efficiently. One option is to exploit the convexity of f as in Section 3 using a recursive regression algorithm [13] or to use recent results on robust regression via sub-space embeddings [6, 15].

Acknowledgements. We thank Michael Crouch for useful discussions about range-summable random variables.

References

[1] K. J. Ahn, S. Guha, and A. McGregor. Analyzing graph structure via linear measurements. In *SODA*, pages 459–467, 2012.

[2] K. J. Ahn, S. Guha, and A. McGregor. Graph sketches: sparsification, spanners, and subgraphs. In *PODS*, pages 5–14, 2012.

[3] A. Andoni, K. D. Ba, P. Indyk, and D. P. Woodruff. Efficient sketches for earth-mover distance, with applications. In *FOCS*, pages 324–330, 2009.

[4] J. Brody, H. Liang, and X. Sun. Space-efficient approximation scheme for circular earth mover distance. In *LATIN*, pages 97–108, 2012.

[5] C. Cabrelli and U. Molter. A linear time algorithm for a matching problem on the circle. *Inf. Process. Lett.*, 66(3):161–164, 1998.

[6] K. L. Clarkson, P. Drineas, M. Magdon-Ismail, M. W. Mahoney, X. Meng, and D. P. Woodruff. The fast cauchy transform: with applications to basis construction, regression, and subspace approximation in l1. *CoRR*, abs/1207.4684, 2012.

[7] G. Cormode, M. N. Garofalakis, P. J. Haas, and C. Jermaine. Synopses for massive data: Samples, histograms, wavelets, sketches. *Foundations and Trends in Databases*, 4(1-3):1–294, 2012.

[8] P. Indyk. A near linear time constant factor approximation for euclidean bichromatic matching (cost). In *SODA*, pages 39–42, 2007.

[9] P. Indyk, A. McGregor, I. Newman, and K. Onak, editors. *Open Problems in Data Streams, Property Testing, and Related Topics*, 2011. Available at: *http://www.cs.umass.edu/ mcgregor/papers/11-openproblems.pdf*.

[10] P. Indyk and E. Price. K-median clustering, model-based compressive sensing, and sparse recovery for earth mover distance. In *STOC*, pages 627–636, 2011.

[11] D. M. Kane, J. Nelson, E. Porat, and D. P. Woodruff. Fast moment estimation in data streams in optimal space. In *STOC*, pages 745–754, 2011.

[12] A. McGregor, editor. *Open Problems in Data Streams and Related Topics*, 2007. Available at: *www.cse.iitk.ac.in/users/sganguly/data-stream-probs.pdf*.

[13] A. McGregor, A. Rudra, and S. Uurtamo. Polynomial fitting of data streams with applications to codeword testing. In *STACS*, pages 428–439, 2011.

[14] Y. Rubner, C. Tomasi, 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.

[15] C. Sohler and D. P. Woodruff. Subspace embeddings for the l_1-norm with applications. In *STOC*, pages 755–764, 2011.

[16] S. Tirthapura and D. P. Woodruff. Rectangle-efficient aggregation in spatial data streams. In *PODS*, pages 283–294, 2012.

[17] E. Verbin and Q. Zhang. Rademacher-sketch: A dimensionality-reducing embedding for sum-product norms, with an application to earth-mover distance. In *ICALP (1)*, pages 834–845, 2012.