Sublinear Time Algorithms for Earth Mover's Distance

Khanh Do Ba	Huy L. Nguyen	Huy N. Nguyen	Ronitt Rubinfeld
MIT, CSAIL	MIT	MIT, CSAIL	MIT, CSAIL
doba@mit.edu	hlnguyen@mit.edu	huy2n@mit.edu	ronitt@csail.mit.edu

January 3, 2010

Abstract

We study the problem of estimating the Earth Mover's Distance (EMD) between probability distributions when given access only to samples of the distribution. We give closeness testers and additive-error estimators over domains in $[0, 1]^d$, with sample complexities independent of domain size – permitting the testability even of continuous distributions over infinite domains. Instead, our algorithms depend on the dimension of the domain space and the quality of the result required. We also prove lower bounds showing the dependencies on these parameters to be essentially optimal. Additionally, we consider whether natural classes of distributions exist for which there are algorithms with better dependence on the dimension, and show that for highly clusterable data, this is indeed the case. Lastly, we consider a variant of the EMD, defined over tree metrics instead of the usual ℓ_1 metric, and give tight upper and lower bounds.

1 Introduction

In traditional algorithmic settings, algorithms requiring linear time and/or space are generally considered to be highly efficient; in some settings, even polynomial time and space requirements are acceptable. However, today this is often no longer the case. With data being generated at rates of terabytes a second, sublinear time algorithms have become crucial in many applications. In the increasingly important area of massive data algorithmics, a number of models have been proposed and studied to address this. One of these arises when the data can be naturally viewed as a probability distribution (e.g., over IP addresses, or items sold by an online retailer, etc.) that allows independent and identically distributed (i.i.d.) samples to be drawn from it. This is the model on which this paper focuses.

Perhaps the most fundamental problem in this model is that of testing whether two distributions are close. For instance, if an online retailer such as Amazon.com wishes to detect changes in consumer habits, one way of doing so might be to see if the distribution of sales over all offered items this week, say, is significantly different from last week's distribution. This problem has been studied extensively, mostly under the ℓ_1 and ℓ_2 distances, with special focus on the dependence of the time and sample complexity on n, the size of the domain. Algorithms with sublinear dependence on the time and sample complexity exist to distinguish whether two distributions are identical or ε-far from each other [5, 9, 19]. However, under the ℓ_1 distance, for instance, the sample complexity, though sublinear, is $\Omega(n^{2/3})$ (where n is the domain size) [5, 19], which may be prohibitively large. Furthermore, in many situations the ℓ_1 (or ℓ_2) distance is not a good measurement for the changes in distributions. For example, at Amazon.com, a change in consumer habits that results in with a drop in the sale of PCs and rise in the sale of Macs must be more alarming than a drop in the sale of blue t-shirts and rise in the sale pink t-shirts. However, under the ℓ_1 metric, these changes in sale distributions might be considered as the same.

Fortunately, there is a natural metric on the underlying domain, under which nearby points should be treated as "less different" than faraway points. This motivates a metric known as Earth Mover's Distance (EMD), first introduced in the vision community as a measure of (dis)similarity between images that more accurately reflects human perception than more traditional ℓ_1 [12]. It has since proven to be important in computer graphics and vision [13, 14, 15, 7, 17, 18, 16], and has natural applications to other areas of computer science. As a result, its computational aspects have recently drawn attention from the algorithms community as well [2, 11, 6, 10]. However, previous work has generally focused on the more classical model of approximating the EMD when given the input distributions explicitly; that is, when the probability of any domain element can be queried. As far as we know, no work has been done on estimating EMD and closeness testing with respect to EMD in the setting in which one has access only to i.i.d. samples from the distributions.

In this model, it is easy to see that we cannot hope to compute a multiplicative approximation, even with arbitrarily many samples, since that would require us to distinguish between arbitrarily close distributions and identical ones. However, if we settle for additive error, we show in this paper that, in contrast to the ℓ_1 distance, we can estimate EMD using a number of samples *independent* of the domain size. Instead, our sample complexity depend only on the *dimension* of the domain space and the quality of the results required. The consequence is that this allows us to effectively deal with distributions over extremely large domains, and even, under a natural generalization of EMD, continuous distributions over infinite domains[1].

Specifically, if p and q are distributions over $M \subset [0,1]^d$, we can[2]

- test whether $p = q$ or $EMD(p,q) > \varepsilon$ with $\tilde{O}((2d/\varepsilon)^{2d/3})$ samples, and
- estimate $EMD(p,q)$ to within an additive error of ε with $\tilde{O}((4d/\varepsilon)^{d+2})$ samples.

We also prove a lower bound of $\Omega((1/\varepsilon)^{2d/3})$ samples for closeness testing with respect to EMD, implying our tester above is essentially optimal in its dependence on $1/\varepsilon$. When d is small, the bounds behave slightly different. In particular, when $d = 1$ or 2, upper and lower bounds for the tester converge at a tight $\Theta((1/\varepsilon)^2)$. Additionally, in the appendix we consider assumptions on the data that might make the problem easier, and give an improved algorithm for highly clusterable input distributions.

Besides the continuous domain, we also consider the EMD over tree metrics, which can be considered as a generalization of the ℓ_1 setting. While simple, tree metrics are extremely powerful. Bartal [3] showed how to probabilistically embed arbitrary graph metrics to tree metrics with distortion $O(\log^2 n)$. The distortion was later improved by Fakcharoenphol, Rao, and Talwar [8] to $O(\log n)$. Thanks to these results, many computational problems on graph metrics can be reduced to the same problems on tree metrics. In this paper, we give an optimal (up to polylogarithmic factors) algorithm for estimating the EMD over tree metrics. For tree metrics with bounded diameter, we give an optimal reduction from closeness testing for the EMD over tree metrics to closeness testing for ℓ_1.

[1] It is easy to see that every continuous distribution in the hypercube can be additively-approximated well by a small support distribution. In particular, for every quality parameter $\varepsilon > 0$, a continuous distribution can be approximated with an additive error ε by a distribution on the grid points of the grid of cell side ε.

[2] In what follows we slightly abuse the big-O notations and write $f(d,\varepsilon) = O(g(d,\varepsilon))$ (or $f(d,\varepsilon) = \Omega(g(d,\varepsilon))$ to mean that for any $d > 0$, there exists $\varepsilon_0 > 0$ such that for all $\varepsilon < \varepsilon_0$, $f(d,\varepsilon)$ is bounded above (or below) by $C \cdot g(d,\varepsilon)$ where C is some universal constant.

2 Preliminaries

Definition 1. *A* supply-demand network *is a directed bipartite graph $G = (S \cup T, E)$ consisting of supply vertices S and demand vertices T, with supply (probability) distribution p on S and demand (probability) distribution q on T, and edge set $E = S \times T$, with associated weights $w : E \to \mathbb{R}^+$. A* satisfying flow *for G is a mapping $f : E \to \mathbb{R}^+$ such that for each $s \in S$ and each $t \in T$,*

$$\begin{aligned} \sum_{t' \in T} f((s,t')) &= p(s), \text{ and} \\ \sum_{s' \in S} f((s',t)) &= q(t). \end{aligned}$$

The cost of satisfying flow f is given by

$$C(f) = \sum_{e \in E} f(e)w(e).$$

We define the Earth Mover's Distance (EMD) as follows.

Definition 2. *Let p and q be probability distributions on points in a finite metric space (M, δ). Then let G be the supply-demand network given by supply vertices $S = \{s_x \mid x \in M\}$ and demand vertices $T = \{t_x \mid x \in M\}$, with supply distribution $\hat{p} : s_x \mapsto p(x)$ and demand distribution $\hat{q} : t_x \mapsto q(x)$, and edge weights $w : (s_x, t_y) \mapsto \delta(x, y)$. Define $EMD(p, q)$ to be the minimum cost of all satisfying flows for G.*

Intuitively, $EMD(p, q)$ is the minimum cost to "move" the weight of distribution p around so that it matches distribution q where the cost of moving a weight w over a distance d is defined to be wd. It is straightforward to verify that the EMD as defined above is a metric on all probability distributions on M. Note, moreover, that to upperbound the EMD it suffices to exhibit any satisfying flow.

It is also possible to define the EMD between any two probability measures p and q on an infinite (continuous) metric space $(\mathcal{M}, \delta_{\mathcal{M}})$, $\mathcal{M} \subset [0, 1]^d$, by using the Wasserstein metric:

$$EMD(p, q) = \inf_{\gamma \in \Gamma(p,q)} \int_{\mathcal{M} \times \mathcal{M}} \delta_{\mathcal{M}}(x, y) d\gamma(x, y),$$

where $\Gamma(p, q)$ denotes the collection of all measures on $\mathcal{M} \times \mathcal{M}$ with marginals p and q on the first and second factors, respectively. By using an $\varepsilon/4$-net, $(\mathcal{M}, \delta_{\mathcal{M}})$ can be discretized into a finite metric (M, δ) in such a way that the EMD between any two probability distributions only changes by at most $\varepsilon/2$; thus, an EMD-closeness tester over (M, δ) with error parameter $\varepsilon/2$ is also a valid EMD-closeness tester over $(\mathcal{M}, \delta)$ with error parameter ε. Henceforth, we will therefore consider only the finite case, understanding that any result obtained will also apply to the continuous case.

Finally, we define a closeness tester and additive-error estimator for the EMD in the usual way as follows.

Definition 3. *Let p, q be two distributions on metric space (M, δ). An* EMD-closeness tester *is an algorithm which takes as input samples from p and q, together with parameter $\varepsilon > 0$, and guarantees with probability at least $2/3$ that (1) if $p = q$, then it accepts, and (2) if $EMD(p, q) > \varepsilon$, then it rejects. An* additive-error estimator for EMD *is an algorithm which, given the same inputs, returns a value in the range $[EMD(p, q) - \varepsilon, EMD(p, q) + \varepsilon]$ with probability at least $2/3$.*

3 EMD over hypercubes

In this section we will consider the EMD over a (finite) space $M \subset [0,1]^d$ endowed with the ℓ_1 metric (note that the ℓ_1 metric used here is to measure the distance between two points in the hypercube and is different from the ℓ_1 distance between two distributions). We will give both a closeness tester and an additive-error estimator, and prove matching lower bounds.

3.1 Closeness testing

We start with several lemmas and facts that will allow us to tie the EMD with respect to ℓ_1.

Lemma 4. *If p and q are distributions on any finite space M, where M has minimum distance γ and diameter 1, then*

$$\frac{||p-q||_1}{2} \cdot \gamma \leq EMD(p,q) \leq \frac{||p-q||_1}{2}.$$

Proof. Observe that there is a minimum-cost satisfying flow f from S to T (as defined in Definition 2) such that the total amount sent by f through edges with non-zero cost is exactly $\|p-q\|_1/2$. The lemma directly follows, since such non-zero cost is at least δ and at most 1. □

Definition 5. *Given distribution p over $M \subset [0,1]^d$ and a positive integer i, let $G^{(i)}$ be a grid with side length $\frac{1}{2^i}$ over $[0,1]^d$ centered at the origin. Define the* i-coarsening of p, *denoted $p^{(i)}$, to be the distribution over the grid cells of $G^{(i)}$ such that, for each grid cell c of $G^{(i)}$, $p^{(i)}(c) = \sum_{u \in c} p(u)$.*

The $p^{(i)}$'s can be thought of as coarse approximations of p where all points in the same grid cell are considered to be the same point. We then have the following result from [11] relating the EMD of two distributions to a weighted sum of the ℓ_1 distances of their i-coarsenings.

Fact 6. *[11] For any two distributions p and q over $M \subset [0,1]^d$, for any $\varepsilon > 0$*

$$EMD(p,q) \leq d\left(\sum_{i=1}^{\log(2d/\varepsilon)} \frac{1}{2^{i-1}} \cdot \left\|p^{(i)} - q^{(i)}\right\|_1\right) + \frac{\varepsilon}{2}.$$

Having thus established the relationship between the EMD and ℓ_1 metrics, we can make use of the result from [5] for testing closeness of distributions in ℓ_1. Alternatively, via a simple Chernoff bound analysis (similar to [4]), we can show that a whole distribution can be approximated efficiently, giving us another closeness tester for ℓ_1 with worse dependence on n but better dependence on ε. Putting these together, we have the following:

Fact 7. *Given access to samples from two distributions p and q over M, where $|M| = n$, and parameters $\varepsilon, \delta > 0$, there exists an algorithm, denoted ℓ_1-`Closeness-Tester`(p,q,ε,δ), that takes $\tilde{O}(\min\{n^{2/3}\varepsilon^{-4}, n\varepsilon^{-2}\})$ samples and guarantees with probability at least $1-\delta$ that (1) if $p = q$, then it accepts, and (2) if $|p-q\|_1 > \varepsilon$, then it rejects.*

We then use Algorithm 1 as a subroutine for our EMD-closeness tester, giving us the following theorem.

Theorem 8. *Algorithm 1 is an EMD-closeness tester for distributions over $M \subset [0,1]^d$ that has sample complexity*

(i) $\tilde{O}((1/\varepsilon)^2)$ *if $d = 1$ or 2,*

Algorithm 1: On input p, q, and ε

1 **for** $i = 1$ **to** $\log(2d/\varepsilon)$ **do**

2 **if** ℓ_1-`Closeness-Tester`$(p^{(i)}, q^{(i)}, \frac{\varepsilon 2^{i-2}}{d\log(2d/\varepsilon)}, \frac{1}{3\log(2d/\varepsilon)})$ rejects **then reject**

3 **end**

4 **accept**

(ii) $\tilde{O}((1/\varepsilon)^3)$ *if* $d = 3$,

(iii) $\tilde{O}((1/\varepsilon)^4)$ *if* $d = 4$ *or* 5, *and*

(iv) $\tilde{O}((2d/\varepsilon)^{2d/3})$ *if* $d \geq 6$.

Proof. We start with correctness. If $p = q$, then $p^{(i)} = q^{(i)}$ for all i, so by the union bound, the probability that the algorithm rejects is at most $\log(2d/\varepsilon) \cdot \frac{1}{3\log(2d/\varepsilon)} = 1/3$.

If, on the other hand, $EMD(p, q) > \varepsilon$, then by Fact 6,

$$d\left(\sum_{i=1}^{\log(2d/\varepsilon)} \frac{1}{2^{i-1}} \cdot \left\|p^{(i)} - q^{(i)}\right\|_1\right) > \frac{\varepsilon}{2}.$$

It follows by the averaging argument that there exists an index i such that

$$\left\|p^{(i)} - q^{(i)}\right\|_1 > \frac{\varepsilon 2^{i-2}}{d\log(2d/\varepsilon)}.$$

Hence, for that index i, the ℓ_1-closeness tester in Line 2 will reject (with probability $2/3$).

Now let us analyze the number of samples the algorithm needs. In the i^{th} iteration of the main loop, $p^{(i)}$ and $q^{(i)}$ has a domain with $n_i = 2^{di}$ elements, and we need to run an ℓ_1-closeness tester with a distance parameter of $\varepsilon_i = \frac{\varepsilon 2^{i-2}}{\log(2d/\varepsilon)}$. Applying Fact 7, we get a sample complexity of the minimum of

$$\tilde{O}(n_i^{2/3}\varepsilon_i^{-4}) = \tilde{O}\left(2^{(2d/3-4)i}\left(\frac{d}{\varepsilon}\right)^4\right)$$

and

$$\tilde{O}(n_i\varepsilon_i^{-2}) = \tilde{O}\left(2^{(d-2)i}\left(\frac{d}{\varepsilon}\right)^2\right).$$

If $d \leq 2$, both quantities are maximized when $i = 1$; if $d \geq 6$, they are maximized when $i = \log(2d/\varepsilon)$; and if $d = 3, 4$ or 5, the first is maximized when $i = 1$ and the second when $i = \log(2d/\varepsilon)$. The complexities in the theorem immediately follow. □

3.2 Additive-error estimation

As we have seen, for ℓ_1-closeness testing, sometimes it is to our advantage to simply estimate each probability value, rather than use the more sophisticated algorithm of [5]. This seemingly naive approach has another advantage: it gives an actual numeric estimate of the distances, instead of just an accept/reject answer. Here, we use this approach to obtain an additive-error estimator of the EMD between two distributions over $M \subset [0, 1]^d$ as follows.

Algorithm 2: On input p, q, ε

1 Let G be the grid on $[0,1]^d$ with side length $\frac{\varepsilon}{4d}$, and let P and Q be the distributions induced by p and q on G, with weights in each cell concentrated at the center
2 Take $O((4d/\varepsilon)^{d+2})$ samples from P and $O((4d/\varepsilon)^{d+2})$ samples from Q, and let $\tilde{P}$ and $\tilde{Q}$ be the resulting empirical distributions
3 **return** $EMD(\tilde{P}, \tilde{Q})$

Theorem 9. *Algorithm 2 is an additive-error estimator for EMD with sample complexity* $O((4d/\varepsilon)^{d+2})$.

Proof. Observe that $|G| = (4d/\varepsilon)^d$, so G has $2^{(4d/\varepsilon)^d}$ subsets. By the Chernoff bound, with the $O((4d/\varepsilon)^{d+2})$ samples from p, we can guarantee for each $S \subseteq G$ that $|P(S) - \tilde{P}(S)| > \frac{\varepsilon}{4d}$ with probability at most $2^{-(4d/\varepsilon)^d}/3$. By the union bound, with probability at least $2/3$, all subsets of G will be approximated to within an additive $\frac{\varepsilon}{4d}$. In that case,

$$\|P - \tilde{P}\|_1 = \sum_{c \in G} |P(c) - \tilde{P}(c)| = 2 \max_{S \subseteq G} |P(S) - \tilde{P}(S)| \leq \frac{\varepsilon}{2d}.$$

We then have, by Lemma 4, $EMD(P, \tilde{P}) \leq \varepsilon/4$. Further, since each cell has radius $\varepsilon/4$, we have $EMD(p, P) \leq \varepsilon/4$, giving us by the triangle inequality, $EMD(p, \tilde{P}) \leq \varepsilon/2$. Similarly, $EMD(q, \tilde{Q}) \leq \varepsilon/2$, so again by triangle inequality, we get

$$|EMD(p,q) - EMD(\tilde{P}, \tilde{Q})| \leq EMD(p, \tilde{P}) + EMD(q, \tilde{Q}) = \varepsilon,$$

completing our proof. □

3.3 Lower bounds

We can show that our tester is optimal for the 1-dimensional and 2-dimensional domains by a simple argument:

Theorem 10. *Let* $\mathcal{A}$ *be an EMD-closeness tester for distributions over any domain* M *of diameter* 1. *Then* $\mathcal{A}$ *requires* $\Omega((1/\varepsilon)^2)$ *samples.*

Proof. Let $x, y \in M$ be at distance 1 apart. Consider two distributions p and q over $\{x, y\}$ given by $p(x) = p(y) = 1/2$, $q(x) = 1/2 + \varepsilon$ and $q(y) = 1/2 - \varepsilon$. Clearly $EMD(p, q) = \varepsilon$, and it is a folklore fact that distinguishing p from q requires $\Omega((1/\varepsilon)^2)$ samples. □

Next we prove that for high dimensional domains (i.e., $d \geq 6$), our tester is also essentially optimal in its dependence on $1/\varepsilon$. However, tight bounds for the cases where $d = 3$, 4 or 5 are still open.

Theorem 11. *There is no EMD-closeness tester that works on any* $M \subset [0,1]^d$ *with sample complexity* $o((1/\varepsilon)^{2d/3})$.

Proof. Suppose $\mathcal{A}$ is such an EMD-closeness tester that requires only $o((1/\varepsilon)^{2d/3})$ samples. Then consider the ℓ_1-closeness tester for $\varepsilon = \frac{1}{2}$ in Algorithm 3.

Correctness is easy to see: if $p = q$, then clearly $P = Q$ as well and the tester accepts; alternatively, if $\|p - q\|_1 = \frac{1}{2}$, then by Lemma 4 and the observation that $\|P - Q\|_1 = \|p - q\|_1$,

$$EMD(P, Q) \geq \frac{\|P - Q\|_1}{2} \cdot n^{-1/d} = \frac{1}{4} n^{-1/d},$$

Algorithm 3: On input p, q (ε is set to $\frac{1}{2}$)

1 Let G be a grid on $[0,1]^d$ with side length $n^{-1/d}$
2 Let f be an arbitrary injection from $[n]$ into the lattice points of G
3 Let P and Q be distributions on the lattice points of G induced by f on p and q, resp.
4 **return** $\mathcal{A}(P, Q, \frac{1}{4}n^{-1/d})$

so the tester rejects, as required.

To take a sample from P (or Q), we simply take a sample x from p (or q) and return $f(x)$. Hence, the sample complexity of this tester is

$$o\left(\left(\frac{1}{\frac{1}{4}n^{-1/d}}\right)^{2d/3}\right) = o(n^{2/3}).$$

But this contradicts the lower bound for ℓ_1-closeness testing from [5, 19], completing our proof. □

4 EMD over trees

So far we have considered only underlying ℓ_1-spaces. We will now see what can be done for EMD over tree-metrics[3], and prove the following result.

Theorem 12. *If p and q are distributions over the nodes of a tree T, with edge weight function $w(\cdot)$, then there exists an ε-additive-error estimator for $EMD(p,q)$ that requires only $\tilde{O}((Wn/\varepsilon)^2)$ samples, where $W = \max_e w(e)$, n is the number of nodes of T, and $EMD(p,q)$ is defined with respect to the tree metric of T. Moreover, up to polylog factors, this is optimal.*

Proof. First, let us consider an unweighted tree T over n points (i.e., where every edge has unit weight), with distributions p and q on the vertices. Observe that the minimum cost flow between p and q on T is simply the flow that sends through each edge e just enough to balance p and q on each subtree on either side of e. In other words, if T_e is an arbitrary one of the two trees comprising $T - e$,

$$EMD(p,q) = \sum_e |p(T_e) - q(T_e)|.$$

Then, with $\tilde{O}(n^2/\varepsilon^2)$ samples we can, for every T_e, estimate $p(T_e)$ and $q(T_e)$ to within $\pm\frac{\varepsilon}{2(n-1)}$. This gives us an ε-additive estimator for $EMD(p,q)$.

Generalizing to the case of a weighted tree, where edge e has weight $w(e)$, we have

$$EMD(p,q) = \sum_e w(e)|p(T_e) - q(T_e)|.$$

It then suffices to estimate each $p(T_e)$ and $q(T_e)$ term to within $\pm\frac{\varepsilon}{2w(e)(n-1)}$. Thus, $\tilde{O}((Wn/\varepsilon)^2)$ samples suffice, where $W = \max_e w(e)$.

Note that what we get is not only a closeness tester but also an additive-error estimator. In fact, even if we only want a tester, this is the best we can do: in the case where T is a line graph (with diameter $n-1$), the standard biased-coin lower bound implies we need $\Omega(n^2/\varepsilon^2)$ samples. □

[3] In a the tree-metric space, the domain is the set of vertices in a tree and the distance between any two vertices is defined to be the length of the shortest path connecting them in the tree.

For trees with bounded diameter, it is possible to get a better bound for testing. Assuming that the diameter of the tree is bounded by Δ, we have the following theorem.

Theorem 13. *If p and q are distributions over the nodes of an unweighted tree T, then there exists a closeness tester for $EMD(p,q)$ that, assuming an ℓ_1-closeness tester with query complexity $\tilde{O}(n^{2/3}\varepsilon^{\rho})$, requires $\tilde{O}(n^{2/3}\Delta^{-\rho-2/3}\varepsilon^{\rho})$ samples, where n is the number of nodes of T, Δ is the diameter of T, and $EMD(p,q)$ is defined with respect to the tree metric of T.*

Definition 14. *A* k-cover *of a tree T is a set of subtrees of T such that each vertex of T belongs to at least one subtree and the size of each subtree is at most k.*

Lemma 15. *For positive integer k, any tree T of size n has a $2k$-cover of size at most $\lceil n/k \rceil$. Moreover, any pair of subtrees in the cover is edge-disjoint and intersect in at most one vertex.*

Proof. We will use induction on n. For a tree of size up to $2k$, the singleton cover consisting of just itself satisfies our lemma. Now consider a tree T of size $n > 2k$. Let us root T at an arbitrary vertex r, and start walking from r down any leaf-bound path. Let v be the last vertex on this path for which the subtree rooted at v, denoted T_v, is of size at least $2k$, and let C_v denote its children.

If $|T_u| \geq k$ for some $u \in C_v$, then $T - T_u$ is a tree of size at most $n-k$, so by inductive hypothesis it has a $2k$-cover of size at most $\lceil (n-k)/k \rceil = \lceil n/k \rceil - 1$. Adding T_u to this cover makes it a $2k$-cover of T of size at most $\lceil n/k \rceil$, satisfying our lemma.

In case $|T_u| \leq k-1$ for every $u \in C_v$, construct a subset $A \subset C_v$ as follows: add elements $u \in C_v$ to A in arbitrary order until $\sum_{u \in A} |T_u| \geq k$. Because each u we add contributes at most $k-1$ to this sum, when this is procedure terminates, we will have $\sum_{u \in A} |T_u| \in [k, 2k-2]$. The subtree T' formed by v together with every T_u, $u \in A$, is therefore of size $|T'| \in [k+1, 2k-1]$. Applying the inductive hypothesis to the tree $T'' := T - T' + \{v\}$, which has size $|T''| \leq n-(k+1)+1 = n-k$, gives us a $2k$-cover (of T'') of size at most $\lceil n/k \rceil - 1$, so adding T' gives us a $2k$-cover of T of size at most $\lceil n/k \rceil$, completing our proof. □

Proof of Theorem 13. Applying Lemma 15, let $\mathcal{C} = \{T_1, \ldots, T_m\}$ be a 2Δ-cover of T, where Δ is the diameter of T. Define $p^{\mathcal{C}}$ to be the distribution over $\mathcal{C}$ induced by p, where a node of T belonging to multiple subtrees in $\mathcal{C}$ is assigned to one of them arbitrarily. Specifically, and wlog, let

$$p^{\mathcal{C}}(T_i) = \sum_{u \in T_i \setminus (T_1 \cup \cdots \cup T_{i-1})} p(u).$$

Define $q^{\mathcal{C}}$ similarly. Note that $m \leq \lceil n/\Delta \rceil$, and $|T_i| \leq 2\Delta$ for every $i \in [m]$. We can label the edges in each T_i with distinct values from $[2\Delta]$. Note that since the subtrees are edge-disjoint, we do not need to worry about conflict resolution. Now, breaking all edges in T labeled some fixed $i \in [2\Delta]$ results in a forest F_i with at most $2m$ trees, and we can define p^{F_i} and q^{F_i} similarly to the above. We then claim that Algorithm 4 is the specified closeness tester for $EMD_T(p,q)$.

We start by computing the sample complexity of the algorithm. Recall that $p^{\mathcal{C}}$ and $q^{\mathcal{C}}$, as well as p^{F_i} and q^{F_i}, for $i \in [2\Delta]$, are all defined over $O(n/\Delta)$ points. Moreover, we will not assume any independence between our calls to ℓ_1-`Closeness-Tester`, so we can reuse samples from p and q to generate samples for all the above distributions. Hence, our algorithm's sample complexity is dominated by one of the calls to ℓ_1-`Closeness-Tester` inside the loop, or

$$\tilde{O}\left(\left(\frac{n}{\Delta}\right)^{2/3}\left(\frac{\varepsilon}{4\Delta}\right)^{\rho}\right) = \tilde{O}(n^{2/3}\Delta^{-\rho-2/3}\varepsilon^{\rho}),$$

Algorithm 4: On input p, q, and ε

if ℓ_1-`Closeness-Tester`$(p^{\mathcal{C}}, q^{\mathcal{C}}, \frac{\varepsilon}{2\Delta}, \frac{1}{6})$ rejects **then reject**
for $i = 1$ **to** 2Δ **do**
    if ℓ_1-`Closeness-Tester`$(p^{F_i}, q^{F_i}, \frac{\varepsilon}{4\Delta}, \frac{1}{12\Delta})$ rejects **then reject**
end
accept

as claimed.

It remains to prove correctness. Consider first the case where $p = q$. Then by construction, $p^{\mathcal{C}} = q^{\mathcal{C}}$ and $p^{F_i} = q^{F_i}$ for every $i \in [2\Delta]$, so all our calls to ℓ_1-`Closeness-Tester` accept with the specified error probabilities. Taking a union bound over all errors gives us a rejection probability of at most $1/3$, as required.

Next, assume $EMD_T(p, q) > \varepsilon$. We claim that at least one of the calls to ℓ_1-`Closeness-Tester` *should* reject, so that by the same correctness guarantees as above, they all answer as they should and we correctly reject with probability at least $2/3$. Suppose, in contradiction to our claim, that $\|p^{\mathcal{C}} - q^{\mathcal{C}}\|_1 \leq \frac{\varepsilon}{2\Delta}$ and $\|p^{F_i} - q^{F_i}\|_1 \leq \frac{\varepsilon}{4\Delta}$ for every $i \in [2\Delta]$. Then consider the following flow from p to q: (1) within each subtree $T_i \in \mathcal{C}$, move probability weight so as to match p to q as much as possible (modulo the difference in total probability weight, $|p^{\mathcal{C}}(T_i) - q^{\mathcal{C}}(T_i)|$), then (2) move the unmatched probability weight across different subtrees to completely match q. Observe that step (1), similarly to the proof of Theorem 12, is done by moving just enough weight across each edge to balance the two sides, so we can bound the total cost of (1) over all subtrees by $\frac{\varepsilon}{4\Delta} \cdot 2\Delta = \frac{\varepsilon}{2}$. Now, in step (2), we are moving a total probability weight of $\leq \frac{\varepsilon}{2\Delta}$ across a distance of $\leq \Delta$, so this contributes a cost of no more than $\frac{\varepsilon}{2}$. This gives us a total cost of $\leq \varepsilon$, contradicting our assumption. □

Theorem 16. *There is no closeness tester for EMD over a tree of size n and diameter Δ using $o(n^{2/3}\Delta^{-\rho-2/3}\varepsilon^{\rho})$ samples, assuming a lower bound of $\Omega(n^{2/3}\varepsilon^{\rho})$ for ℓ_1-closeness testing.*

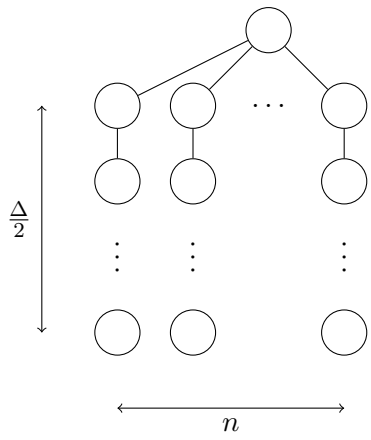

Figure 1: A regular star tree with n leaves and diameter Δ.

Proof. Assume, for contradiction, that there is a closeness tester $\mathcal{A}$ for EMD over a tree (of size n and diameter Δ) using $o(n^{2/3}\Delta^{-\rho-2/3}\varepsilon^{\rho})$ samples. We construct a closeness tester for ℓ_1 over $[n]$ as follows. Let T be a regular star graph (tree) with n wings, each consisting of $\Delta/2$ vertices (see Figure 1). Note that T is of size $n\Delta/2+1$ and diameter Δ. We then map each $i \in [n]$ to the i^{th} leaf of T, and run $\mathcal{A}$ on the induced distributions on T, with closeness parameter $\Delta\varepsilon/2$. Notice that every leaf of T is at distance exactly Δ from every other leaf, so EMD of these distributions (on T) is exactly $\Delta/2$ times the ℓ_1 distance of the originals, and our described ℓ_1-closeness tester is correct. Moreover, $\mathcal{A}$, and therefore the ℓ_1-closeness tester, uses only

$$o((n\Delta)^{2/3}\Delta^{-\rho-2/3}(\Delta\varepsilon)^{\rho}) = o(n^{2/3}\varepsilon^{\rho})$$

samples, contradicting the lower bound for ℓ_1-closeness testing. □

5 Clusterable distributions

In this section, we return to the domain of hypercubes. For general distributions over hypercubes, the main technique of our algorithms is to forcibly divide the distributions' support into several small "clusters". It is therefore natural to consider what improvements are possible when the distributions are inherently clusterable. We are able to obtain a substantial improvement from the exponential dependence on the dimension d that we had in the general case. But first, we need the following stand-alone lemma.

Lemma 17. *Let p and q be distributions on M with diameter 1, and $\mathcal{M} = \{M_1, \ldots, M_k\}$ be a partition of M wherein $\mathrm{diam}(M_i) \leq \Gamma$ for every $i \in [k]$. Let P and Q be distributions on $\mathcal{M}$ induced by p and q, resp. Then $EMD(p,q) \leq \frac{||P-Q||_1}{2} + \Gamma$.*

Proof. Let us define distribution p' by moving some of the probability weight of p *between* M_i's (taking from and depositing anywhere within the respective M_i's) in such a way that p' induces Q on $\mathcal{M}$. This is effectively a flow from P to Q where all distances are bounded by 1, so by Lemma 4 it can be done at cost at most $\frac{||P-Q||_1}{2}$. It follows that $EMD(p,p') \leq \frac{||P-Q||_1}{2}$.

Then, having equalized the probability weights of each M_i, let us move the probability weight of p' *within* each M_i to precisely match q. This might require moving everything (i.e., 1), but the distance anything is moved is at most Γ, so $EMD(p',q) \leq \Gamma$ and the lemma follows by triangle inequality. □

In the following, p and q are assumed to be distributions over $[0,1]^d$.

Theorem 18. *If the combined support of distributions p and q can be partitioned into k clusters of diameter $\varepsilon/2$, and we are given the k centers, then there exists an EMD-closeness tester for p and q that requires only $\tilde{O}(k^{2/3}(d/\varepsilon)^4)$ samples.*

Proof. Let us denote the set of centers by $\mathcal{C} = \{C_1, \ldots, C_k\}$. Consider the distributions P and Q on $\mathcal{C}$ induced by p and q, respectively, by assigning each point to its nearest center. If $EMD(p,q) > \varepsilon$, by Lemma 17, $\frac{\|P-Q\|_1}{2}(d) > \varepsilon/2$. We can, of course, obtain samples from P and Q by sampling from p and q, respectively, and returning the nearest center. Our problem thus reduces to ℓ_1-testing for $(\frac{\varepsilon}{d})$-closeness over k points, which requires $\tilde{O}(k^{2/3}(d/\varepsilon)^4)$ samples using the ℓ_1-tester from [5]. □

If we do not assume knowledge of the cluster centers, then we are still able to obtain the following only slightly weaker result.

Theorem 19. *If the combined support of p and q can be partitioned into k clusters of diameter $\varepsilon/4$, then even without knowledge of the centers there exists an EMD-closeness tester for p and q that requires only $\tilde{O}(kd/\varepsilon + k^{2/3}(d/\varepsilon)^4) \leq \tilde{O}(k(d/\varepsilon)^4)$ samples.*

To prove this, we need the following result by Alon et al. that was implicit in Algorithm 1 from [1].

Lemma 20. *(Algorithm 1 from [1]) There exists an algorithm which, given distribution p, returns $k' \leq k$ representative points if p is (k, b)-clusterable, or rejects with probability $2/3$ if p is γ-far from $(k, 2b)$-clusterable, and which requires only $O(k \log k/\gamma)$ samples from p. Moreover, if the k' points are returned, they are with probability $2/3$ the centers of a $(k, 2b)$-clustering of all but a γ-weight of p.*

Proof. (of Theorem 19) By the lemma, if our distributions are $(k, \varepsilon/4)$-clusterable, using $\tilde{O}(kd/\varepsilon)$ samples we obtain a $(k', \varepsilon/2)$-clustering of all but an $\frac{\varepsilon}{4d}$-fraction of the support of p and q, with centers $\mathcal{C}'$. Note that the unclustered probability mass contributes at most $\varepsilon/4$ to the EMD. The theorem then follows from an identical argument as that of Theorem 18 (since we now know the centers). □

Note that in general, if we assume (k, b)-clusterability, this implies $((2b/\varepsilon)^d k, \varepsilon/2)$-clusterability (by packing ℓ_1 balls), where knowledge of the super-cluster centers also implies knowledge of the sub-cluster centers. Similarly, in the unknown centers case, (k, b)-clusterability implies $((4b/\varepsilon)^d k, \varepsilon/4)$-clusterability. Unfortunately, in both cases we reintroduce exponential dependence on d, so clusterability only really helps when the cluster diameters are as assumed above.

6 Open directions

While we have provided some initial investigation on closeness testing of distributions under the EMD metric, the complete understanding of this problem is still wide open. Below is a list of several open problems suggested by our work.

- *Bounds for low dimensional space*. We have shown matching lower and upper bounds for closeness testing when the distributions are over a d-dimensional hypercube for $d = 1$ and $d = 2$. It is open whether one can obtain matching bounds for $d = 3, 4$, and 5.
- *Better bounds for approximating the distance*. In the hypercube domain case, we gave a simple algorithm for estimating the distance between two distributions up to an additive error. Improvement on the sample complexity of this algorithm and a lower bound would be very interesting. It is also open whether one can get a multiplicative approximation of the distance. Since we cannot hope to be able to distinguish identical distributions from distributions of infinitesimally small distance, the sample complexity of the algorithm would have to depend on the distance between the two distributions. It would be even more interesting if one can get a constant multiplicative approximation. Because the approach of reducing EMD to ℓ_1 incurs a logarithmic distortion, some other approach is needed to overcome this barrier.

References

[1] N. Alon, S. Dar, M. Parnas, and D. Ron. Testing of clustering. In *FOCS '00: Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, page 240, Washington, DC, USA, 2000. IEEE Computer Society.

[2] Alexandr Andoni, Piotr Indyk, and Robert Krauthgamer. Earth mover distance over high-dimensional spaces. In *SODA '08: Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 343–352, Philadelphia, PA, USA, 2008. Society for Industrial and Applied Mathematics.

[3] Y. Bartal. Probabilistic approximation of metric spaces and its algorithmic applications. In *FOCS '96: Proceedings of the 37th Annual Symposium on Foundations of Computer Science*, page 184, Washington, DC, USA, 1996. IEEE Computer Society.

[4] T. Batu, L. Fortnow, E. Fischer, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *FOCS '01: Proceedings of the 42nd IEEE symposium on Foundations of Computer Science*, page 442, Washington, DC, USA, 2001. IEEE Computer Society.

[5] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *FOCS '00: Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, page 259, Washington, DC, USA, 2000. IEEE Computer Society.

[6] Moses S. Charikar. Similarity estimation techniques from rounding algorithms. In *STOC '02: Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388, New York, NY, USA, 2002. ACM.

[7] Scott Cohen and Leonidas Guibas. The earth mover's distance under transformation sets. In *ICCV '99: Proceedings of the International Conference on Computer Vision-Volume 2*, page 1076, Washington, DC, USA, 1999. IEEE Computer Society.

[8] Jittat Fakcharoenphol, Satish Rao, and Kunal Talwar. A tight bound on approximating arbitrary metrics by tree metrics. In *STOC '03: Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 448–455, New York, NY, USA, 2003. ACM.

[9] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. *Electronic Colloquium on Computational Complexity (ECCC)*, 7(20), 2000.

[10] Piotr Indyk. A near linear time constant factor approximation for euclidean bichromatic matching (cost). In *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 39–42, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[11] Piotr Indyk and Nitin Thaper. Fast image retrieval via embeddings. In *3rd International Workshop on Statistical and Computational Theories of Vision*. ICCV, 2003.

[12] Shmuel Peleg, Michael Werman, and Hillel Rom. A unified approach to the change of resolution: Space and gray-level, 1989.

[13] Y. Rubner, L. J. Guibas, and C. Tomasi. The earth movers distance, multi-dimensional scaling, and color-based image retrieval. In *APRA Image Understanding Workshop*, pages 661–668, May 1997.

[14] Y. Rubner and C. Tomasi. Texture metrics. In *Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on*, volume 5, pages 4601–4607 vol.5, Oct 1998.

[15] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Computer Vision, 1998. Sixth International Conference on*, pages 59–66, 1998.

[16] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40(2):99–121, November 2000.

[17] M. A. Ruzon and C. Tomasi. Color edge detection with the compass operator. volume 2, page 166 Vol. 2, 1999.

[18] M. A. Ruzon and C. Tomasi. Corner detection in textured color images. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1039–1045 vol.2, 1999.

[19] Paul Valiant. Testing symmetric properties of distributions. In *STOC '08: Proceedings of the 40th annual ACM symposium on Theory of computing*, pages 383–392, New York, NY, USA, 2008. ACM.