Sublinear Algorithms for Earth Mover's Distance

by

Khanh Do Ba

B.A. Mathematics and Computer Science
Dartmouth College, 2006

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
MAR 0 5 2009
LIBRARIES

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2009

©2009 Massachusetts Institute of Technology. All rights reserved.

Signature of Author: .....
Department of Electrical Engineering and Computer Science
December 18, 2008

Certified by: .....
Piotr Indyk
Associate Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by: .....
Terry P. Orlando
Professor of Electrical Engineering and Computer Science
Chairman, Committee on Graduate Students

ARCHIVES

Abstract

We study the problem of estimating the Earth Mover's Distance (EMD) between probability distributions when given access only to samples. We give closeness testers and additive-error estimators over domains in $[0, \Delta]^d$, with sample complexities independent of domain size – permitting the testability even of continuous distributions over infinite domains. Instead, our algorithms depend on other parameters, such as the diameter of the domain space, which may be significantly smaller. We also prove lower bounds showing our testers to be optimal in their dependence on these parameters. 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 optimal algorithms.

1 Introduction

In traditional algorithmic settings, algorithms requiring linear time and/or space are generally considered to be highly efficient; sometimes 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 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, and algorithms sublinear in time and sample complexity exist to distinguish whether two distributions are identical or ε-far from each other [4, 7, 18]. However, under the L_1 distance, for instance, the sample complexity, though sublinear, can be no smaller than $n^{2/3}$ (where n is the domain size), which may be prohibitively large [4, 18].

Fortunately, in many situations 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 [11]. It has since proven to be important in computer graphics and vision [13, 12, 14, 6, 16, 17, 15], 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, 9, 5, 8]. 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 exact probability of any domain element can be queried. As far as we know, no work has been done on estimating and closeness testing of EMD when given access only to i.i.d. samples of 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 complexities depend only on the *diameter* of the domain space, which in many cases can be significantly smaller. 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.

Specifically, if p and q are distributions over $M \subset [0, \Delta]^d$ (where d is a constant), we can

- estimate $EMD(p, q)$ to within an additive error of ε with $\tilde{O}((\Delta/\varepsilon)^{d+O(1)})$ samples,
- distinguish whether $p = q$ or $EMD(p, q) > \varepsilon$ with $\tilde{O}((\Delta/\varepsilon)^{2d/3+O(1)})$ samples, and

- if q is known, distinguish whether $p = q$ or $EMD(p, q) > \varepsilon$ with $\tilde{O}((\Delta/\varepsilon)^{d/2+O(1)})$ samples.

We also give lower bounds that imply these results to be essentially optimal (up to small $poly(\Delta/\varepsilon)$ factors). In the case of $d = 1$ or 2, upper and lower bounds for both testers become $\Theta((\Delta/\varepsilon)^2)$.

Additionally, we consider natural assumptions on the data that might make the problem easier, and give an improved algorithm in the case our input distributions is highly clusterable. Finally, it is natural to consider the EMD over domains endowed with a metric other than L_1 distance. We give an optimal (upto polylogarithmic factors) algorithm for estimating EMD over tree metrics.

2 Preliminaries

2.1 Earth Mover's Distance

We start with the following definition.

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 distribution p on S and demand 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 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.*

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.

Since the magnitude of the EMD depends on the distances in the underlying metric space M, we must assume that M is of bounded diameter. In particular, we will in this paper focus mostly on the case where $M \subset [0, \Delta]^d$, for some $\Delta > 0$, endowed with the L_1 metric.

Finally, we define a closeness tester for the EMD in the usual way as follows.

Definition 3 *Let p, q be two distributions on (finite) metric space M. An* EMD-closeness tester *is an algorithm which takes as input samples from p and q, together with a real number $\varepsilon > 0$, and guarantees that*

(1) if $p = q$, then it accepts with probability at least $2/3$, and

(2) if $EMD(p, q) > \varepsilon$, then it rejects with probability at least $2/3$.

Note that it is also possible to define EMD for any two probability measures p and q in an infinite (continuous) metric space $(\mathcal{M}, \delta_{\mathcal{M}})$, $\mathcal{M} \subset [0, \Delta]^d$, by using 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 $\varepsilon/4$-net, $(\mathcal{M}, \delta_{\mathcal{M}})$ can be discretized into a finite metric (M, δ) in such a way that the EMD distance of any two probability measures only increases or decreases at most $\varepsilon/2$. Therefore, an EMD-closeness tester with additive error $\varepsilon/2$ for (M, δ) is also a valid EMD-closeness tester for $(\mathcal{M}, \delta)$ with additive error ε.

2.2 Properties of EMD

Let us get a firmer handle on the EMD by relating it to L_1 distance, via the following lemmas.

Lemma 4 *If p and q are distributions on (M, δ), there exists a minimum cost satisfying flow f from S to T (as defined in Def. 2) such that the total amount sent by f across edges with non-zero cost is exactly $||p - q||_1/2$.*

Proof Let $a = \sum_x \min\{p(x), q(x)\}$ and $A = \sum_x \max\{p(x), q(x)\}$, so that $A - a = ||p - q||_1$ and $A + a = 2$. Observe that the total amount sent from S to T is 1, and the maximum possible amount sent across edges with zero cost is a, which leaves at least $1 - a = ||p - q||_1/2$ to be sent across non-zero cost edges.

On the other hand, suppose that for some x, an optimal flow through the edge (s_x, t_x) is less than $\min\{p(x), q(x)\}$. Then there exist points y and z such that f sends at least $\alpha > 0$ from s_x to t_y and from s_z to t_x. We can replace this partial flow, which costs $\alpha(\delta(x, y) + \delta(x, z))$, with one sending α from s_x to t_x and from s_z to t_y, which costs $\alpha(\delta(y, z))$. Doing so, we will not increase the total cost (by triangle inequality), nor will we affect the rest of the flow. We can therefore repeated the procedure until we obtain an optimal flow that saturates every zero-edge. ■

Corollary 5 *If p and q are distributions on M, where M has minimum distance δ and diameter Δ, then*

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

Lemma 6 *Let p and q be distributions on M with diameter Δ, 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 $EMD(p,q) \leq \frac{||P-Q||_1}{2} \cdot \Delta + \Gamma$.*

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 Δ, so by Lemma 4 it can be done at cost at most $\frac{||P-Q||_1}{2} \cdot \Delta$. It follows that $EMD(p,p') \leq \frac{||P-Q||_1}{2} \cdot \Delta$.

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. ■

2.3 Some tools from previous work

Here, for completeness, we state some results from previous work that we will make use of later. First, we define some useful distributions described in [9] for testing closeness between distributions over subsets of $[0,\Delta]^d$.

Definition 7 *Given distribution p over $M \subset [0,\Delta]^d$ and a positive integer i, let $G^{(i)}$ be a grid with side length $\frac{\Delta}{2^i}$ over $[0,\Delta]^d$ 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 lemma from [9] relating the EMD of two distributions to a weighted sum of the L_1 distances of their i-coarsenings.

Lemma 8 *For any two distributions p and q over $M \subset [0,\Delta]^d$,*

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

Having established the various relationships between the EMD and the L_1 distance, we will make use of the result from [4] below for testing closeness of distributions in L_1 as a subroutine.

Theorem 9 *Given access to samples from two distributions p and q over M, where $|M| = n$, there exists an algorithm that takes $O(n^{2/3}\varepsilon^{-4}\log n \log(1/\delta))$ samples and (1) accepts with probability at least $1-\delta$ if $p = q$ and (2) rejects with probability at least $1-\delta$ if $\|p-q\|_1 > \varepsilon$.*

Alternatively, via a simple Chernoff bound analysis similar to [3], we can show that a whole distribution can be approximated efficiently, giving us another closeness tester for L_1.

Lemma 10 *Given access to samples from a distribution p over M, where $|M| = n$, and parameters $\varepsilon, \delta, t > 0$, there exists an algorithm that takes $O(t^{-1}\varepsilon^{-2}\log n \log(1/\delta))$ samples and outputs a distribution $\tilde{p}$ over M such that with probability at least $1 - \delta$*

$$(1-\varepsilon)\max\{p(i), t\} \leq \tilde{p}(i) \leq (1+\varepsilon)\max\{p(i), t\}$$

for every $i \in M$.

As a result, we can simply estimate p and q with $\tilde{p}$ and $\tilde{q}$, respectively, and compute their L_1 distance to get the following alternative tester.

Theorem 11 *Given access to samples from two distributions p and q over M, where $|M| = n$, there exists an algorithm that takes $O(n\varepsilon^{-2}\log n \log(1/\delta))$ samples and (1) accepts with probability at least $1 - \delta$ if $p = q$ and (2) rejects with probability at least $1 - \delta$ if $\|p - q\|_1 > \varepsilon$.*

3 Closeness tester

In this section, we consider the EMD-closeness testing problem when the domain is $M \subset [0, \Delta]^d$. The main idea behind the algorithm is to embed EMD into the L_1 metric and use an L_1 closeness tester (Theorems 9 and 11) to test the resulting distributions. Recall from the preliminaries, $p^{(i)}$ and $q^{(i)}$ are the i-coarsening approximations of p and q. We have the following algorithm, where the subroutine $\texttt{L}_1\texttt{-Closeness-Tester}(p, q, \varepsilon, \delta)$ is an L_1-closeness tester on distributions p and q with distance parameter ε and failure probability δ.

$\texttt{EMD-Closeness-Tester}(p, q, \varepsilon)$
1 **for** $i = 1$ **to** $\log(2\Delta d/\varepsilon)$ **do**
2 **if** $\texttt{L}_1\texttt{-Closeness-Tester}(p^{(i)}, q^{(i)}, \frac{\varepsilon 2^{i-2}}{\Delta d \log(2\Delta d/\varepsilon)}, \frac{1}{3\log(2\Delta d/\varepsilon)})$ rejects **then**
3 **reject**
4 **accept**

Note that our subroutine takes advantage of whichever tester (Theorem 9 or 11) requires fewer samples. Specifically, when d is small, the domains of the i-coarsenings of p and q are small and ε is the bottleneck, so we use Theorem 11. On the other hand, if d is large, the sizes of these domains become the bottleneck and we use Theorem 9. This gives us the following theorem.

Theorem 12 *The above is an EMD-closeness tester for distributions over $M \subset [0, \Delta]^d$ that takes $\tilde{O}((2\Delta d/\varepsilon)^{2d/3})$ samples when $d \geq 6$ and $\tilde{O}((\Delta/\varepsilon)^2)$ samples when $d \leq 2$.*

Proof 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 $\log \frac{2\Delta d}{\varepsilon} \cdot \frac{1}{3\log \frac{2\Delta d}{\varepsilon}} = \frac{1}{3}$.

If, on the other hand, $EMD(p, q) > \varepsilon$, then by Lemma 8,

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

It follows by the pigeonhole principle that there exists an index i such that

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

Hence, for that index i, the L_1-closeness tester in Step 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 $\varepsilon_i = \frac{\varepsilon 2^{i-2}}{\Delta \log(2\Delta d/\varepsilon)}$. Consider the following two cases:

- $d \geq 6$: Using the algorithm of Theorem 9, we get a sample complexity of

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

  This quantity is maximized when $i = \log(2\Delta d/\varepsilon)$, which gives us a total complexity of $\tilde{O}((2\Delta d/\varepsilon)^{2d/3})$.

- $d \leq 2$: Using the algorithm of Theorem 11, we get a sample complexity of

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

  This quantity is maximized when $i = 1$, giving us a total complexity of $\tilde{O}((\Delta/\varepsilon)^2)$.

■

If one of the distributions is explicitly known, we can use the corresponding L_1-closeness tester with sample complexity $O(n^{1/2}\varepsilon^{-2}\log n)$ from [3] to similarly get the following theorem.

Theorem 13 *There exists an EMD-closeness tester for distributions over $M \subset [0,\Delta]^d$, where one is explicitly known, that takes $\tilde{O}((2\Delta/\varepsilon)^{d/2})$ samples when $d \geq 4$.*

4 Additive-error estimation

We have seen that in L_1-closeness testing, sometimes it is to our advantage to simply estimate each probability value, rather than use the more sophisticated algorithm of [4]. 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 approximation of the EMD of two unknown distributions over $M \subset [0,\Delta]^d$ as follows.

`EMD-Approx`(p, q, ε)

1 Let G be the grid on $[0,\Delta]^d$ 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 $O((4d\Delta/\varepsilon)^{d+2})$ samples from P and Q, and let $\tilde{P}$ and $\tilde{Q}$ be the resulting empirical distributions
3 **return** $EMD(\tilde{P}, \tilde{Q})$

Theorem 14 `EMD-Approx` *takes* $O((4d\Delta/\varepsilon)^{d+2})$ *samples from* p *and* q *and, with probability* $2/3$*, outputs an* ε*-additive approximation of* $EMD(p,q)$*.*

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

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

We then have, by Corollary 5, $EMD(P, \tilde{P}) \leq \varepsilon/4$. 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. ■

5 Lower bounds

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

Theorem 15 *Let* $\mathcal{A}$ *be an EMD-closeness tester of distributions over any domain* $M \subset [0, \Delta]$*. Then* $\mathcal{A}$ *requires* $\Omega((\Delta/\varepsilon)^2)$ *samples.*

Proof Consider two distributions p and q over the domain $\{0, \Delta\}$, where p is the distribution that puts the weight of $\frac{1}{2}$ at 0 and Δ and q is the distribution that puts the weight of $1/2 + \varepsilon/\Delta$ at 0 and the weight of $1/2 - \varepsilon/\Delta$ at Δ. Clearly $EMD(p,q) = \varepsilon$, and it is a classic result that distinguishing p from q requires $\Omega((\Delta/\varepsilon)^2)$ samples. ■

Clearly this also implies the same lower bound for 2-dimensional domains, making our algorithm optimal in those cases. Next we prove that our d-dimensional tester is also essentially optimal in its dependence on Δ/ε.

Theorem 16 *There is no EMD-closeness tester that works on any* $M \subset [0, \Delta]^d$ *that takes* $o((\Delta/\varepsilon)^{2d/3})$ *samples.*

Proof Suppose $\mathcal{A}$ is an EMD-closeness tester that requires only $o((\Delta/\varepsilon)^{2d/3})$ samples. Then consider the following L_1-closeness tester for $\varepsilon = 1$:

L_1-Tester$(p, q, \varepsilon = 1)$

1 Let G be a grid on $[0, \Delta]^d$ with side length $\Delta n^{-1/d}$

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** $\mathcal{A}(P, Q, \frac{1}{2}\Delta n^{-1/d})$

Correctness is easy to see: if $p = q$, then clearly $P = Q$ as well and the tester accepts; alternatively, if $\|p-q\|_1 = 1$, then by Corollary 5 and the observation that $\|P-Q\|_1 = \|p-q\|_1$,

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

so the tester rejects, as required.

Now, 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

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

But this contradicts the lower bound for L_1-closeness testing from [4, 18], completing our proof. ■

As in the previous subsection, if one of the distributions is known, we can similarly use the weaker $\Omega(n^{1/2})$ lower bound for uniformity testing to obtain the following theorem.

Theorem 17 *There is no EMD-closeness tester that works on any $M \subset [0, \Delta]^d$, where one of the input distributions is explicitly known, that takes $o((\Delta/\varepsilon)^{d/2})$ samples.*

6 Clusterable distributions

Since the general technique of our algorithms is to forcibly divide the input distributions into several small "clusters," it is natural to consider what improvements are possible when the distributions are naturally clusterable.

First, consider an easy case in which we know the clustering. That is, suppose both distributions p and q are $(k, \varepsilon/2)$-clusterable (i.e., their combined support can be partitioned into k clusters of diameter $< \varepsilon$), and we are given the k centers, $\mathcal{C} = \{C_1, \ldots, C_k\}$. 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 6, $\frac{\|P-Q\|_1}{2}(d\Delta) > \varepsilon/2$. 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 $(\frac{\varepsilon}{d\Delta})$-closeness over k points, which requires $\tilde{O}(k^{2/3}(d\Delta/\varepsilon)^4)$ samples using the L_1-tester from [4]. This gives us the following theorem, and a substantial improvement on our previously exponential dependence on the dimension 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 $\tilde{O}(k^{2/3}(d\Delta/\varepsilon)^4)$ samples.*

Now, supposed we do not know the clustering. We remove the assumption of knowledge of $\mathcal{C}$ by using the following result by Alon et al.

Theorem 19 *(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.*

Thus, if our distributions are $(k, \varepsilon/4)$-clusterable, using $\tilde{O}(kd\Delta/\varepsilon)$ samples we obtain a $(k', \varepsilon/2)$-clustering of all but an $\frac{\varepsilon}{4d\Delta}$-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 argument above then follows through, giving us the following theorem.

Theorem 20 *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 $\tilde{O}(kd\Delta/\varepsilon + k^{2/3}(d\Delta/\varepsilon)^4) \leq \tilde{O}(k(d\Delta/\varepsilon)^4)$ samples.*

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 $((8b/\varepsilon)^d k, \varepsilon/4)$-clusterability. Unfortunately, in both cases we reintroduce exponential dependence on d, so clusterability only really helps when the cluster diameters are as assumed above.

7 EMD over tree-metrics

So far we have considered only underlying L_1-spaces (or, almost equivalently, L_p). We will now see what can be done for EMD over tree-metrics.

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 $p(T_e)$ to within $\pm\frac{\varepsilon}{2(n-1)}$. Similarly for $q(T_e)$. This gives us the following ε-additive estimator for $EMD(p, q)$.

Note that what we 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.

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 $W = \max_e w(e)$. This gives us

Theorem 21 *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 $W = \max_e w(e)$ and $EMD(p,q)$ is defined with respect to the tree metric of T. Moreover, upto polylog factors, this is optimal.*

Acknowledgments

This thesis is the result of joint work with Huy L. Nguyen, Huy N. Nguyen and Ronitt Rubinfeld.

References

[1] N. Alon, S. Dar, M. Parnas and D. Ron. Testing of clustering. *SIAM J. Discrete Math*, 16:393–417, 2003.

[2] A. Andoni, P. Indyk and R. Krauthgamer. Earth Mover Distance over high-dimensional spaces. *Proc. SODA*, 343–352, 2008.

[3] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld and P. White. Testing random variables for independence and identity. *Proc. FOCS*, 442–451, 2001.

[4] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith and P. White. Testing that distributions are close. *Proc. FOCS*, 259–269, 2000.

[5] M. S. Charikar. Similarity estimation techniques from rounding algorithms. *Proc. STOC*, 380–388, 2002.

[6] S. Cohen and L. Guibas. The Earth Mover's Distance under transformation sets. *Proc. ICCV*, 1076–1083, 1999.

[7] O. Goldreich and D. Ron. On Testing Expansion in Bounded-Degree Graphs. *ECCC*, 7-20, 2000.

[8] P. Indyk. A near linear time constant factor approximation for Euclidean Bichromatic Matching (Cost). *Proc. SODA*, 39–42, 2007.

[9] P. Indyk and N. Thaper. Fast image retrieval via embeddings. *3rd International Workshop on Statistical and Computational Theories of Vision*, 2003.

[10] E. Levina and P. Bickel. The Earth Mover's Distance is the Mallows Distance: some insights from statistics. *Proc. ICCV*, 251–256, 2001.

[11] S. Peleg, M. Werman and H. Rom. A unified approach to the change of resolution: space and gray-level. *Trans. Pattern Analysis and Machine Intelligence*, 11: 739–742.

[12] Y. Rubner and C. Tomasi. Texture metrics. *Proc. ICSMC*, 4601–4607, 1998.

[13] Y. Rubner, C. Tomasi and L. J. Guibas. The Earth Mover's Distance, multi-dimensional scaling, and color-based image retrieval. *Proc. ARPA Image Understanding Workshop*, 661–668, 1997.

[14] Y. Rubner, C. Tomasi and L. J. Guibas. A metric for distributions with applications to image databases. *Proc. IEEE ICCV*, 59–66, 1998.

[15] 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.

[16] M. Ruzon and C. Tomasi. Color edge detection with the compass operator. *Proc. CVPR*, 2:160–166, 1999.

[17] M. Ruzon and C. Tomasi. Corner detection in textured color images. *Proc. ICCV*, 2:1039–1045, 1999.

[18] P. Valiant. Testing Symmetric Properties of Distributions *ECCC*, 2007.