Optimal Random Matchings on Trees and Applications

Jeff Abrahamson[1], Béla Csaba[2⋆], and Ali Shokoufandeh[1⋆⋆]

[1] jeffa,ashokouf@cs.drexel.edu
Dept. of Computer Science
Drexel University
Philadelphia, PA
[2] bela.csaba@wku.edu
Dept. of Mathematics
Western Kentucky University
Bowling Green, KY

Abstract. In this paper we will consider tight upper- and lower-bounds on the weight of the optimal matching for random point sets distributed among the leaves of a tree, as a function of its cardinality. Specifically, given two n sets of points $R = \{r_1, ..., r_n\}$ and $B = \{b_1, ..., b_n\}$ distributed uniformly and randomly on the m leaves of λ-Hierarchically Separated Trees with branching factor b such that each one of its leaves are of depth δ, we will prove that the expected weight of optimal matching between R and B is $\Theta(\sqrt{nb}\sum_{k=1}^{h}(\sqrt{b}\lambda)^k)$, for $h = \min(\delta, \log_b n)$. Using a simple embedding algorithm from $\mathbb{R}^d$ to HSTs, we are able to reproduce the results concerning the expected optimal transportation cost in $[0,1]^d$, except for $d = 2$. We also show that giving random weights to the points does not affect the expected matching weight by more than a constant factor. Finally, we prove upper bounds on several sets for which showing reasonable matching results would previously have been intractable, e.g., the Cantor set, and various fractals.

Key words: Random Matching, Hierarchically Separated Trees, Supremum Bounds

1 Introduction

The problem of computing a large similar common subset of two point sets arises in many areas of computer science, ranging from computer vision and pattern recognition, to bio-informatics [2, 12, 4]. Most of recent related work concerns the design of efficient algorithms to compute rigid transformations for establishing correspondences between two point sets in $\mathbb{R}^d$ subject to minimization of a

⋆ Part of this research was done while the author worked at the Analysis and Stochastics Research Group at the University of Szeged, Hungary. Partially supported by OTKA T049398.

⋆⋆ Partially supported by NSF grant IIS-0456001 and by ONR grant ONR-N000140410363.

distance measure. In comparison, less attention has been devoted to extremal matching problems related to random point sets such as: *Presented with two random point sets, how do we expect matching weight to vary with data set size?*

Perhaps the most seminal work in extremal random matching is the 1984 paper of Ajtai, Komlós and Tusnády [1] presenting a very deep and important result which has found plenty of applications since then. They considered two sets of points X_n and Y_n chosen uniformly at random in $[0,1]^2$, with $|X_n| = |Y_n| = n$, and determined (asymptotic) bounds on the sequence $\{\mathbf{E}M\}$, where M is the optimal matching weight, or *transportation cost* between X_n and Y_n: $M = \min_\sigma \sum_i ||X_i - Y_{\sigma(i)}||_2$ where σ runs through all the possible permutations on $[n]$. Leighton and Shor [8] addressed the problem of 2 dimensional grid matching where shortly after Ajtai et al, they analyzed the maximum cost of any edge in the matching instead of the sum. Shor and Yukick [14] extended this minimax grid matching result to dimensions greater than two. Shor [13] applied the AKT result to obtain bounds on the average case analysis of several algorithms. Talagrand [15] introduced the notion of majorizing measures and as an illustration of this powerful technique derived the theorem of Ajtai et al. Rhee and Talagrand [9] have explored upward matching (in $[0,1]^2$): the case where points from X must be matched to points of Y that have greater x- and y-coordinates. They have also explored a similar problem in the cube [10]. In [16] Talagrand gave insight to exact behavior of expected matching weight for dimensions $d \geq 3$ for arbitrary norms.

In this paper we will introduce the random matching problem on *hierarchically separated trees*. The notion of hierarchically (well-)separated tree (HST) was introduced by Bartal [3]. An λ-HST is a rooted weighted tree with two additional properties: (1) edge weights between nodes and their children are the same for any given parent node, and (2) the ratio of incident edge weights along a root-leaf path is λ (so edges get lighter by a factor of λ as one approaches leaves). We primarily consider balanced trees, that is, the branching factor of all nodes other than the leaves is an integer b, and also require that every leaf is of depth δ. Using the notion of balanced λ-HST, we can state the first contribution of this manuscript on the expected transportation cost of optimal matching $\mathbf{E}M_T(R,B)$:

Theorem 1. *Let $T = T(b,\delta,\lambda)$ be a balanced HST, and R and B two randomly chosen n-element submultisets of the set of leaves of T and define $h = \min(\delta, \log_b n)$. Then there exist positive constants K_1 and K_2 such that*

$$K_1\sqrt{bn}\sum_{k=1}^{h}(\sqrt{b}\lambda)^k \leq \mathbf{E}M_T(R,B) \leq K_2\sqrt{bn}\sum_{k=1}^{h}(\sqrt{b}\lambda)^k.$$

Theorem 1 will also allow us to approach and duplicate the upper-bound results of optimal matching for point sets distributed in $[0,1]^d$ found in the literature easily (see [7]), with a slightly loose result in the single (and most interesting) case of $d = 2$. Since we use crude approximations of $[0,1]^d$ by HSTs, we cannot expect much more.

On the other hand this method is general enough to attack the randomized matching problem in general for finite metric spaces. It can always give upper bounds (by using Theorem 2 or Corollary 1). If the metric space is sufficiently symmetric (e.g., fractals), one can get reasonable lower bounds by applying the theorem of Fakcharoenphol et al. ([5]) on approximating a finite metric space by HSTs. We further extend the upper bound of the transportation cost to the case of weighted point sets. This model is commonly used in texture mapping in computer vision, see [11].

The final application of the newly developed machinery will include extending upper-bound matching results to finite approximations of certain fractals. We generalize Theorem 1 for non-uniformly distributed point sets and for subtrees of balanced trees as well.

2 The Upper- and Lower-Bounds for Matching on HSTs

In this section, our modus operandi will be to prove upper- and lower-bounds for the weight of the matching problems on HSTs. The trees considered in the paper are a somewhat restricted variation of HSTs defined as follows:

Definition 1. *Let b, δ be positive integers and $0 < \lambda < 1$ be a real number. We call a rooted tree T a balanced (b, δ, λ)-HST, if every edge parented to the root has unit weight, every edge not parented to the root has weight λ times the weight of the edge immediately closer to the root, every non-leaf node has the same number of children (which we will call the branching factor b), and every leaf has the same depth δ.*

We remark that having the same depth δ for every leaf of T can be assumed without loss of generality, and as we will see, in several cases the branching factor is naturally bounded.

Given a balanced HST T, let $R = \{r_1, ..., r_n\}$ and $B = \{b_1, ..., b_n\}$ respectively denote the multisets of n red and n blue points chosen among the leaves of T. We define a *matching* between R and B as a one-to-one mapping σ between them. The weight of optimal matching (optimal transportation cost) with respect to T will be defined as $M_T(R, B) = \min_\sigma \left(\sum_{1 \le i \le n} d_T(r_i, b_{\sigma(i)}) \right)$, where $d_T(r, b)$ is the length of the path between leaves containing points r and b in T. Note that $M_T(R, B)$ is the Earth Mover Distance of R and B on the metric defined by T. For a pair of matched points (r, b) under a mapping σ, belonging to distinct leaves u_r and u_b in T, we will say the matched pair (r, b) results in a *transit* at vertex v, if v is an ancestor of both u_r and u_b and the path between u_r and u_b passes through v. We will also use τ_v to denote the total number of transits at vertex v in an optimal matching between R and B. Any red-blue pair that is mapped under a matching σ at a leaf of T contributes no weight to the transportation cost. For a vertex v let $\delta(v)$ denote its level in the tree, that is, the number of edges on the path from r to v. Observe that the weight of the

optimal matching can be restated as follows:

$$M_T(R,B) = \sum_{k=0}^{\delta-1} \sum_{v:\delta(v)=k} \tau_v S(k, \delta-1), \tag{1}$$

where $S(i,j) = 2(\lambda^i + \lambda^{i+1} + \ldots + \lambda^j) \leq C_\lambda \lambda^i$ for $0 \leq i \leq j \leq \delta - 1$, and $C_\lambda = \sum_{j \geq 0} \lambda^j$.
Our goal is to estimate tight bounds on the expected optimal transportation cost $\mathbf{E}M_T(R,B)$ for randomly chosen R and B. Throughout the paper we will denote the standard deviation of a random variable X by $\mathbf{D}X$. The following pair of observations will be useful in the proof of Theorem 1:

Observation 1 *Given a balanced (b,δ,λ)-HST tree T, and multiset R of n red points and multiset B of m blue points distributed among the leaves of T, we have $M_T(R,B) \leq \min(n,m)S(1,\delta)$.*

Lemma 1. *Let X be the sum of a finite number of independent bounded random variables. Then $\mathbf{E}\,|X - \mathbf{E}X| = \Theta(\mathbf{D}X)$.*

We omit the details, but comment that to show the upper bound of Lemma 1, one can repeatedly use Chebyshev's inequality, while the lower bound is the consequence of Hölder's inequality.

The process of randomly and uniformly choosing the leaves of a balanced HST T with branching factor b to host the points in R and B can be stated as follows: starting from the root, choose a child of the current vertex uniformly at random among its b children; if the new vertex is not a leaf, repeat this random selection process. Otherwise, this leaf is our random choice. We will distribute the "random" sets R and B among the leaves of T by repeating this procedure independently for every point of $R \cup B$. It is obvious that this procedure results in two random submultisets of the set of leaves of T. For an arbitrary vertex $v \in T$ let R_v and B_v, respectively, denote the cardinality of the set of red (respectively, blue) points that when distributed reach their host leaves in T on a path from the root through v. In particular, R_l is the number of red points assigned to the leaf l, and B_l is the number of blue points assigned to l.

Next, we will estimate the number of transits, τ_r, at root r of a star (HST-) tree T with b leaves $L = \{u_1, u_2, \ldots, u_b\}$ when n red and n blue points are distributed randomly among the elements of L. Let $X_u = R_u - B_u$ for the leaf u, then $\sum_{u \in L} X_u = 0$ and

$$\tau_r = \sum_{u \in L} \max\{X_u, 0\} = -\sum_{u \in L} \min\{X_u, 0\}.$$

It follows that $\sum_{u \in L} |X_u| = 2 \sum_{u \in L} \max\{X_u, 0\}$, and hence

$$\mathbf{E}\tau_r = \frac{1}{2} \sum_{u \in L} \mathbf{E}|X_u|.$$

Observe that X_u is the combination of $2n$ independent indicator random variables

$$X_u = \sum_{j=1}^{n} R_u(j) - \sum_{j=1}^{n} B_u(j),$$

where $R_u(j) = 1$ if and only if the j^{th} red points reaches leaf u; similarly, we define $B_u(j)$. Hence, $\mathbf{E}X_u = 0$, and Lemma 1 can be applied. Setting $\beta = (1/b - 1/b^2)$, it is an easy exercise to verify that $\mathbf{D}R_u(j) = \mathbf{D}B_u(j) = \sqrt{\beta}$ for every $1 \leq j \leq n$, and hence $\mathbf{D}X = \sqrt{2n\beta}$. In summary, we have

Lemma 2. *There exist positive constants c_1 and c_2 such that $c_1 b\sqrt{n\beta} \leq \mathbf{E}\tau_r \leq c_2 b\sqrt{n\beta}$ for a star T with root r on b leaves, when n red and n blue points are distributed randomly and uniformly among its leaves.*

We note that Lemma 2 proves Theorem 1 when $\delta = 1$. We need a generalization of the above, when $\sum_{u\in L}(R_u - B_u) = \Delta \neq 0$. In this case there will be $|\Delta|$ points which will remain unmatched in the star tree. The number of transits at r is easily seen to be

$$\tau_r = \frac{1}{2}\left(\sum_{u\in L} |X_u| - |\Delta|\right),$$

thereby we get

Lemma 3. *The expected number of transits at root r of a star with leaf set L is*

$$\mathbf{E}\tau_r = \frac{1}{2}\mathbf{E}\left(\sum_{u\in L} |X_u| - |\sum_{u\in L}(R_u - B_u)|\right).$$

Next, we present the proof of Theorem 1 for two randomly chosen n-element submultisets R and B among the leaves of a balanced (b, δ, λ)-HST T. The following simple combinatorial lemma is crucial for the proof.

Lemma 4. *Let R, B and T be as above, and let $k \geq 1$. Then $\mathcal{T}_{k-1}$, the total number of transits at level $k-1$ is*

$$\mathcal{T}_{k-1} = \sum_{\delta(v)=k-1} \tau_v = \frac{1}{2}\sum_{\delta(u)=k} |X_u| - \frac{1}{2}\sum_{\delta(u')=k-1} |X_{u'}|.$$

Proof: The lemma follows easily by induction on the depth of the tree, we omit the details. □

Now we are ready to prove our main result

Proof of Theorem 1: Since $M_T(R, B)$ is a finite sum (see Equation 1), we can restate it as the sum of the expectation at each level of tree T, i.e.,

$$\mathbf{E}M_T(R, B) = \sum_{k=0}^{\delta-1} \sum_{v:\delta(v)=k} \mathbf{E}\tau_v \Theta(C_\lambda \lambda^k).$$

Applying Lemma 4 we get

$$\mathbf{E}M_T(R,B) = \sum_{k=0}^{\delta-1} \Theta(C_\lambda \lambda^k) \mathbf{E}\left(\sum_{\delta(u)=k} |X_u| - \sum_{\delta(u')=k-1} |X_{u'}| \right).$$

Therefore, it suffices to compute $\mathbf{E}|X_u|$ for every $u \in T$. Notice that we are in a situation very similar to that of the star tree. At level k we have b^k vertices, hence the expected number of transits at level k is $b^k \mathbf{E}|X_u|$, where u is an arbitrary vertex at level k. Let $\beta_k = \frac{1}{b^k}(1 - \frac{1}{b^k})$. Applying Lemma 1 we get that the expected number of transits at level k is of order

$$\mathcal{T}_k = b^{k-1}\sqrt{n}(b\sqrt{\beta_k} - \sqrt{\beta_{k-1}}).$$

Simple calculation shows that

$$b^{k/2}\frac{\sqrt{nb}}{2} \leq \mathcal{T}_k \leq 2b^{k/2}\sqrt{nb}.$$

This will allow us to conclude that

$$K_1\sqrt{bn}\sum_{k=0}^{\delta-1}(\lambda\sqrt{b})^k \leq \mathbf{E}M_T(R,B) \leq K_2\sqrt{bn}\sum_{k=0}^{\delta-1}(\lambda\sqrt{b})^k. \tag{2}$$

If $\delta \leq \log_b n$ then the above proves Theorem 1. So, assume that $\delta > \log_b n$. In this case there is at least one vertex w such that $\delta(w) = \log_b n$. Then using Observation 1 the expected transportation cost of the matching for the subtree T_w rooted at w can be bounded as

$$M_{T_w}(R_w, B_w) \leq \min(|R_w|, |B_w|)S(\delta(w), \delta).$$

Therefore, we get the following upper-bound for the expected matching length in T_w:

$$\begin{aligned}
\mathbf{E}M_{T_w}(R_w, B_w) &\leq C_\lambda \lambda^{\log_b n} \sum_{k=0}^{n} \Pr(R_w = k) \times k \\
&= C_\lambda \lambda^{\log_b n} \sum_{k=0}^{n} \frac{k}{k!} \\
&\leq eC_\lambda \lambda^{\log_b n}.
\end{aligned}$$

Here we used the fact that if $\delta(w) = \log_b n$ then R_w has Poisson distribution. Observing that there are b^k vertices at level k, we have

$$\mathbf{E}M_T(R,B) \leq K_2\sqrt{bn} \sum_{k=0}^{\log_b n - 1} (\lambda\sqrt{b})^k + eC_\lambda n \lambda^{\log_b n}.$$

Now we are in the position to estimate the precise upper-bound on $\mathbf{E}M_T(R,B)$. We will consider three distinct cases, depending on the value of $\sqrt{b}\lambda$:
Case I: If $\sqrt{b}\lambda < 1$ then $\lambda^{\log_b n} < n^{-1/2}$, and hence, $neC_\lambda\lambda^{\log_b n} < eC_\lambda\sqrt{n}$. Since in this case $\sum_{k=0}(\lambda\sqrt{b})^k$ is a constant, we get the desired upper-bound.
Case II: If $\sqrt{b}\lambda = 1$ we will have

$$\sum_{k=0}^{\log_b n-1}(\lambda\sqrt{b})^k = \log_b n,$$

and $\lambda^{\log_b n} = n^{-1/2}$, which again gives us the upper-bound of the theorem.
Case III: If $\sqrt{b}\lambda > 1$ then

$$\sqrt{n}\sum_{k=0}^{\log_b n-1}(\lambda\sqrt{b})^k = \sqrt{n}\frac{(\lambda\sqrt{b})^{\log_b n}-1}{\lambda\sqrt{b}-1} = O(n\lambda^{\log_b n}),$$

which implies the desired upper-bound.
The lower bound of Theorem 1 follows trivially from the fact that if truncating the lower-bound of the sum in (2) at the $\log_b n$-th term (which has only non-negative elements) will result in the desired lower-bound. □

An important generalization emerges when the points of R and B are not necessarily uniformly distributed among the leaves of T. Given any non-leaf vertex of T we can distribute the red and blue points among its children according to an arbitrary probability distribution. Conversely, it is easy to see that given any probability distribution on the leaves one can find appropriate probabilities for every non-leaf vertex of T in order to arrive at the desired distribution of the red and blue points at the leaf-level. This gives rise to the following theorem:

Theorem 2. *Let $T = T(b,\delta,\lambda)$ be a balanced HST, and $\mathbf{P}$ a probability distribution on the leaves of T. Let R and B be two n-element submultisets of the set of leaves of T chosen randomly, independently from $\mathbf{P}$. Then there exist a positive constant K_3 (depending only on λ) such that*

$$\mathbf{E}M_T(R,B) \le K_3\sqrt{bn}\sum_{k=0}^{\delta-1}(\sqrt{b}\lambda)^k.$$

Sketch of the proof: The proof follows the same line of argument as Theorem 1, except that in addition we use the following elementary inequality: if $a_1, a_2, \ldots, a_t \in [0,1]$, $\sum a_i \le 1$ then

$$\sum_{1\le i\le t}\sqrt{a_i(1-a_i)} \le t\sqrt{\sum a_i/t(1-\sum a_i/t)}.$$

Applying the above inequality we can perform the following balancing algorithm: first, we make the probability of choosing an arbitrary child of the root equal to the reciprocal of the number of its children, that is, we choose uniformly

among the children of the root. Then we repeat the above for all the subtrees originating from these children. Proceeding top-down, at the end we achieve that every leaf of the tree has the same chance to be chosen, moreover, we never decreased the expected number of transitions at any intermediate vertex. This implies the theorem. □

We also note the following consequence of Theorem 2, which follows by choosing certain edge probabilities to be 0.

Corollary 1. *If T' is an arbitrary subtree of a balanced (b, δ, λ)-HST T, then the expected optimal transportation cost on T' is upper bounded by the expected optimal transportation cost on T.*

Observe that one cannot expect any reasonable general lower bound in case of a non-uniform distribution or for subtrees: if the subtree T' is a path, the transportation cost is zero.

3 The Case of Matching in $[0, 1]^d$

As a first application of the theory we developed in Section 2 we reproduce the results of [7] concerning the expected optimal transportation cost in the d-dimensional unit cube. We remark that for finding nearest neighbors in the Euclidean space Indyk and Thaper [6] used similar ideas for approximating the d-dimensional unit cube with HSTs.

We begin by presenting the general idea for approximating $[0, 1]^d$ by a balanced HST. The number of iterations of this proces will be equivalent to the depth δ of the tree. In the j^{th} step we construct a grid G_j with 2^{jd} cells, the edge length of a cell is 2^{-j}. G_j is a refinement of G_{j-1} for every j: we obtain the cells of G_j by dividing each cell of G_{j-1} into 2^d subcells of equal volume. We stop when $j = \delta$. The tree is going to have 2^{jd} vertices at level j, each vertex corresponds to a cell of G_j. A vertex v at level j will be adjacent to a vertex w at level $j + 1$ if and only if the cell of v in G_j contains the cell of w in G_{j+1}. The weight of edge (v, w) will be 2^{1-j}. Clearly, the construction will result in a balanced $(2^d, \delta, 1/2)$-HSTs. Moreover, the resulting HST will dominate the distances of the lattice points of G_δ: the Euclidean distance of any two lattice points is at most as large as their distance in the HST. Finally, we will approximate a set of points in $[0, 1]^d$ by *discretizing* the point set: for every point we assign it to the closest available lattice point.

We will first consider the case of the unit interval, i.e., $d = 1$:

Proposition 1. *Given n red points and n blue points distributed uniformly at random on $[0, 1]$, the expected weight of an optimal matching is $O(\sqrt{n})$.*

Proof: We approximate the $[0, 1]$ interval with an equidistant set of $O(n^2)$ lattice points, as is described above. We will approximate this metric space by a balanced $(2, 2 \log n, 1/2)$-HST T, whose leaves are the lattice points. The discretization overhead associated with approximating the red and blue points with

the leaves is no more than the cost of moving each point to the nearest leaf, i.e., $2n \cdot 1/(2n) = 1$. Applying Theorem 1 with parameters $b = 2$ and $\lambda = 1/2$, will result in the desired bound. □

In the plane our HST technique offers loose results. Ajtai et al. [1] showed that the expected weight of the optimal matching in $[0,1]^2$ is $\Theta(\sqrt{n \log n})$. In Proposition 2, we use Theorem 1 to obtain the bound of $O(\sqrt{n} \log n)$.

Proposition 2. *Let $B = \{b_i\}_{i=1}^n$ and $R = \{r_i\}_{i=1}^n$ be sets of blue and red points distributed uniformly at random in $[0,1]^2$ and let M_n be the expected weight of an optimal matching of B against R. Then $\mathbf{E}M_n = O(\sqrt{n} \log n)$.*

Proof: As discussed in the general process, we construct the 2-dimensional grid, then the balanced $(4, 1/2, 2\log_4 n)$-HST T. The discretization overhead associated with approximating the red and blue points with the leaves of T is again negligible for $\delta \geq 2\log_4 n$. Applying Theorem 1 with parameters $b = 4$ and $\lambda = 1/2$, we get the upper bound of $O(\sqrt{n} \log n)$. □

We now jump to real dimension 3 and above, showing (Proposition 3) that expected weight of optimal matching is $O(n^{(d-1)/d})$,

Proposition 3. *Let $B = \{b_i\}_{i=1}^n$ and $R = \{r_i\}_{i=1}^n$ be sets of blue and red points distributed uniformly at random in $[0,1]^d$, $d \geq 3$, and let M_n be the expected weight of an optimal matching of B against R. Then $\mathbf{E}M_n = O(n^{(d-1)/d})$.*

Proof: As before, we construct a sufficiently dense grid. Its lattice points will be used to aproximate the real vectors of $[0,1]^d$. The finite metric space of the lattice points will be dominated by a balanced HST $T = T(2^d, 3\log_{2^d} n, 1/2)$. We are in the position to apply Theorem 1 with parameters $b = 2^d$ and $\lambda = 1/2$, and get the bound of $O(n^{(d-1)/d})$. □

Observe, that for $d \neq 2$ our seemingly crude approximations by HSTs result in tight bounds up to a constant factor (see e.g., [1]).

4 Optimal Matching for Weighted Point Sets

In this section we will estimate the expected weight of the optimal weighted matching for point sets $R = \{r_1, ..., r_n\}$ and $B = \{b_1, ..., b_n\}$ distributed uniformly and at random among the leaves of an HST T. We assume that every leaf u of T is associated with a randomly and independently chosen mass $m(u) \in [0,1]$. Then the total transportation cost is defined to be

$$M_{T,m}(R,B) = \min_\sigma \left(\sum_{1 \leq i \leq n} d_T(r_i, b_{\sigma(i)}) \min\{m(r_i), m(b_{\sigma(i)})\} \right)$$

We will use the following folklore result: if x and y are chosen randomly, independently from $[0,1]$, then their expected distance is $\mathbf{E}|x - y| = 1/3$.

Then we have the following

Theorem 3. *Let $T = T(b, \delta, \lambda)$ be a balanced HST with set of leaves L, and R and B two randomly chosen n-element submultisets of L. Let $m : L \to [0, 1]$ be a function, its values are drawn randomly and independently, and define $h = \min(\delta, \log_b n)$. Then there exist positive constants K_4 and K_5 such that*

$$K_4\sqrt{bn}\sum_{k=1}^{h}(\sqrt{b}\lambda)^k \leq \mathbf{E}M_{T,m}(R, B) \leq K_5\sqrt{bn}\sum_{k=1}^{h}(\sqrt{b}\lambda)^k.$$

Sketch of the proof: The proof follows the same line of arguments of Theorem 1, except that when computing the expected transportation cost, one has to multiply the number of transitions not only by the edge weight of T but the expected mass which is to be moved. Since this latter number is $1/3$ on the average and was chosen independently from the distribution of the points, we conclude the theorem. □

5 The Case of Finite Approximation of Fractals

The machinery developed in Section 2 is general enough to consider matching on a finite approximation of a self-similar set. The notion of a finite approximation of fractals is best explained through an example. Recall that, the Cantor set is formed by repeatedly removing the open middle third of each line segment in a set of line segments, starting with $[0, 1]$. If we stop this process after α iterations, we will refer to the resulting set as an α-approximation of the Cantor set.

Next, consider sets $R = \{r_1, ..., r_n\}$ and $B = \{b_1, ..., b_n\}$ of red and blue points respectively, distributed uniformly at random on the δ-approximation of the Cantor set, with $\delta \geq 2\log n$. We are interested in the expected weight of an optimal matching between R and B. We can think of the δ-approximation of the Cantor set as being embedded into a balanced $(2, \delta, 1/3)$-HST T over the unit interval. We have $b = 2$ since at every step we double the number of subintervals, and $\lambda = 1/3$, because the length of these subintervals shrink by a factor of $1/3$. The discretization overhead associated with approximating the red and blue points with the leaves is no more than the cost of moving each point to the nearest leaf: $\leq 2n \cdot 1/(2n) = 1$. We can apply Theorem 1 with parameters $b = 2$ and $\lambda = 1/3$, and conclude that $\mathbf{E}M_T(R, B) = O(\sqrt{n})$.

The tree metric of $T(2, \delta, 1/3)$ dominates the Euclidean metric on the Cantor set. Therefore, the expected optimal matching weight of n blue and n red points distributed randomly on the Cantor set is no heavier than the same points distributed on $[0, 1]$ itself. We have proved the following

Theorem 4. *Let $R = \{r_1, ..., r_n\}$ and $B = \{b_1, ..., b_n\}$ be sets of red and blue points distributed uniformly and at random in the δ-approximation of the Cantor set process with $\delta \geq 2\log n$. Then the expected weight of an optimal matching between R and B is $O(\sqrt{n})$.*

Next, we consider the $\log_3 n$-approximation of a Sierpinski triangle. Here a balanced HST with branching factor $b = 3$, $\lambda = 1/2$, and depth δ ($\delta \geq$

$2\log_3 n$) dominates the Euclidean metric, and provides a good approximation after discretization. A similar argument to Cantor set will allow us to prove the following

Theorem 5. *Let $R = \{r_1, ..., r_n\}$ and $B = \{b_1, ..., b_n\}$ be sets of red and blue points distributed uniformly and at random in the interior of the δ-approximation of a Sierpinski triangle for δ large enough. Then the expected weight of an optimal matching between R and B is $O(\sqrt{n})$.*

Note the lack of the $\log n$ factor in the upper-bound. The expected optimal matching weight in a triangle would be $O(\sqrt{n \log n})$ by the result of Ajtai et al.

As a final example for the application of Theorem 1 on fractals, we will consider the *Menger sponge*. A Menger sponge results from recursively dividing the unit cube into $3^3 = 27$ sub-cubes, removing the middle cube on each face and the cube in the center, then recursing on each sub-cube. To find an upper bound on the expected weight of matchings on the Menger sponge, consider a balanced HST T with $\lambda = 1/3$ (the diameter decreases by a factor of $1/3$ at every recursion step), branching factor $b = 20$, and depth $\delta \geq 3\log_{20} n$ (this depth is sufficiently large to provide good approximation in the discretization). T is dominating, therefore, an upper bound on the expected weight of the optimal matching is stated in the following

Theorem 6. *Let $R = \{r_1, ..., r_n\}$ and $B = \{b_1, ..., b_n\}$ be sets of red and blue points distributed uniformly and at random in the interior of the $3\log_{20} n$- approximation of a Menger sponge. Then the expected weight of an optimal matching between R and B is $O(n^{1-\log_{20} 3})$.*

6 Conclusions

In this paper we presented a tight bound on the expected weight of transportation cost for matching of points on balanced HSTs. We extended our upper-bounds for subtrees of balanced HSTs, and for non-uniform distributions. Using low-distortion embedding of $\mathbb{R}^d$ to HSTs, we reproduce the results concerning the expected optimal transportation cost in the $[0, 1]^d$, except for the case of $d = 2$ for which we have a discrepancy of a factor of $\sqrt{\log n}$. We also proved upper-bounds on several sets for which showing reasonable matching results would previously have been intractable. By existing approximation theorems for finite metric spaces we could give bounds on the expected transportation cost in any finite metric space. We plan to consider the analogues of other related matching problems, for example up-right matchings, etc.

References

1. M. Ajtai, J. Komlós, and G. Tusnády. On optimal matchings. *Combinatorica*, 4(4):259–264, 1984.

2. H. Alt and L. Guibas. Discrete geometric shapes: Matching, interpolation, and approximation. In J. Sack and J. Urrutia, editors, *Handbook of Comput. Geom.*, pages 121–153. Elsevier Science Publishers B.V. North-Holland, Amsterdam, 1996.
3. Y. Bartal. Probabilistic approximation of metric spaces and its algorithmic applications. In *FOCS 1996: 37th Annual Symposium on Foundations of Computer Science*, pages 184–193, 1996.
4. S. Cabello, P. Giannopoulos, C. Knauer, and G. Rote. Matching point sets with respect to the earth mover's distance. *Comput. Geom. Theory Appl.*, 39(2):118–133, 2008.
5. J. Fakcharoenphol, S. Rao, and K. Talwar. A tight bound on approximating arbitrary metrics by tree metrics. In *STOC 2003: Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 448–455, 2003.
6. P. Indyk and N. Thaper. Fast image retrieval via embeddings. In *The 3rd International Workshop on Statistical and Computational Theories of Vision*, 2003.
7. R. M. Karp, M. Luby, and A. Marchetti-Spaccamela. A probabilistic analysis of multidimensional bin packing problems. In *STOC '84: Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 289–298, 1984.
8. T. Leighton and P. W. Shor. Tight bounds for minimax grid matching, with applications to the average case analysis of algorithms. In *STOC 1986: Proceedings of the Eighteenth Annual ACM Symposium on Theory of Computing*, pages 91–103, 1986.
9. W. T. Rhee and M. Talagrand. Exact bounds for the stochastic upward matching problem. *Transactions of the American Mathematical Society*, 307(1):109–125, 1988.
10. W. T. Rhee and M. Talagrand. Matching random subsets of the cube with a tight control on one coordinate. *The Annals of Applied Probability*, 2(3):695–713, 1992.
11. Y. Rubner and C. Tomasi. Texture based image retrieval without segmentation. *IEEE International Conference on Computer Vision*, pages 1018–1024, 1999.
12. Y. Rubner, C. Tomasi, and L. Guibas. The earth movers distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–122, 2000.
13. P. W. Shor. The average-case analysis of some on-line algorithms for bin packing. In *FOCS 1984: 25th Annual Symposium on Foundations of Computer Science*, pages 193–200, 1986.
14. P. W. Shor and J. E. Yukich. Minimax grid matching and empirical measures. 19(3):1338–1348, 1991.
15. M. Talagrand. Matching random samples in many dimensions. *Annals of Applied Probability*, 2(4):846–856, 1992.
16. M. Talagrand. Matching theorems and empirical discrepancy computations using majorizing measures. *J. ACM*, 7:455–537, 1994.