CS-67720 Metric Embedding Theory and Its Algorithmic Applications

Lecture 10

Lecturers: Yair Bartal, Nova Fandina *Fall 2018/19*
For comments contact fandina@gmail.com

We continue our search for high quality embeddings of a finite metric spaces into tree metrics. Thus far we considered embedding of an input space into a metric defined over a tree graph, with relaxed guarantees on distortion (scaling distortion, prioritized distortion, Ramsey-type embeddings, etc.). In this lecture we discuss another approach of embedding: embedding into a distribution over tree metrics.

10.1 Probabilistic Embeddings

The notion of probabilistic embedding was formally defined and studied by Bartal in [2]:

Definition 10.1. *Let (X, d_x) be any metric space, and let S be a collection of metric spaces. We say that X probabilistically embeds into S with distortion $\alpha \geq 1$ if:*

1. *For all $(Y, d_y) \in S$ there is a non-contractive embedding $f_Y : X \to Y$,*
2. *There exists a probability distribution D over embeddings $\{f_Y | Y \in S\}$ such that for every $u, v \in X$, $E_{f_Y \sim D}[d_Y(f_Y(u), f_Y(v))] \leq \alpha \cdot d_X(u, v)$.*

We remark that the first requirement is crucial for algorithmic applications (we will see that later on).

Claim 10.1. *A weighted cycle on n-nodes is probabilistically embeddable into the line metric, with distortion 2.*

For comparison, recall that any embedding of an n-point cycle into a tree metric must have (a worst case) distortion at least $\Omega(n)$.

Proof. The algorithm randomly chooses an edge in the graph and removes it to obtain a line metric. Particularly, the algorithm removes an edge $e \in C$ with probability $\frac{w(e)}{w}$, where w denotes the total weight of the edges of C_n. For every edge e let L_e denote the line metric obtained by removing the edge e from the cycle. Consider any nodes $u, v \in C_n$. If the algorithm chooses an edge e on the lightest path between u and v, then $d_{L_e}(f_e(u), f_e(v)) \leq w$. Otherwise, $d_{L_e}(f_e(u), f_e(v)) = d(u, v)$. Therefore,

$$E[d(f_e(u), f_e(v))] \leq \frac{d(u,v)}{w} \cdot w + \left(1 - \frac{d(u,v)}{w}\right) \cdot d(u, v) = 2d(u, v) - (d(u, v))^2 \frac{1}{w} < 2 \cdot d(u, v). \qquad \square$$

In the next few lectures we will prove that any metric space on n points can be embedded into a distribution of trees (ultrametrics) with expected distortion $O(\log n)$. In [2] Bartal gave the first construction of probabilistic embedding into a distribution of ultrametrics, with distortion $O(\log n \log \Phi)$, where Φ is the aspect ratio of the input space. Later, in [3] he improved the result to $O(\log n \log \log n)$ distortion. Finally, in [6] the authors provided construction with the optimal expected distortion $O(\log n)$. (The lower bound of $\Omega(\log n)$ was already shown in [2]). In [4] Bartal gave a construction with optimal $O(\log n)$ distortion, using different ideas.

10.2 Probabilistic Embeddings into Trees

Theorem 10.2. *Every metric space on n points can be probabilistically embedded into ultrametric with expected distortion of $O(\log n)$.*

Remarks. Since tree metrics are essentially ℓ_1 metrics, and ℓ_1 metric is additive, the above result implies an embedding of any finite metric space into ℓ_1 with worst case distortion $O(\log n)$.

In addition, the above probabilistic embedding can be derandomized [5]: there is an efficient algorithm that generates a distribution over $n \log n$ trees, such that the expected distortion according to this distribution is $O(\log n)$. There is also a line of work that investigates the possibility of probabilistic embedding of any finite graph metric into its *spanning* tree. The state of the art so far [1] is expected distortion $O(\log n \log \log n)$, while the lower bound is known to be $\Omega(\log n)$.

10.2.1 Definitions

First we give all the necessary definitions.

Definition 10.2 (Δ bounded partition of X)**.** *Let X be a finite metric space. A collection P of subsets $S_1, S_2, \ldots S_t \subseteq X$, such that $\forall 1 \leq i \neq j \leq t,\ \ S_i \cap S_j = \emptyset$ and $\cup S_i = X$ is called a partition of X. Each S_i is called a cluster of the partition.*

The partition P is Δ-bounded if $\forall i, diam(S_i) \leq \Delta$. For $x \in X$ let $P(x)$ denote the cluster of the partition that contains the element x.

Definition 10.3 (Probabilistic Δ bounded partition)**.** *A Δ - bounded probabilistic partition $\mathcal{P}$ is a distribution over a set $\{P_i\}$ of Δ-bounded partitions.*

Padding parameter of a probabilistic partition.
A Δ- bounded probabilistic partition $\mathcal{P}$ has padding parameter $\gamma \geq 0$, if $\forall x \in X$, and $\forall r > 0$, $Pr_{\mathcal{P}}[B(x,r) \not\subseteq P(x)] \leq \gamma \cdot \frac{r}{\Delta}$. In more general case, $\gamma : X \to \mathbb{R}^+$.

Definition 10.4 (Bundle of probabilistic partitions)**.** *Let X be an n point metric space and $k > 1$. Let $\Delta_0 = diam(X)$, and $\Delta_i = \Delta_0 k^{-i}$. A bundle of probabilistic partitions $\mathcal{H}$ is a collection of probabilistic partitions $\{\mathcal{P}_i\}$ such that each $\mathcal{P}_i$ is a Δ_i-bounded probabilistic partition with padding parameter γ_i.*

Definition 10.5 (Special bundle)**.** *A bundle $\mathcal{H}$ is called a special bundle if every γ_i is a function, such that for all $x \in X$ it satisfies one of the following:*

1. $\gamma_i(x) > 0$, *and* $\forall r > 0$ *it holds that* $Pr_{\mathcal{P}_i}[B(x,r) \nsubseteq P(x)] \leq \gamma_i(x)\frac{r}{\Delta_i}$ *(standard requirement)*
2. $\gamma_i(x) = 0$, *and* $\forall 0 < r < \frac{\Delta_i}{16}$, $Pr_{\mathcal{P}_i}[B(x,r) \nsubseteq P(x)] = 0$ *(i.e., the promise is only for the small radii).*

Definition 10.6 (Padding parameter of a bundle)**.** *The padding parameter of a (special) bundle* $\mathcal{H}$ *is defined by* $\gamma(\mathcal{H}) = \max\{\max_{x\in X}\{\sum_i \gamma_i(x)\}, 1\}$

10.2.2 The proof of Theorem 10.2

We will prove the following two theorems from which Theorem 10.2 follows.

Theorem 10.3. *Let X be an n point metric space, and let $\mathcal{H}$ be a special bundle of probabilistic partitions of X, with any $k \geq 2$ and with padding parameter $\gamma(\mathcal{H})$. Then X admits a probabilistic embedding into k-HST trees, with expected distortion $O(k \cdot \gamma(\mathcal{H}))$.*

Theorem 10.4. *Let X be any n point metric space. Then there exists a special bundle of probabilistic partitions $\mathcal{H}$ with padding parameter $\gamma(\mathcal{H}) = O(\log n)$.*

We start with proving Theorem 10.3.

Proof of Theorem 10.3.
Assume w.l.o.g. that $k \geq 16$. We recursively construct the probabilistic embedding of X into k-HST as follows.

Let $Z \subseteq X$ be a current subspace (at the beginning $Z = X$), and let $i = \max\{i \geq 1 \mid \Delta_{i-1} \geq diam(Z)\}$. Therefore, $\Delta_i \leq diam(Z) \leq \Delta_{i-1} = k\Delta_i$.

Consider $\mathcal{P}_i$ - the Δ_i-bounded probabilistic partition of X from the given bundle of partitions. Pick a Δ_i-bounded partition P_i according to the distribution $\mathcal{P}_i$, and let $C_1, C_2, \ldots, C_t$ be the clusters of the chosen partition *induced* on Z, i.e., C_j is an intersection of Z with a j-th cluster in the partition P_i. For each C_j recursively construct a probabilistic embedding into k-HST. As a result we obtain k-HST trees $T_1, \ldots, T_t$. Denote by $w_1, \ldots, w_t$ the roots of these trees. Construct a k-HST tree for Z by defining a new root w, and letting $w_1, \ldots, w_t$ to be its children. Set the label of w to be $\Delta(w) = k\Delta_i$. From the recursive construction it holds that $\Delta(w_s) \leq \Delta_i$. Thus the result is indeed a k-HST.

Note that the embedding is non-contractive. If x, y are in different clusters of the partition (during some step of the recursion) then the distance between them can only grow since the label of the root is defined to be at least the diameter of the current subspace; and if x, y are in the same cluster then by induction's assumption the distance between them does not decrease.

We have to show that for all $x, y \in X$, $E\left[\frac{d_T(x,y)}{d(x,y)}\right] = O(k\gamma(\mathcal{H}))$. We will show by induction that at every step $i \geq 1$ of the recursion, $\forall x, y \in Z$, $E\left[\frac{d_{T_i}(x,y)}{d(x,y)}\right] = O\left(k \cdot \left(\sum_{j\geq i}\gamma_j(x) + 1\right)\right)$, where T_i is the random tree obtained on the step i of the recursion. We say that the recursive algorithm performs step i of the recursion when it considers $Z \subseteq X$, such that $\Delta_i \leq diam(Z) \leq k\Delta_i$.

If so, consider the i-th (when $i = 1$, $Z = X$) step of the recursion. Denote by $Z \subseteq X$ the set being considered during the step i, and assume by induction that for every cluster (induced by probabilistic partition on Z) the above claim holds. Let T_i denote the random tree the algorithm builds at the step i.

For any $x, y \in Z$ let $A_i(x, y)$ denote the event: x, y are separated *for the first time* to be in different clusters during the i-th step of the algorithm. Namely, the ball $B(x, d(x, y)) \subseteq Z$, and $B(x, d(x, y)) \not\subseteq (P_i(x) \cap Z)$ (the ball is padded at level's $i-1$ partition, but not padded in level's i partition intersected with Z). If $d(x, y) > \Delta_i/16$, then

$$\frac{d_{T_i}(x, y)}{d(x, y)} \le \frac{\Delta_{i-1}}{d(x, y)} = \frac{k\Delta_i}{d(x, y)} \le 16k,$$

with probability 1.

Otherwise, if $d(x, y) \le \Delta_i/16$, then

$$E_{T_i}[d_{T_i}(x, y)] = Pr_{T_i}[x, y \text{ in cluster } C_s] \cdot E[d_{T_s}(x, y) \mid x, y \text{ in cluster } C_s] + Pr_{T_i}[A_i(x, y)]k\Delta_i$$

$$\le E_{T_s}[d_{T_s}(x, y) \mid x, y \text{ in cluster } C_s] + Pr_{T_i}[A_i(x, y)]k\Delta_i \le^{\text{ind. ass.}}$$
$$\le O\left(k \cdot \left(\sum_{l \ge j \ge i+1} \gamma_j(x) + 1\right)\right) d(x, y) + Pr_{\mathcal{P}_i}[A_i(x, y)]k\Delta_i.$$

Since $d(x, y) \le \Delta_i/16$, by the definition of the padding parameter γ_i it holds that:

$$Pr[A_i(x, y)] = Pr_{T_i}[B(x, d(x, y)) \not\subseteq P(x)] \le \frac{\gamma_i(x) \cdot d(x, y)}{\Delta_i},$$

resulting in

$$E[d_{T_i}(x, y)] \le O\left(k\left(\sum_{l \ge j \ge i+1} \gamma_j(x) + 1\right)\right) d(x, y) + k\gamma_i(x)d(x, y) = O\left(k\left(\sum_{l \ge j \ge i} \gamma_j(x)) + 1\right)\right) d(x, y).$$

Now, for $i = 1$, and for any pair $x \neq y \in X$ it holds that $E[d_T(x, y)] = O(k\gamma(\mathcal{H}))d(x, y)$. $\square$

We continue by proving the following theorem:

Theorem 10.5. *Let X be an n point metric space, and $\Delta > 0$. There exists a Δ-bounded probabilistic partition of X with padding parameter $\gamma = O(\log n)$.*

Then, it follows that there exists a bundle $\mathcal{H}$ with $\gamma(\mathcal{H}) = O(\log n \log_k \Phi)$ padding parameter (since we can partition X until we get a trivial partition, i.e. each cluster contains one point of the space). Therefore, by Theorem 10.3, there is a probabilistic embedding of any finite metric space into k-HSTs, with distortion $O(k \log n \log_k \Phi)$.

After proving this result, we will focus on improving it to obtain the optimal $O(\log n)$ distortion.

Proof. Let $X_0 = X$. Pick an arbitrary $v_1 \in X_0$. Randomly choose a radius $r_1 \geq 0$, according to a distribution we will describe soon. Let $C_1 = B(v_1, r_1)$, and define the fist cluster by $\hat{C}_1 = C_1 \cap X$. Denote by $X_1 = X_0 \setminus \hat{C}_1$ and pick an arbitrary $v_2 \in X_1$. Randomly and independently choose a radius r_2, according to the same distribution. Let $C_2 = B(v_2, r_2)$, and define the second cluster by $\hat{C}_2 = C_2 \cap X_1$. Generally, continue this process to define$C_i = B(v_i, r_i)$, and the cluster $\hat{C}_i = C_i \cap X_{i-1}$, $X_i = X_{i-1} \setminus \hat{C}_i$.

The radii r_i are chosen from the following distribution: divide $[0, \infty)$ into the intervals of length $\frac{\Delta}{4 \log n}$, the k-th interval $I_k = [(k-1)\frac{\Delta}{4 \log n},\ k\frac{\Delta}{4 \log n}]$. Pick the interval I_k with probability $\frac{1}{2^k}$, and from the chosen interval, independently pick an r_i according to the uniform probability.

Note, that with high probability the above algorithm constructs a Δ-bounded partition: If some $r_i > \frac{\Delta}{2}$, then it was chosen from I_k, with $k > 2 \log n$. The probability to choose such I_k is bounded by $\sum_{j > 2 \log n} \frac{1}{2^j} \leq \frac{1}{n^2}$. Since there might be at most n clusters in a partition, it holds that $Pr[\exists\ \textit{an unbounded cluster}] = \frac{n}{n^2} = \frac{1}{n}$.

Note that we could slightly change the distribution of r_i: pick the interval $I_{k > 2 \log n}$ with probability 0, and the interval $I_{k = 2 \log n}$ with appropriate probability $(2/n^2)$. This process would create Δ-bounded partition with probability 1. We will use this observation later (in the refined partitions we will see in the next calss).

It remains to show that the padding parameter is $O(\log n)$. We will show that:

$$\forall x \in X,\ r > 0, Pr_D[B(x, r) \not\subseteq P(x)] \leq 24 \cdot \log n \frac{r}{\Delta}.$$

In other words, we wish to prove that our probabilistic partition satisfies the following:

$$\forall x \in X,\ \forall 0 < \delta \leq 1,\ \ for\ \ \eta^\delta := \frac{\delta}{24 \log n},\ \ Pr_D[B(x, \eta^\delta \Delta) \not\subseteq P(x)] \leq \delta.$$

If we manage to prove this, then given any $r > 0$ choose $\delta = 24 \frac{r}{\Delta} \log n$ and get the required. Note that it is enough to consider $\delta < 1$ as otherwise the inequity trivially holds.

Let $x \in X, 0 < \delta < 1$, let us denote $\hat{r} = \eta^\delta \Delta$. Next, we classify the balls C_i as follows. Starting from the first ball $C_1 = B(v_1, r_1)$, we say that C_1 **is a bad for** $B(x, \hat{r})$, if $B(x, \hat{r}) \cap B(v_1, r_1) \neq \emptyset$, and $B(x, \hat{r}) \cap B(v_1, r_1) \neq B(x, \hat{r})$.
This event occurs iff $r_1 \in [d(x, v_j) - \hat{r}, d(x, v_1) + \hat{r})$.
C_1 **is a good for** $B(x, \hat{r})$, if $B(x, \hat{r}) \subseteq B(v_1, r_1)$. This event occurs iff $r_1 \geq d(x, v_j) + \hat{r}$.
C_1 **is a neutral for** $B(x, \hat{r})$, if $B(x, \hat{r}) \cap B(v_1, r_1) = \emptyset$. This event occurs iff $r_1 < d(x, v_j) - \hat{r}$.

The next ball C_i is being classified according to the same rule, given that all the previous balls $C_1, \ldots, C_{i-1}$ have been classified as neutral.

The proof of the Theorem 10.5 will follow from the following lemma.

Lemma 10.6. *For all $x \in X$, $0 < \delta < 1$, for all $j \geq 1$ it holds that*

$Pr[C_j\ is\ bad\ for\ B(x, \hat{r}) \mid C_1, \ldots C_{j-1}\ neutral] \leq \delta \cdot Pr[C_j\ is\ good\ for\ B(x, \hat{r}) | C_1, \ldots C_{j-1}\ neutral]$.

Assume this lemma is correct, then:

$Pr[B(x, \hat{r}) \not\subseteq P(x)] = Pr[\exists\ a\ bad\ C_j] = Pr[C_1\ is\ bad] + Pr[C_2\ is\ bad \mid C_1\ is\ neutral] \cdot Pr[C_1\ is\ neutral] +$

$$+Pr[C_3 \textit{ is bad} \mid C_1, C_2 \textit{ are neutral}] \cdot Pr[C_1, C_2 \textit{ are neutral}] + \ldots +$$
$$+Pr[C_t \textit{ is bad} \mid C_1, C_2, \ldots C_{t-1} \textit{ are neutral}] \cdot Pr[C_1, C_2, \ldots C_{t-1} \textit{ are neutral}] \leq^{\textit{by lemma}}$$
$$\leq \delta(Pr[C_1 \textit{ is good}] + Pr[C_2 \textit{ is good} \mid C_1 \textit{ is neutral}] \cdot Pr[C_1 \textit{ is neutral}] +$$
$$+Pr[C_3 \textit{ is good} \mid C_1, C_2 \textit{ are neutral}] Pr[C_1, C_2 \textit{ are neutral}] + \ldots$$
$$+Pr[C_t \textit{ is good} \mid C_1, C_2, \ldots C_{t-1} \textit{ are neutral}] Pr[C_1, C_2, \ldots C_{t-1} \textit{ are neutral}]) \leq$$
$$\leq \delta \cdot Pr[\exists \textit{ a good } C_j] \leq \delta.$$

Proof of Lemma 10.6. First, we have

$$Pr[C_j \textit{ is bad} \mid C_1, \ldots, C_{j-1} \textit{ neutral}] \leq Pr[r_j \in (d(x, v_j) - \hat{r}, d(x, v_j) + \hat{r}) \mid C_1, \ldots, C_{j-1} \textit{ neutral}].$$

Recall that we consider $\delta < 1$, meaning $\hat{r} = \eta^\delta \Delta < \frac{\Delta}{24 \log n}$. Hence r_j can be in at most two intervals I_{l-1}, I_l, for some l, where we assume that $d(x, v_j) \in I_l$. Therefore,

$$Pr[r_j \in (d(x, v_j) - \hat{r}, d(x, v_j) + \hat{r}) \mid C_1, \ldots, C_{j-1} \textit{ neutral}] \leq \left(2^{-(l-1)} + 2^{-l}\right) \frac{2\eta^\delta \Delta}{\frac{\Delta}{4 \log n}} = \frac{\delta}{2^l}.$$

On the other hand,

$$Pr[C_j \textit{ is good} \mid C_1, \ldots, C_{j-1} \textit{ neutral}] = Pr[r_j \geq d(x, v_j) + \hat{r} \mid C_1, \ldots, C_{j-1} \textit{ neutral}] \geq \sum_{m > l} 2^{-m} = \frac{1}{2^l}.$$

□

This completes the proof of the theorem. □

References

[1] Ittai Abraham and Ofer Neiman. Using petal-decompositions to build a low stretch spanning tree. In *Proceedings of the 44th Symposium on Theory of Computing Conference, STOC 2012, New York, NY, USA, May 19 - 22, 2012*, pages 395–406, 2012.

[2] Yair Bartal. Probabilistic approximations of metric spaces and its algorithmic applications. In *37th Annual Symposium on Foundations of Computer Science, FOCS '96, Burlington, Vermont, USA, 14-16 October, 1996*, pages 184–193, 1996.

[3] Yair Bartal. On approximating arbitrary metrices by tree metrics. In *Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing, Dallas, Texas, USA, May 23-26, 1998*, pages 161–168, 1998.

[4] Yair Bartal. Graph decomposition lemmas and their role in metric embedding methods. In *Algorithms - ESA 2004, 12th Annual European Symposium, Bergen, Norway, September 14-17, 2004, Proceedings*, pages 89–97, 2004.

[5] Moses Charikar, Chandra Chekuri, Ashish Goel, Sudipto Guha, and Serge Plotkin. Approximating a finite metric by a small number of tree metrics. In *Proceedings of the 39th Annual Symposium on Foundations of Computer Science*, FOCS '98. IEEE Computer Society, 1998.

[6] Jittat Fakcharoenphol, Satish Rao, and Kunal Talwar. A tight bound on approximating arbitrary metrics by tree metrics. *Journal of Computer and System Sciences*, 69(3):485 – 497, 2004. Special Issue on {STOC} 2003.