On the Distortion of Embedding Perfect Binary Trees into Low-dimensional Euclidean Spaces

Dona-Maria Ivanova

under the direction of
Mr. Zhenkun Li
Department of Mathematics
Massachusetts Institute of Technology

Research Science Institute
August 2, 2016

Abstract

The paper considers the problem of embedding binary trees into $\mathbb{R}^d$ for a fixed positive integer d. This problem is part of a more general question concerning distortions of embedding of finite metric spaces into another metric spaces studied by Bourgain. Matousek's observation, that for embedding trees into an infinite-dimensional Euclidean space the upper bound of the distortion is achieved by binary trees motivated us to study the embeding of this particular class of trees into $\mathbb{R}^d$. We extend a result of Kumamoto and Miyano by showing that the distortion of an optimal embedding of binary tree with n vertices in $\mathbb{R}^d$ is $\Theta\left(\frac{n^{1/d}}{log_2 n}\right)$ and provide a construction achieving it.

Summary

Graph theory is the mathematical study of mathematical structures and their properties. The characteristic, which we are examining, distortion, can be thought of as follows: We draw a graph into a finite dimensional space. How much can we make the image correspond to the structure of the graph? In the specific case of graphs known as binary trees, we manage to bound the distortion from below using geometrical arguments, and from above – by providing an example. Our construction is optimal since the upper and lower bounds are equivalent up to a constant, not depending on the number of vertices in the tree.

1 Introduction

We study the distortion of embeddings of perfect binary trees into the Euclidean space $\mathbb{R}^d$ for a fixed positive integer d. The distortion is a characteristic of the graph, describing the minimal difference between the natural metric on the graph and the induced Euclidean metric on its image. Our question is a specific case of a more general problem concerning the distortion of embeddings of any finite metric space into an arbitrary metric space. Bourgain [1] showed that every finite metric space on n points embeds in some Euclidean space with $O(\log n)$ distortion. However, in order to achieve the minimal distortion, high dimensions are needed and little is known about the distortion of embeddings into lower dimensional spaces.

Matoušek [2] proved that any connected planar graph with n vertices can be embedded into $\mathbb{R}^2$ with $O(n)$ distortion. Later, Babilon, Matoušek, Maxova and Valtr [3] proved that any tree can be embedded into $\mathbb{R}^2$ with $O\left(\sqrt{n}\right)$ distortion. Our goal is to improve this bound for binary trees. Our motivation to study this particular class of trees is Matoušek's [4] observation that any tree can be embedded into a Euclidean space, with no restriction on the dimension, with distortion $O(\log\log n)$ and this upper bound is achieved for binary trees. Kumamoto and Miyano [5] showed that binary trees can be embedded into a line with distortion $O\left(\frac{n}{log_2 n}\right)$.

We construct another embedding, achieving the same distortion as in [5] in the 1-dimensional case, which we further use to generalize for higher dimensions. In particular, for any fixed d, we find an embedding into $\mathbb{R}^d$ with $O\left(\frac{n^{\frac{1}{d}}}{\log_2 n}\right)$ distortion. Moreover, we prove that the distortion of embedding binary trees into $\mathbb{R}^d$ is precisely $\Omega\left(\frac{n^{\frac{1}{d}}}{\log_2 n}\right)$, so our construction gives an optimal distortion up to a constant, independent of n.

In Section 2 we introduce some basic definitions and notations, used in our paper. Some general results and more specific definitions are given in Section 3. Here is introduced a

lemma, showing that any embedding of a binary tree with n vertices into $\mathbb{R}^d$ has distortion $\Omega\left(\frac{n^{\frac{1}{d}}}{\log_2 n}\right)$. Our goal is to find a particular embedding, achieving this lower bound. In Section 4 our initial construction of an embedding with distortion $O\left(n^{\frac{1}{d}}\right)$ is presented. We make use of this construction in Section 5 to develop a more elaborate one, which defines an embedding with distortion $O\left(\frac{n^{\frac{1}{d}}}{\log_2 n}\right)$. In Section 6 we summarize our results and discuss some avenues for future development of this project.

2 Background

In this section we introduce some basic definitions and notations.

2.1 Some basic graph theory definitions

For a given graph G we denote the set of its vertices by $V(G)$ and the set of its edges by $E(G)$. For each vertex $v \in V(G)$ its *valency* equals the number of edges emanating from v. A *tree* is any undirected connected graph on n vertices with $n-1$ edges. *Binary trees* are trees whose vertices have valency at most 3. The *leaves* of a tree are the vertices of valency 1. A *simple path* is a path in the graph which does not contain any $e \in E(G)$ more than once. It is a well-known result that for any two vertices v and u of a tree, there exists an unique simple path of edges connecting them. The number of edges forming this path is the *distance in the graph* between these two vertices, denoted by $d_G(v,u)$.

Definition 1. Let h be a positive integer. A *perfect binary tree* is a tree T_h for which

- there is exactly one vertex of valency 2, called *the root* of the tree;
- all leaves of the tree have the same distance h to the root;
- all other leaves have valency 3.

For a perfect binary tree the distance from a vertex v to the root is defined as the height of v and is denoted by $h(v) = d_G(v, v_0)$. Let h denote the height of the leaves. For every non-leaf vertex v there exist two vertices u_1 and u_2 such that $d_G(v, u_1) = d_G(v, u_2) = 1$ and $h(u_1) + 1 = h(u_2) + 1 = h(v)$. Define v as the parent of u_1 and u_2 and u_1 and u_2 as the left and the right child of v.

Remark. The labeling left and right is just for distinguishing between the two children and does not correspond to any difference in the properties of v_1 and v_2. However, when drawing the tree in the plane, it usually corresponds to their positioning.

An ancestor of v is a vertex u, such that the unique path between v and u does not contain any vertices of the same height.

2.2 Other definitions

Definition 2. An *embedding of a connected graph* is a map $f : G \to \mathbb{R}^d$ such that for each $v \in V(G)$ we have $f : v \to f(v)$ where $f(v) \in \mathbb{R}^d$.

Denote by $\|f(u) - f(v)\|$ the *standard Euclidean metric* on $\mathbb{R}^d$ for the images of v and u.

Definition 3. A *non-contracting* embedding f of a graph G is an embedding such that

$$\|f(u) - f(v)\| \geq d_G(u, v)$$

for any two vertices $v, u \in V(G)$

Remark. In other words, two vertices of the graph are mapped into points in the Euclidean space at distance greater than or equal to their distance in the graph.

Definition 4. The *distortion of an embedding* is given by

$$\operatorname{distor}(f) = \frac{\max\limits_{v,u} \frac{\|(f(v)-f(u)\|}{d_G(v,u)}}{\min\limits_{v,u} \frac{\|(f(v)-f(u)\|}{d_G(v,u)}},$$

where $v \neq u$ runs over all pairs of distinct vertices of G.

Remark. The distortion of an embedding is a characteristic which describes the difference between the natural metric on the graph and the induced Euclidean metric on the image of the graph.

For simplicity, let $g(v,u) = \frac{\|(f(v)-f(u)\|}{d_G(v,u)}$. Then $\max g = \max\limits_{v,u \in V(G)} \frac{\|(f(v)-f(u)\|}{d_G(v,u)}$ and $\min g = \min\limits_{v,u} \frac{\|(f(v)-f(u)\|}{d_G(v,u)}$, where v and u run over all pairs of distinct vertices of G. By definition, a non-contracting embedding f has corresponding g with $\min g \geq 1$, so $\operatorname{distor}(f) = \frac{\max g}{\min g} \leq \max g$.

Definition 5. The *distortion of a graph* G is given by

$$D(G) = \inf_{f \in \mathcal{E}(G)} \operatorname{distor}(f),$$

where $\mathcal{E}(G)$ denotes the set of all embeddings of the graph G in $\mathbb{R}^d$.

In other words, the distortion of a graph is the minimal distortion among all of its embeddings.

Let us illustrate Definitions 2 – 5 with an example.

Consider the embeddings as described in Figure 1. For f_1 we have that

$$g(v_1,v_2) = \tfrac{2}{1}, \quad g(v_4,v_1) = \tfrac{1}{1}, \quad g(v_2,v_3) = \tfrac{\sqrt{2}}{1},$$
$$g(v_2,v_4) = \tfrac{\sqrt{5}}{2}, \quad g(v_3,v_4) = \tfrac{1}{1}, \quad g(v_1,v_3) = \tfrac{\sqrt{2}}{2}.$$

Therefore

$$\operatorname{distor}(f_1) = \frac{\max g}{\min g} = \frac{2}{\frac{\sqrt{2}}{2}} = \sqrt{2}.$$

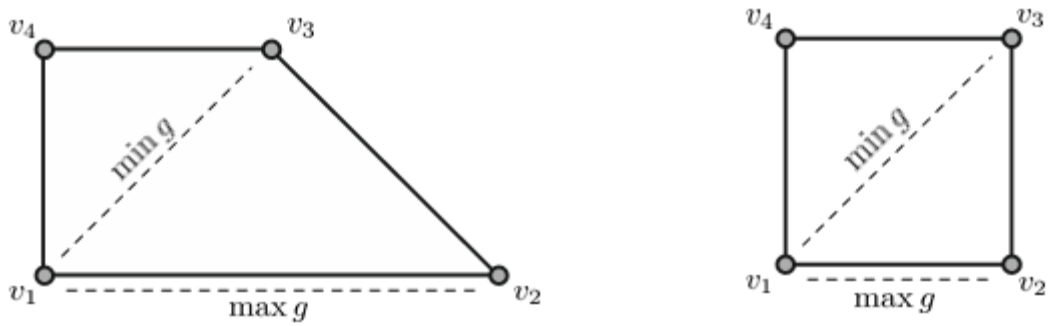


Figure 1: Two different embeddings f_1 and f_2 of the same graph G into $\mathbb{R}^2$

We can similarly compute

$$\text{distor}(f_2) = \frac{1}{\frac{\sqrt{2}}{2}} = \frac{\sqrt{2}}{2}.$$

A geometrical argument implies that any other embedding of G has distortion at least $\frac{\sqrt{2}}{2}$, thus the distortion of the graph G is

$$D(G) = \text{distor}(f_2) = \frac{\sqrt{2}}{2}.$$

3 Preliminaries

We present some general results and definitions, which are either not widely used, or specific for our paper.

Proposition 3.1. *Let G be a finite connected graph. There exist $v', u' \in V(G)$ with $d_G(v', u') = 1$ such that $\max\limits_{v,u \in V(G)} \frac{\|f(v)-f(u)\|}{d_G(v,u)} = \|f(v') - f(u')\|$.*

Proof. Suppose that $\max\limits_{v,u \in V(G)} g(v,u)$ is attained at v_1 and v_k where $d_G(v_1, v_k) = k - 1 > 1$. Pick a vertex v_2 satisfying $d_G(v_1, v_2) = 1$ and $d_G(v_2, v_k) = k - 2$. Such a vertex always exists

because G is connected. From the triangle inequality we have

$$\|f(v_1) - f(v_2)\| + \|f(v_2) - f(v_k)\| \geq \|f(v_1) - f(v_k)\|.$$

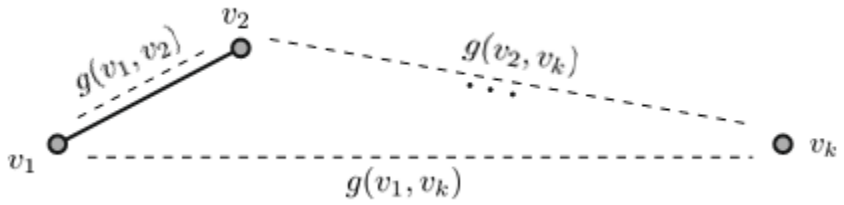


Figure 2: The embedding f

Hence at least one of the following holds:

Case 1. $\|f(v_1) - f(v_2)\| \geq \frac{1}{k-1}\|f(v_1) - f(v_2)\| \iff g(v_1, v_2) \geq g(v_1, v_k)$.

Note that here $d_G(v_1, v_2) = 1$ and $g(v_1, v_2) = \max g$.

Case 2. $\|f(v_2) - f(v_k)\| > \frac{k-2}{k-1}\|f(v_1) - f(v_2)\| \iff g(v_2, v_k) > g(v_1, v_k)$.

This contradicts the assumption that $g(v_1, v_k) = \max g$.

□

Proposition 3.2. *A perfect binary tree T_h of height h has $n = 2^{h+1} - 1$ vertices.*

Proof. The perfect binary tree T_h has 2^k vertices of any height $1 \leq k \leq h$. By summation of geomertic series we calculate the total number of vertices to be $n = 2^{h+1} - 1$. □

Remark. Because we are interested in estimating the distortion up to a constant, independent of n, we consider a sufficiently large perfect binary tree and discuss only the asymptotic bounds. Proposition 3.2 shows that we may use the approximation $h \approx \log_2 n$.

Lemma 3.3. *Any embedding of a perfect binary tree T_h with n verices into the d-dimensional Euclidean space has distortion $\Omega\left(\frac{n^{\frac{1}{d}}}{\log_2 n}\right)$.*

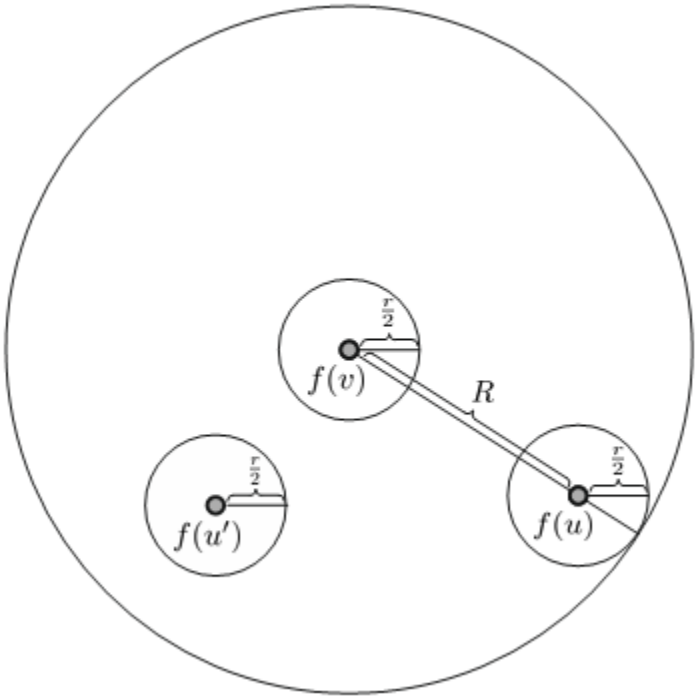


Figure 3: An arbitrary embedding of T_h

Proof. Suppose $f : T_h \to \mathbb{R}^d$ is an arbitrary embedding. Let $v, u \in V(G)$ be vertices with $\|f(v) - f(u)\| = \max g$. Denote this distance by R. Also let r be the minimal distance between any two image points of the embedding. For each image point $f(w)$, we draw a d-ball of radius $\frac{r}{2}$ centered at $f(w)$. By assumption, all those balls are disjoint from each other. Now draw a large ball of radius $R + \frac{r}{2}$ centered at $f(v)$. If there is a small ball with radius $\frac{r}{2}$ and center $f(u')$, which is not fully contained in the one with radius $R + \frac{r}{2}$, then $\|f(v) - f(u')\| > \|f(v) - f(u)\|$, which is a contradiction. Hence the large ball with center $f(v)$ and radius $R+\frac{r}{2}$ contains all the small balls. A straightforward volume argument implies that $(R + \frac{r}{2})^d \geq n(\frac{r}{2})^d$. Therefore

$$\frac{R}{r} \geq \frac{1}{2} \cdot n^{\frac{1}{d}}.$$

Recall that $g(v, u)$ denotes $\frac{\|(f(v)-f(u)\|}{d_G(v,u)}$ and note we have $\min g \leq \frac{r}{1}$ and $\max g \geq \frac{R}{2h}$. Hence

we can estimate the minimal possible distortion of any embedding of T_h by

$$\operatorname{distor}(f) = \frac{\max g}{\min g} \geq \frac{1}{2h} \cdot \frac{R}{r} \geq \frac{\frac{1}{2} \cdot n^{\frac{1}{d}} - 1}{2h} = \Omega\left(\frac{n^{\frac{1}{d}}}{\log_2 n}\right).$$

Since the embedding f is picked arbitrarily, this implies $D(T_h) = \Omega\left(\frac{n^{\frac{1}{d}}}{\log_2 n}\right)$. □

Definition 6. To each vertex v of height $h(v)$ we associate a sequence

$$\{a_1, a_2, \ldots, a_{h(v)-1}, a_{h(v)}, \ldots, a_h, \}$$

such that

$$a_i = \begin{cases} 1, & \text{if the ancestor of } v \text{ of height } i \text{ is a right child;} \\ -1, & \text{if the ancestor of } v \text{ of height } i \text{ is a left child;} \\ 0, & \text{if } \ i \geq h(v). \end{cases}$$

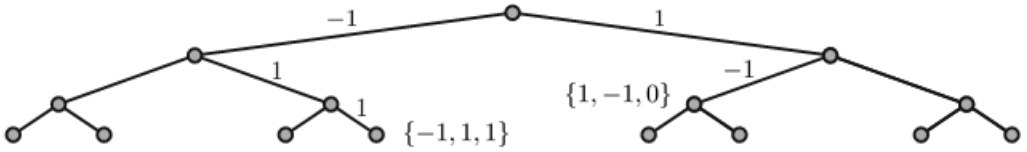


Figure 4: An example for the associated sequences

Proposition 3.4. *Suppose $v, u \in V(T_h)$ and let the two associated sequences be $\{a_1, \ldots, a_h\}$ and $\{b_1, \ldots, b_h\}$, respectively. Suppose j is the smallest index such that $a_{j+1} \neq b_{j+1}$. Then it holds that $d_G(v, u) = h(v) + h(u) - 2j$.*

Proof. For every $k \leq j$, we have $a_k = b_k$. Therefore the vertex w, associated with the sequence $\{a_1, \ldots, a_j, 0, 0, \ldots, 0\}$ is the common ancestor of v and u of greatest height, which is j. Hence we have

$$d_G(v, u) = d_G(v, w) + g_G(u, w) = h(v) - h(w) + h(u) - h(j) = h(v) + h(u) - 2j.$$

□

Definition 7. Let $\{x_k\}$ denote the sequence such that $x_1 = d$ and

$$x_i = 2d \cdot 2^{i-1} - d$$

for each $1 < i \leq k$

Note. Let us motivate the introduction of this sequence. For the purposes of our project we make use of one primal construction. Suppose we embed the leaves of T_k in a line from left to right, starting with the vertex associated with $\{-1, -1, \ldots, -1, -1\}$, then $\{-1, -1, \ldots, -1, +1\}$, $\{-1, -1, \ldots, +1, -1\}$, $\{-1, -1, \ldots, +1, +1\}$ and so on. For $v, u \in V(T_k)$ such that $f(v)$ and $f(u)$ are embedded consecutively on the line, we embed them so that $\|f(v) - f(u)\| = d \cdot d_G(v, u)$. Then the half-length of the segment, containing all v such that $h(v) = k$ by induction is estimated to be

$$x_k = d \cdot k + x_{k-1} + x_{k-2} + \cdots + x_1$$

where $x_1 = 1$. However, this construction is for embedding into a line, and for the d-dimensional case the construction is similar, but the used sequence starts with

$$x_1 = d.$$

Similarly, for each $1 < i \leq k$ we have

$$x_{i+1} - x_i = d + x_i$$

$$x_i = 2d \cdot 2^{i-1} - d.$$

4 A basis for the d-dimensional case

We present our initial construction of an embedding, which achieves distortion $c \cdot n^{\frac{1}{d}}$, where c is a constant, independent of n. This is higher than the lower bound, found in Lemma 3.3, but is instrumental for our final construction, which defines an embedding with distortion $O\left(\frac{n^{\frac{1}{d}}}{\log_2 n}\right)$.

Let us begin with an intuitive sketch. First, we group the leaves of T_h into 2^{h-d} groups, consisting of 2^d vertices with a common ancestor of height $h-d$. This implies that each two vertices in a group have distance in the graph at most $2d$. For each group, we map the leaves on the vertices of a d-cube with sidelength $2d$. Then, each two of those 2^{h-d} cubes are arranged in a d-lattice with appropriately chosen distances, based on the maximal distance in the graph between two vertices of the cubes. We have now mapped all of the leaves.

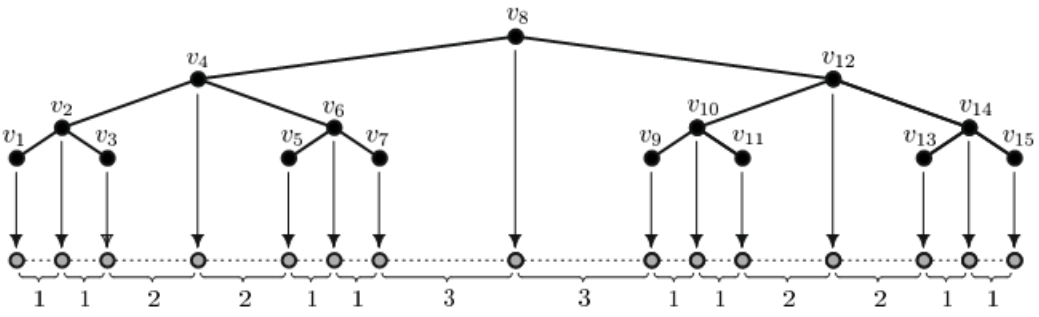


Figure 5: The embedding of T_3 into $\mathbb{R}^1$

Each parent is placed exactly in the middle between its children. This inductively gives the exact position of all of the vertices. We then prove that this embedding is non-contracting and estimate the sidelength of the d-cube, containing the entire perfect binary tree.

Remark. The two-dimensional case is similar to the well-known construction of an H-tree but with distances adjusted to make the embedding non-contracting.

Let us define the exact embedding f of T_h into the d-dimensional Euclidean space explicitly. There exists an integer m such that $md \geq h \geq (m-1)d$.

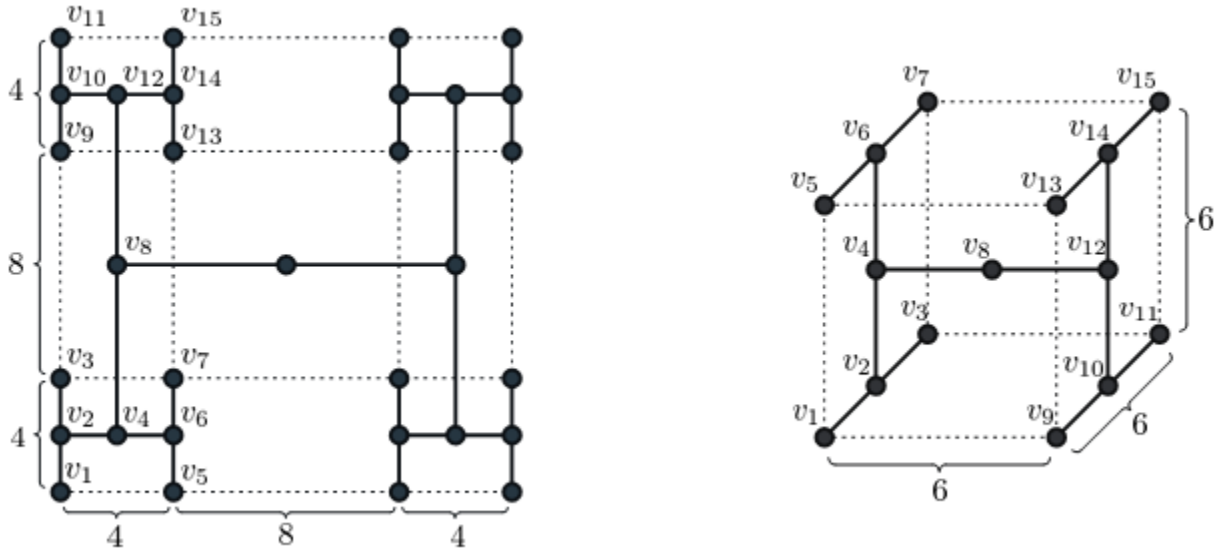


Figure 6: The embeddings of T_3 into $\mathbb{R}^2$ and $\mathbb{R}^3$

Construction 1. For every $v \in V(T)$ with associated sequence $\{a_1, \ldots, a_h\}$ the image point $f(v)$ is written as

$$f(v) = (f_1(v), f_2(v), \ldots, f_d(v)) \in \mathbb{R}^d$$

where for each $1 \leq j \leq d$ we set

$$f_j(v) = \sum_{i=0}^{m-1} a_{id+j} \cdot x_{m-i}.$$

Theorem 4.1. *Fix a positive integer d. Suppose T_h is the perfect binary tree of height h and $n = 2^{h+1} - 1$ vertices. Then there exists a non-contracting embedding f of T_h into the d-dimensional Euclidean space, so that the image of T_h is contained in an d-cube with sidelength $c \cdot n^{\frac{1}{d}}$ for some constant c independent of n.*

Proof. Let us first estimate the sidelength of a d-dimensional cube centered at the origin and covering the image of f. It is at most

$$2(x_m + x_{m-1} + \cdots + x_1) = 2(2d \cdot (2^m - 1) - md)$$
$$\leq 4 \cdot t \cdot 2^m = 8 \cdot d \cdot 2^{m-1} \leq 8d \cdot 2^{\frac{h}{d}} \leq 8d \cdot n^{\frac{1}{d}} \sim n^{\frac{1}{d}}.$$

Next we prove that the constructed embedding is non-contracting, i.e. that $\|\|f(v), f(u)\|\| \geq d_G(v,u)$ holds for every $v, u \in V(T_h)$. Suppose v and u are associated with $\{a_1, \ldots, a_h\}$ and $\{b_1, .., b_h\}$, respectively. Applying 3.4 gives $d_G(v,u) = h(v) + h(u) - 2j$ where j is the smallest index such that $a_{j+1} \neq b_{j+1}$. Assume $d \mid j$. Other cases are similar. We have

$$\begin{aligned}
\|f(v) - f(u)\| &\geq |f_1(v) - f_1(u)| \\
&= \left| \sum_{i=\frac{j}{d}}^{m-1} (a_{id+1} - b_{id+1}) \cdot x_{m-i} \right| \\
&\geq \left| 2x_{m-\frac{j}{d}} - 2 \sum_{i=\frac{j}{d}+1}^{m-1} \cdot x_{m-i} \right| \qquad (1) \\
&\geq 2(x_{m-\frac{j}{d}} - x_{m-\frac{j}{d}-1} - \cdots - x_2 - x_1) \\
&= 2d(m - \frac{j}{d}) \geq 2h - 2j \geq d_G(v,u).
\end{aligned}$$

□

Before moving onto the final construction for the d-dimensional case, let us first examine the distortion of the embedding, obtained by using Construction 1. Suppose v_0 is the root and v_1 is its left child. Then we have

$$f(v_0) = (0, \ldots, 0)$$

$$f(v_1) = (-x_m, 0, \ldots, 0).$$

Hence

$$\|f(v_0) - f(v_1)\| = x_m \geq \frac{d}{4} n^{\frac{1}{d}} - d. \tag{2}$$

Furthermore, suppose v_2 is the left child of v_1 and v_3 is the left child of v_2, we compute similarly that

$$\|f(v_1) - f(v_2)\| = x_{m-1} \geq \frac{d}{8} n^{\frac{1}{d}} - d; \tag{3}$$

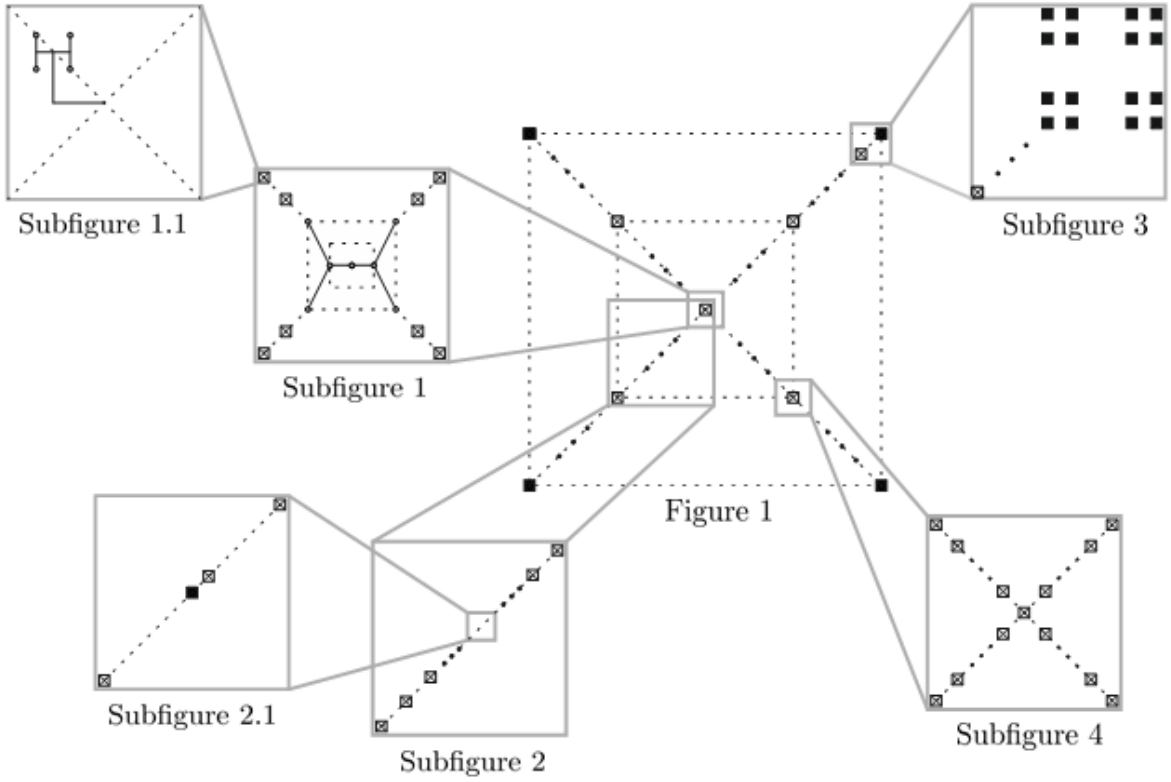
$$\|f(v_2) - f(v_3)\| = x_{m-2} \geq \frac{d}{16} n^{\frac{1}{d}} - d. \tag{4}$$

Since the map f is non-contracting, a straightforward argument shows that $\text{distor}(f) = \Theta(n^{\frac{1}{d}})$, which is higher than the lower bound, computed in Lemma 3.3. From Equations 2,3,4, we see that the large distortion appears between the vertices near the root point, while as the height of the vertex increases, the distance between the image points of a child and its parent, under the embedding f, decreases very quickly. Because of this observation, our strategy is to re-arrange the first few vertices so that they have distances roughly $\frac{n^{\frac{1}{d}}}{h}$ from their parents and children, which would result in achieving the lower bound of the distortion, proved in Lemma 3.3 to be $c \cdot \frac{n^{\frac{1}{d}}}{h}$ where c is a constant, independent of n.

5 Final construction for the d-dimensional case

Sketch. We first embed the binary trees rooted at the vertices of T_h of height $\frac{h}{2}$. There are $2^{\frac{h}{2}}$ such trees and each of them is a binary tree of height $\frac{h}{2}$. Each is embedded in a d-cube with sidelength $\sim n^{\frac{1}{2d}}$ applying Construction 2. (illustrated with black squares in Figure 1). Those we arrange in a d-cube, scaling them with an integer $M \sim n^{\frac{1}{2d}}$ to leave enough space for placing the remaining vertices of T_h of height less than $\frac{h}{2}$ in between the small cubes. The vertices of height at most $\frac{h}{2}$ we arrange in different d-cubes on approximately equal distances on segments, first taking the vertices of height k, where $1 \leq k \leq \frac{h}{4}$ (Subfigure 2), then taking

those of height $\frac{h}{4} < k \leq \frac{h}{8}$ and so on. As a result of this algorithm each vertex of height $\frac{h}{2}$ is placed at the center point of its small d-cube.On Figure 1 is given the embedding into $\mathbb{R}^2$ as an example. The Subfigures illustrate different zoomed parts of the embedding.



To define the embedding explicitly we need to introduce two more notations:

$$m = \left\lceil \frac{h}{2d} \right\rceil, \text{and}$$

$$M = d \cdot 2^{\lfloor \frac{h}{2d} \rfloor + 2} + 2h.$$

We may assume h is even. If it is not, increasing the height by 1 would affect the total distortion only by a constant. Let $v \in V(T_h)$ be associated with the sequence $\{a_1, \ldots, a_h\}$.

We define the image point $f(v) \in \mathbb{R}^d$ as

$$f(v) = (f_1(v), .., f_d(v)),$$

where

$$f_j(v) = M \cdot L_j(v) + S_j(v).$$

Note. We can think of $f_j(v)$ as composed of a "large scale" part $L_j(v)$, scaled by an integer M, and a "small scale" part $S_j(v)$.

Construction 2. For $v \in V(T_h)$ we define the image point $f(v) = (f_1(v), \ldots, f_d(v))$ such that for each $1 \leq j \leq d$ we have

$$f_j(v) = M \cdot L_j(v) + S_j(v).$$

We define the j-th coordinate of the image point $f(v)$ in two cases, depending on the height of the vertex v.

Case 1. For vertices v such that $h(v) \geq \frac{h}{2}$:

$$S_j(v) = \sum_{i=0}^{m-1} a_{id+j} \cdot x_{m-i}$$

$$L_j(v) = \sum_{i=0}^{m-1} a_{id+j} \cdot x_{m-i}.$$

Case 2. For vertices v such that $h(v) < \frac{h}{2}$:

There exists a positive integer p such that

$$(1 - \frac{1}{2^p}) \times \frac{h}{2} \leq h(v) < (1 - \frac{1}{2^{p+1}}) \times \frac{h}{2}.$$

Now the value of the j-th coordinate $f_j(v)$ is defined as $M \cdot L_j(v) + S_j(v)$ whereas

$$S_j(v) = \sum_{i=0}^{m-1} a_{id+j} \cdot x_{m-i}$$

$$L_j(v) = \begin{cases} L'_j(v), & \text{if } \ L'_j(v) \neq L_j(u) \text{ for any } u \text{ such that } \ h(u) \geq \frac{h}{2}; \\ L'_j(v) + 1, & \text{if } \ L'_j(v) = L_j(u) \text{ for some } u \text{ such that } \ h(u) \geq \frac{h}{2}. \end{cases}$$

Here

$$L'_j(v) = \sum_{i=0}^{p-1} a_{id+j} \cdot x_{m-i} + a_{pd+j} \cdot \lfloor \frac{h(v) - (1 - \frac{1}{2p}) \cdot \frac{h}{2}}{\frac{1}{2^{p+1}} \cdot \frac{h}{2}} \cdot x_{m-p} \rfloor.$$

Proposition 5.1. *If $v, u \in V(T)$ such that $L_j(v) \neq L_j(u)$ for a coordinate $1 \leq j \leq d$, then*

$$\|f(v) - f(u)\| > d_G(v, u).$$

Proof.

$$f_j(v) \leq M \cdot L_j(v) + \min_{v,u \in V(T_h)} S_j(v)$$

$$f_j(u) \geq M \cdot L_j(u) + \max_{v,u \in V(T_h)} S_j(v).$$

Therefore

$$\|f_j(v) - f_j(u)\| \geq M \cdot (L_j(v) - L_j(u)) - 2 \max_{v,u \in V(T)} |S_j(v)|$$

But we also have

$$|L_j(v) - L_j(u)| \geq 1 \text{ and}$$

$$M = 2d \cdot \max_{v,u \in V(T_h)} |S_j(v)| + 2h.$$

Hence

$$\|f(v) - f(u)\| \geq M \cdot (L_j(v) - L_j(u)) - 2 \max_{v,u \in V(T_h)} |S_j(v)| \geq M - 2 \max_{v,u \in V(T_h)} |S_j(v)| \geq 2h \geq d_G(v, u).$$

□

Lemma 5.2. *The embedding f, as defined in Construction 2, is non-contracting.*

Proof. By definition for the image point $f(v)$ of the vertex v we have

$$f(v) = (f_1(v), .., f_d(v)) \in \mathbb{R}^d$$

where for each $1 \leq j \leq d$ we have

$$f_j(v) = M \cdot L_j(v) + S_j(v).$$

Assume $L_j(v) \neq L_j(u)$ for some $1 \leq j \leq d$. Then by Proposition 5.1 $\|f(v) - f(u)\| \geq d_G(v, u)$. So we can focus on the case where $L(v) = L(u)$. Then $\|f(v) - f(u)\| = \|S(v) - S(u)\|$.

Case 1. $h(v), h(u) \geq \frac{h}{2}$

By the way the construction is set up, this implies that v and u have a common ancestor of height $\frac{h}{2}$ and

$$h(v), h(u) \geq \frac{h}{2}.$$

Recall that those v and u are mapped in small scale, using Construction 1 for the subtree rooted at a vertex of height $\frac{h}{2}$ which implies that

$$\|f(v) - f(u)\| \geq d_G(v, u).$$

Case 2. $h(v), h(u) < \frac{h}{2}$

Apply Theorem 4.1 for the subtree of height $\frac{h}{2}$ rooted at the root of T_h. Contradiction.

Case 3. $h(v) \geq \frac{h}{2}$ and $h(u) < \frac{h}{2}$

We have defined $L(v)$ for vertices v of height $h(v) \geq \frac{h}{2}$ as

$$L_j(u) = \begin{cases} L'_j(u), & \text{if } L'_j(u) \neq L_j(w) \text{ for any } w \text{ such that } h(w) \geq \frac{h}{2} \\ L'_j(u) + 1, & \text{if } L'_j(u) = L_j(w) \text{ for some } w \text{ such that } h(w) \geq \frac{h}{2} \end{cases}$$

If the first case happens, $L(v) \neq L(u)$. If the second case happens, then we should argue that $L(v) \neq L(u)$. Denote by $\tilde{v}$ the ancestor of v of height $\frac{h}{2}$ and similarly for $\tilde{w}$. Now by Construction 2 we have $L(v) = L(\tilde{v})$ and $L(w) = L(\tilde{w})$. If $L(\tilde{v}) = L(\tilde{w})$ then $L(v) = L(w) \neq L(u)$. Hence $L(\tilde{v}) \neq L(\tilde{w})$ holds. By Theorem 4.1 for the subtree of height $\frac{h}{2}$ rooted at the root of T_h there exists j such that $\|L_j(\tilde{v}) - L_j(\tilde{w})\| \geq 2$. Therefore $L(v) \neq L(u)$. □

Lemma 5.3. *Let f be the embedding, defined by using Construction 2. Then we have $\max_{v,u \in V(T)} g \leq c \cdot \frac{\sqrt{n}}{\log_2 n}$ where c is a constant, independent of n.*

Proof. Let us recall the intuitive description of the construction. We have first embedded the binary trees rooted at the vertices of T_h of height $\frac{h}{2}$. Those are $2^{\frac{h}{2}}$ of number and each of them is a binary tree of height $\frac{h}{2}$. Each is embedded in a d-cube with sidelenght $\sim n^{\frac{1}{2d}}$. Those we have arranged in a big d-cube, scaling them with an integer $M \sim n^{\frac{1}{2d}}$. Therefore the sidelenght of the big d-cube, containing all $v \in V(T_h)$ is $\sim n^{\frac{1}{d}}$. Recall that our strategy for embedding the vertices v such that $h(v) < \frac{h}{2}$ is to arrange them on approximately equal distances in different lines, which have approximately the same lenght as the sidelenght of the big cube. Therefore we divide by $2 \cdot \frac{h}{2} \sim \log_2 n$. Now for each $v, u \in V(T_h)$ such that $h(v), h(u) < \frac{h}{2}$ and $d_G(v, u) = 1$ we have $\|f(v) - f(u)\| \sim \frac{n^{\frac{1}{d}}}{\log_2 n}$. We show that this is actually the $\max g$ as well. For a more detailed proof, refer to Appendix A. □

Theorem 5.4. *Fix any positive integer d. Suppose T_h is a perfect binary tree with n vertices. Then the distortion of the graph T_h is $D(T_h) = \Theta\left(\frac{\sqrt{n}}{\log_2 n}\right)$.*

Proof. We have presented a non-contracting embedding f of T_h into the d-dimensional Eu-

clidean space in Construction 1. Also, in Lemma 5.3 we prove that $\max\limits_{v,u\in V(T_h)} g(v,u) = c \cdot \frac{n^{\frac{1}{d}}}{\log_2 n}$ for some constant c independent of n. Therefore we obtain

$$\operatorname{distor} f = c \cdot \frac{n^{\frac{1}{d}}}{\log_2 n}$$

which implies

$$D(T_h) = O\left(\frac{n^{\frac{1}{d}}}{\log_2 n}\right).$$

On the other hand, we also have $D(T_h) = \Omega\left(\frac{n^{\frac{1}{d}}}{\log_2 n}\right)$ by Lemma 3.3. Thus we conclude that

$$D(T_h) = \Theta\left(\frac{n^{\frac{1}{d}}}{\log_2 n}\right).$$

□

6 Conclusion

By a geometrical argument we prove that any embedding of a binary tree with n vertices into $\mathbb{R}^d$ has distortion $\Omega\left(\frac{n^{\frac{1}{d}}}{\log_2 n}\right)$. We have also constructed an embedding, achieving distortion $O\left(\frac{n^{\frac{1}{d}}}{\log_2 n}\right)$. Therefore our main result is proving that the distortion of embedding binary trees into d-dimensional Eucliden space is $\Theta\left(\frac{n^{\frac{1}{d}}}{\log_2 n}\right)$. There are many possible avenues for future development of this project. For instance, we may estimate tgrowth of the constant as a function of the number of dimensions d. We can also try to modify the embedding into $\mathbb{R}^2$ to have all of the edges of the tree embedded into strict lines. A natural next step is to look at k-ary trees, binary trees being a partial instance of these for $k = 2$.

7 Acknowledgments

I would like to thank my mentor Mr. Zhenkun Li for suggesting the topic of this paper and guiding my work. I would like to express my deepest gratitude towards my tutor Dr. John Rickert for all of his help for writing the paper. I am also very indebted to Dr. Tanya Khovanova, Prof. Gerovitch, Prof. Jerison and Prof. Moitra . I want to thank as well Jenny Sendova, Stanislav Atanasov, Rayna Gadzheva and Peter Gaydarov for all of the fruitful conversations and invaluable help in writing this paper. I am also thankful to America for Bulgaria Foundation, St. Cyril and St. Methodius International Foundation, Evrika Foundation and The Union of Bulgarian Mathematician whose sponsorship made my stay at RSI possible. Lastly, I would like to thank RSI, MIT and CEE for providing me with the facilities to conduct this research.

References

[1] J. Bourgain. On Lipschitz embedding of finite metric spaces in Hilbert space. *Israel J. Math*, 52:46–52, 1985.

[2] J. Matoušek. Bi-Lipschitz embeddings into low-dimensional Euclidean spaces. *Commentationes Mathematicae Universitatis Carolinae*, 33(3):589–600, 1990.

[3] J. M. R. Babilon, J. Matoušek and P. Valtr. Low-distortion embeddings of trees. *Proc. Graph Drawing*, 42:343–351, 2002.

[4] J. Matoušek. On embedding trees into uniformly convex Banach spaces. *Israel Journal of Mathematics*, 144:221–237, 1999.

[5] E. M. M. Kumamoto. Optimal distortion of embedding of complete binary trees into lines. *Information Processing Letters*, 112:365–370, 2012.

Appendix A Detailed proof of Lemma 5.3

From Proposition 3.1 we have that it is enough to estimate the distance for vertices v and u such that v is a child of u. Hence we have $h(v) = h(u) + 1$.

Case 1. $h(u) < \frac{h}{2}$ and $h(u) + 1 \geq \frac{h}{2}$

There exists p such that

$$(1 - \frac{1}{2^p}) \cdot \frac{h}{2} \leq h(u) < (1 - \frac{1}{2^{p+1}}) \cdot \frac{h}{2}$$

Now we have that

$$\frac{h}{2} - 1 < h(u) \leq (1 - \frac{1}{2^{p+1}}) \cdot \frac{h}{2}$$

$$h \leq 2^{p+2}$$

From Construction 1. we have

$$L_j(u) = \sum_{i=0}^{p-1} a_{id+j} \cdot x_{m-i} + \left\lfloor \frac{h(u) - (1 - \frac{1}{2^p}) \cdot \frac{h}{2}}{\frac{1}{2^{p+1}} \frac{h}{2}} \cdot x_{m-p} \right\rfloor \cdot a_{pd+j}$$

$$L_j(v) = \sum_{i=0}^{m-1} a_{id+j} \cdot x_{m-i}.$$

Now we have

$$\begin{aligned}
|L'_j(u) - L'_j(v)| &= |\sum_{i=0}^{p-1} a_{id+j} \cdot x_{m-i} - \sum_{i=0}^{m-1} a_{id+j} \cdot x_{m-i} + \left\lfloor \frac{h(u) - (1 - \frac{1}{2^p}) \cdot \frac{h}{2}}{\frac{1}{2^{p+1}} \frac{h}{2}} \cdot x_{m-p} \right\rfloor \cdot a_{pd+j}| \\
&\leq \left| \sum_{i=p}^{m-1} a_{id-j} \cdot x_{m-i} \right| + \left| \left\lfloor \frac{h(u) - (1 - \frac{1}{2^p}) \cdot \frac{h}{2}}{\frac{1}{2^{p+1}} \frac{h}{2}} \cdot x_{m-p} \right\rfloor \right| \\
&\leq |x_{m-p}| + \sum_{i=p+1}^{m-1} |x_{m-i}| + x_{m-p} \leq 3|x_{m-p}| \\
&= 12d \cdot \frac{2^m}{2^{p+2}} \leq 12d \cdot \frac{2^m}{h} \leq 12d \cdot \frac{2^{\frac{h}{2d}} + 1}{h} \leq 24d \cdot \frac{n^{\frac{1}{2d}}}{h}
\end{aligned}$$

Case 2.1. $(1-\frac{1}{2^p})\cdot\frac{h}{2} \le h(u), h(u)+1 < (1-\frac{1}{2^{p+1}})\cdot\frac{h}{2}$

We have

$$\begin{aligned}
|L'_j(u) - L'_j(v)| &= \lfloor \frac{h(u)+1-(1-\frac{1}{2^p})\cdot\frac{h}{2}}{\frac{1}{2^{p+1}}\frac{h}{2}} \cdot x_{m-p} \rfloor - \lfloor \frac{h(u)-(1-\frac{1}{2^p})\cdot\frac{h}{2}}{\frac{1}{2^{p+1}}\frac{h}{2}} \cdot x_{m-p} \rfloor \\
&\le \frac{h(u)+1-(1-\frac{1}{2^p})\cdot\frac{h}{2}}{\frac{1}{2^{p+1}}\frac{h}{2}} \cdot x_{m-p} - \frac{h(u)-(1-\frac{1}{2^p})\cdot\frac{h}{2}}{\frac{1}{2^{p+1}}\frac{h}{2}} \cdot x_{m-p} \\
&= \frac{2^{p+2}}{h} \cdot x_{m-p} = d \cdot \frac{2^m - p + p + 2}{h} \\
&= 4d\frac{2^{\lceil \frac{h}{2d} \rceil}}{h} \le 8d \cdot \frac{2^{\frac{h}{2d}}}{h} \le 8d \cdot \frac{n^{\frac{1}{2d}}}{h}.
\end{aligned}$$

Case 2.2. $(1-\frac{1}{2^p})\cdot\frac{h}{2} \le h(u) < (1-\frac{1}{2^{p-1}})\cdot\frac{h}{2} \le h(u)+1$

Case 2.2.1 $\frac{1}{2^{p+2}}\cdot\frac{h}{2} > 1$

Then we can compute that $(1-\frac{1}{2^{p+1}})\cdot\frac{h}{2} < h(u)+1 < (1-\frac{1}{2^{p+2}})\cdot\frac{h}{2}$.

$$L'_j(u) = \sum_{i+0}^{p-1} [a_{id+j} \cdot x_{m-i}] + \lfloor \frac{h(u)-(1-\frac{1}{2^p})\cdot\frac{h}{2}}{\frac{1}{2^{p+1}}\cdot\frac{h}{2}} \cdot x_{m-p} \rfloor \cdot a_{pd+j}$$

$$L'_j(v) = \sum_{i+0}^{p-1} [a_{id+j} \cdot x_{m-i}] + a_{pd+j} \cdot x_{m-p} + \lfloor \frac{h(u)+1-(1-\frac{1}{2^p})\cdot\frac{h}{2}}{\frac{1}{2^{p+1}}\cdot\frac{h}{2}} \cdot x_{m-(p+1)} \rfloor \cdot a_{(p+1)d+j}$$

Now we have

$$\begin{aligned}
|L'_j(u) - L'_j(v)| &\leq |x_{m-p} - \frac{h(u) - (1-\frac{1}{2^p})\cdot\frac{h}{2}}{\frac{1}{2^{p+1}\cdot\frac{h}{2}}} \cdot x_{m-p} + 1| + |\frac{1}{\frac{1}{2^{p+2}}\cdot\frac{h}{2}} \cdot x_{m-(p+1)}| \\
&\leq |\frac{2^{p+2}\cdot\frac{h}{2} - h(u) + (1-\frac{1}{2^p})\cdot\frac{h}{2}}{2^{p+2}\cdot\frac{h}{2}} \cdot x_{m-p} + 1| + |\frac{1}{\frac{1}{2^{p+2}}\cdot\frac{h}{2}} \cdot x_{m-(p+1)}| \\
&= |\frac{(1-\frac{1}{2^{p+2}})\cdot\frac{h}{2} - h(u)}{2^{p+2}\cdot\frac{h}{2}} \cdot x_{m-p} + 1| + |\frac{1}{\frac{1}{2^{p+2}}\cdot\frac{h}{2}} \cdot x_{m-(p+1)}| \\
&\leq |\frac{x_{m-p}}{2^p \cdot h} + 1| + |\frac{1}{\frac{1}{2^{p+2}}\cdot\frac{h}{2}} \cdot x_{m-(p+1)}| \\
&\leq c \cdot \frac{n^{\frac{1}{2d}}}{h}.
\end{aligned}$$

Case 2.2.2 $\frac{1}{2^{p+2}\cdot}\frac{h}{2} < 1$

There exists an integer $q > p$ such that

$$(1 - \frac{1}{2^q}) \cdot \frac{h}{2} \leq h(u) + 1 \leq (1 - \frac{1}{2^{q+1}}).$$

We have

$$L'_j(u) = \sum_{i=0}^{p-1} [a_{id+j} \cdot x_{m-i}] + a_{pd+j} \cdot \lfloor \frac{h(u) - (1-\frac{1}{2^p})\cdot\frac{h}{2}}{\frac{1}{2^{p+1}}\cdot\frac{h}{2}} \cdot x_{m-p} \rfloor$$

$$L'_j(v) = \sum_{i=0}^{q-1} [a_{id+j} \cdot x_{m-i}] + a_{qd+j} \cdot \lfloor \frac{h(u) + 1 - (1-\frac{1}{2^q})\cdot\frac{h}{2}}{\frac{1}{2^{q+1}}\cdot\frac{h}{2}} \cdot x_{m-p} \rfloor.$$

$$
\begin{aligned}
|L_j(v) - L_j(u)| &\leq |\sum_{i=p}^{q-1} a_{id+j} \cdot x_{m-i}| + x_{m-p} + x_{m-q} \\
&\leq 3x_{m-p} + |\sum_{i=p+1}^{q-1} a + id + j \cdot x_{m-i}| \\
&\leq 4x_{m-p} \\
&\leq c \cdot \frac{n^{\frac{1}{2d}}}{h}.
\end{aligned}
$$

In all of the cases we end up with

$$|L'_j(v) - L'_j(u)| \leq c_1 \cdot \frac{n^{\frac{1}{2d}}}{h}.$$

Recall that if for some vertex v $L_j(v) \neq L'_j(v)$ then $L_j(v) = L'_j(v) + 1$. Therefore we have $|L_j(v) - L_j(u)| \leq |L'_j(v) - L'_j(u)| + 2$. This will not influence the value of the distortion only by a constant, independant of n, so we can assume $|L_j(v) - L_j(u)| \approx |L'_j(v) - L'_j(u)|$. Now we have

$$|L_j(v) - L_j(u)| \leq c_1 \cdot \frac{n^{\frac{1}{2d}}}{h}.$$

$$\|L(v) - L(u)\| \leq \sum_{i=1}^{d} |L_j(v) - L_j(u)| \leq c_2 \cdot \frac{n^{\frac{1}{2d}}}{h}.$$

Recall that

$$f(v) = M \cdot L(v) + S(v)$$

where $M \sim n^{\frac{1}{2d}}$ and Lemma (ref) implies that $\max_{v \in V(T)} S(v) \sim n^{\frac{1}{2d}}$ as well. Hence we have

$$f(v) \leq c \cdot \frac{n^{\frac{1}{2}}}{h}.$$