CORSO DI LAUREA IN MATEMATICA
Università di Pisa

The random Euclidean minimum spanning tree

Tesi di Laurea Triennale

CANDIDATO:
Mario Correddu

RELATORE:
Dott. **Dario Trevisan**

Anno Accademico 2019/2020

Contents

Introduction

The minimum spanning tree is one of the easiest problem of combinatorial optimisation to define and consists , given a connected graph, in finding the spanning acyclic subgraph (tree) with the smallest total edge weight. Its applications span a multitude of fields, the most obvious being computer science and communication networks, but it is also widely used as a tool for clustering and classification problems, image processing and automatic speech recognition, to name a few. A common instance of this problem in applications, is when the nodes of the graph are points in $\mathbb{R}^d$ and the length of edges of the graph are a power of the euclidean norm of the points, called Euclidean minimum spanning tree.

The aim of this work is to study the properties of the Euclidean MST where the points are not supposed to be fixed, but are uniform random variables in $[0,1]^d$. While there is a good amount of articles that study the properties of the MST on the complete random graph generated by the points, very few works about the bipartite case have been published. However, recently in [2] it was studied the bipartite problem when $d = 1$ and found an exact formula for the average cost . Part of the work on this thesis was dedicated to try to extend the results to a generic d and to our knowledge we discovered this new result:

Theorem. *In the Euclidean bipartite toroidal model:*

$$\beta_T^B(d) := \lim_{n\to\infty} \frac{E[T_n]}{n^{\frac{d-1}{d}}} = \frac{\Gamma\left(\frac{1}{d}\right)}{d(c_d)^{\frac{1}{d}}} + \sum_{k=2}^{\infty} \frac{\Gamma\left(k+\frac{1}{d}-1\right)}{dk} \sum_{j=0}^{k-2} \frac{\eta_{j,k}(d)}{j!(k-j-1)!}$$

$$\eta_{j,k}(d) = \int_{\Theta_{j,k}(1)} \left(g_{j,1}(bu_1,\ldots,bu_j) + g_{k-j-1,1}(ru_1,\ldots,ru_{k-j-1})\right)^{-(k+\frac{1}{d}-1)} dbu_1\ldots dbu_j dru_1\ldots dru_{k-j-1}$$

where

- $E[T_n]$ *is the average cost of a bipartite MST on* n *random uniform points in the torus* $[0,1]^d/\sim$
- $\Theta_{j,k}(1)$ *is the subset made by every element of* $\Theta_k(z)$ *having exactly j blue nodes, which*

in turn is the subset of $\mathbb{R}^{d\times(k-1)}$ *made by every red and blue points,* $\{x_1,\dots x_{k-1}\}$*, such that there exists a tree that connects the points with edges of length* ≤ 1

- $g_{s,z}(x_1\dots x_{s-1})$ *is the the volume of* $\bigcup_{j=0}^{s-1} B(x_j,z)$.

In chapter 1, after a brief section dedicated to some essential definitions of graph theory, we focus on the minimum spanning tree problem showing some of its defining characteristics and the classic algorithms used to solve it.

In chapter 2 we explore the properties of the deterministic Euclidean minimum spanning tree, in particular we show a bound for its cost that depends only on the number of nodes, we prove that the degree is limited in any dimension d and also prove some relations that bind two MSTs of different sets of vertices.

In Chapter 3 we study the asymptotics of the mean cost of the random Euclidean MST. By using two different approaches, one that considers the points in the cube and the other that assume the points to be in a torus, we prove a result of convergence of the weighted mean cost for the MST problem.

In chapter 4 we switch our attention to the MST problem when the graph is bipartite. First we present an algorithm that is a direct generalisation of the algorithm presented in [2] in dimension d. Then we prove some deterministic bounds for the cost and use them to show some bounds on the constant of the mean weight. Then we concentrate on the maximum degree of the nodes and prove a bound when the blue points are on a grid and the red ones are uniformly distributed. Finally we prove the result of convergence of the mean weight when the points are in the torus.

Chapter 1

The minimum spanning tree problem

This chapter is dedicated to provide an introduction of the general problem, defining the setting of graph theory and every element needed for the following analysis as done in [3]. After a brief exposition of the essential definitions, there is a description of some classic algorithms used to solve the minimum spanning tree problem in generic graphs, that will be later used as theoretical tools to prove our statements.

1.1 Graphs

Definition 1.1. A graph G is a pair of sets (V,E) where V is the finite set of the vertices (or nodes) of the graph and $E \subset V \times V$ is the set of the edges of the graph. Given a function $W : E \to \mathbb{R}$, $G = (V,E,W)$ is called a weighted graph.

We say that $u, v \in V$ are adjacent if $(u, v) \in E$, while $w \in V$ is incident to $e = (x, y) \in E$ if $x = w$ or $y = w$. We call $N(u)$ the set of the nodes adjacent to u while $\#N(u)$ is called the degree of u.

A path is a sequence of edges $(e_1,\dots,e_k)$ such that $e_i = (v_{i-1}, v_i)$ for a sequence of vertices $(v_0,\dots, v_k)$. If $v_0 = v_k$ the path is called a cycle. Graphs containing no cycles are called acyclic. From paths we can define what connection means in graphs: two nodes u and v are connected if there is a path whose starting node is u and ending node is v. A graph is then connected if every couple of vertices is connected.

If $E = V \times V$, namely for any couple of vertices there is an edge that connects them, then G is called complete.

Given a graph $G = (V,E)$ and $G' = (V,E')$, G' is a subgraph of G if $V \subset V$ and $E' \subset E$. If G is a weighted graph, the weight (equivalently cost or length) of G' is defined as

$$\sum_{e \in E'} W(e).$$

A maximal connected subgraph of a graph G is called a component of G.

A multigraph is a graph where there can be more than one edge that connects two node. We note that if we are dealing with a weighted multigraph two edges connecting the same nodes can have different lengths.

1.2 The problem

Given a graph $G = (V, E)$ a spanning tree $T = (V', E')$ is a connected acyclic subgraph of G such that $V' = V$.

Given a connected weighted graph $G = (V, E, W)$, T is a minimum spanning tree of G if it is a spanning tree of G of minimum weight among all the possible spanning trees of G.

We note that we can suppose, without loss of generality, that the weights of G are all different, as one can set an ordering even among the edges of same weight.

We now present some basic property of the MST functional:

Given a graph G a cut is any set $C \subset E$ of edges such that $G' = (V, E - C)$ is not connected.

Theorem 1.1. *Given a graph $G = (V, E, W)$ (where all weights are different), a minimum spanning tree $T = (V, E')$, and an edge $e \in E$, we have:*

a) *$e \in E'$ if and only if there is a cut C such that $e \in C$ and e is the edge of minimum weight in C*

b) *$e \notin E'$ if and only if there is a cycle C' such that $e \in C'$ and e is the edge of maximum weight of C'*

Proof. **a)**If e is an edge of T, then by removing it from the tree we get a partition of the nodes V' and V''. We consider the cut C made by all edges that connects a node in V' to a node in V''. If e was not the minimum among the edges in C, then one could replace e with another edge and get a tree of cost less than the cost of the minimum spanning tree.

On the other hand if there is a cut C where e is of minimum length, then if $e \notin E'$, when added to E' it forms a cycle. This means that there is at least another node e' in the cycle that is in C. Thus, deleting e' from T and adding e generates another spanning tree of cost strictly inferior than the cost of T .

b) if $e \notin E'$, then $E \cup \{e\}$ has a cycle C' and $e \in C'$. If there was an edge $e' \in C'$ such that $W(e') > W(e)$ then $E \cup \{e\} \setminus \{e'\}$ would be a spanning tree of cost less than the cost of T.

If there is a cycle C' such that e' is the edge of maximum weight in C', then if e was also an edge of the minimum spanning tree, one could remove it and add one of the edges of the cycle that is also in the cut generated by e. The tree generated would have weight strictly inferior to the one of T.

□

1.3 Algorithms

In this section we will present some classic algorithms used to solve the MST problem. Apart from being useful as they provide an easy solution to the minimum spanning tree problem, they are also used as a tool for the theoretical aspect of the problem.

Kruskal's algorithm

Kruskal's algorithm starts with $E' = \emptyset$ and considers the set of the edges ordered according to their weight, in ascending order. Then, one by one, every edge (u, v) is either added to (V, E') or skipped:

- if u and v are already connected in (V, E'), then (u, v) cannot be added or it would generate a cycle;

- if u and v are not already connected, then it is possible to find a cut that divides them and, since the edges are ordered and no other edge of the cut was added before, then (u, v) is an edge of the minimum spanning tree.

After $n - 1$ edges have been added, (V, E') is a tree and the algorithm stops.

The following is a pseudo code for Kruskal algorithm where $G = (V, E)$, $Sort(E)$ is a function that returns the elements of the set E in descending order and $Component(E', u, v)$ is a function that returns true if u and v are in the same component in (V, E') and false otherwise:

```
function $Kruskal$ ( $G$) {
    $E' = \varnothing$ ;
    $X = Sort(E)$;
    while $\big(|E'| < n-1\big)$ {
        remove from $X$ the first edge $(u,v)$;
        if ( $Component(E',u,v)$ )
            $E' = E' \cup \{(u, v)\}$;
        }
    return $T = (V,E')$
}
```

The most critical point of the algorithm is how to implement $Component$, but one can use a data structure to keep track of the components at every iteration, granting cost $O(1)$ for every call of $Component$. Thus the cost of the algorithm is given by the sorting, that can be done in $O(|E|log(|E|)$ and the while cycle that contains only $O(1)$ operations leading to a cost of $O(|E|)$ for the whole cycle. Since $|E|$ can be at most V^2 then the run-time of $Kruskal$ is $O(|E|log(|V|)$.

Prim's algorithm

While Kruskal's algorithm adds one by one every edge until the tree is built, Prim's algorithm works on the nodes of the graph: given V' and E' the nodes and the edges added until that step, Prim's algorithm adds to E' the edge (u,v) of minimum weight in the cut $(V-V',V')$ as well as v to N'. The process is repeated until $V' = V$.

By the end of the algorithm we will have a connected graph (N,E') that contains only edges of the minimum spanning tree thanks to the cut property **a**), meaning that (N,E') is the minimum spanning tree of G.

In order implement it one can just calculate the minimum edge that crosses the cut at every iteration, but the run-time would be of $O(|E||V|)$. A more efficient way to do this is to keep track, for every $v \in V-V'$, of the minimum edge that connects V' to v at every iteration, for example by storing the vertices in a priority queue. By doing this, every time a node v is added to V’ one has to just check the edges that connect v to $V-V'$ and update the queue. If the costs of the operations inside the queue are low enough, for example by using an heap, it is possible to get a run-time of $O(|E|log(V))$ for the whole algorithm.

Here a pseudo code for the algorithm where PQ is the priority queue, $decreasekey(PQ,u,v)$ rearranges the order of the elements of the queue according to the weight of (u,v), $weight[v]$

the weight of the smallest edge that connects the node u to the tree and $pred[v]$ is the node $u \in N'$ such that $(u, v) \in PQ$. $pred$ is also the vector that represents the tree.

```
procedure Prim( G ) {
    choose a node r in N
    foreach( $i \in V$ ) {
        $pred[i] = r$;
        $weight[i] = \infty$;
    }
    $PQ = \{r\}$;
    while($PQ \neq \emptyset$){
        let $u$ be the first element of the queue
        $PQ = PQ - u$;
        foreach($(u, v), v \in V - V'$) {
            if ($w(u, v) < weight[v]$) {
                $decreasekey(PQ, u, v)$;
                $pred[v] = u$;
                $weight[v] = w(u, v)$;
                if ($v \notin PQ$) then $PQ = PQ \cup \{v\}$;
            }
        }
    }
}
```

Chapter 2

Euclidean MST

In this chapter we consider complete graphs where the vertices are points in $[0,1]$ and the length of the edges are a function ψ of the Euclidean norm. This particular case comes out as one of the most natural to study and it can be traced as far back as when the minimum spanning tree problem was first defined. Here we will show some very unique and useful properties of this case, as presented in [5], that arise due to the geometrical nature of the problem.

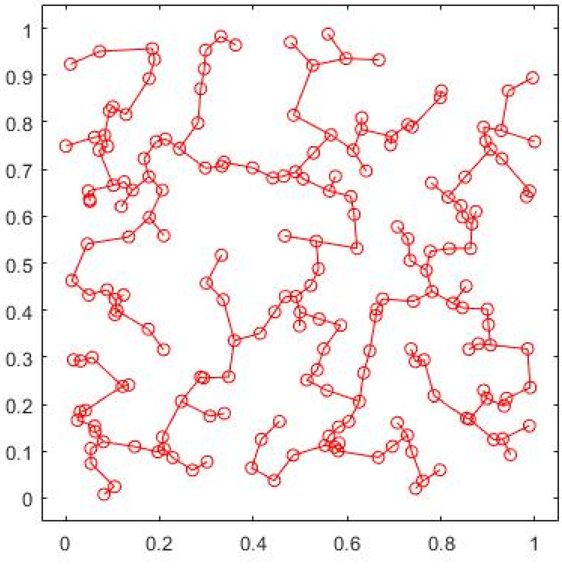


Figure 2.1: A Euclidean minimum spanning tree in the plane

Given n distinct points $S = \{x_1, x_2, \dots x_n\} \subset \mathbb{R}^d$ and $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ we call $G(S) = (S, S \times S, W)$ the graph generated from S with edge weight function ψ, where $W((x_i, x_j)) = \psi(|x_i - x_j|)$

and $|\cdot|$ is the Euclidean norm on R^d. We call $MST(S)$ the minimum spanning tree of G and $M(S)$ the length of $MST(S)$. We will also refer to $MST(S)$ as the minimum spanning tree of S.

Let $\psi(x) = |x|$, we define $v_d(x)$ as the function counting the number of edges of the minimum spanning tree of $\{x_1, \dots, x_n\}$ whose length is greater than x, namely:

$$\#\left\{e \middle| |e| > x\right\}$$

Through Lemma 2.1 we capture some information about the lengths of the edges, proving that they do not depend on n.

Lemma 2.1. There is a constant γ_d that depends only on the dimension d such that:

$$v_d(x) \le \gamma_d x^{-d}$$

for every $0 < x < \infty$.

Proof. First we note, as a consequence of the pigeonhole principle, that there is a constant α_d such that for every set of k points in $[0,1]^d$ one can pick two points x_i x_j such that

$$\left|x_i - x_j\right| < \alpha_d k^{-\frac{1}{d}}. \tag{2.1}$$

In fact if for any constant α we had a configuration such that $\left|x_i - x_j\right| \ge \alpha_d k^{-\frac{1}{d}}$ for any $i \neq j$, then there would be at least a region of the form $B(x_j, 2^{-1}\alpha_d k^{-\frac{1}{d}}) \cap [0,1]^d$ around each point x_j, devoid of any other point, where $B(x,r)$ is the d-ball of centre x and radius r. The area of these regions is minimised when the point x_j is in a corner of $[0,1]^d$ and equals to $c_d 2^{-2d}\alpha_d^d k^{-1}$, where c_d is the area of unit ball in $\mathbb{R}^d$. Then by summing every area around every point we would get that

$$c_d 2^{-2d}\alpha_d^d \le 1$$

for any constant α_d which is absurd.

Let then $\{x_1, \dots x_n\} \in [0,1]^d$, and E' the set of the edges of the minimum spanning tree of the graph generated by x_i.

We now consider Prim's algorithm.

If e_j, $1 < j < n$ are the edges added by the algorithm ordered according to when they are selected, when e_j is added there are exactly $(n - j + 1)$ connected components in the graph built until that time. Then by choosing a vertex from every component we can apply inequality 2.1 and get:

$$\left|e_j\right| \leq \alpha_d (n-j+1)^{-\frac{1}{d}}.$$

Thus if e_{i_j} $1 < j < k$ are any edges of E', since they are added in ascending order we have:

$$\sum_{j=1}^{k} \left|e_{i_j}\right| \leq \sum_{n-k}^{n-1} \left|e_j\right| \leq \alpha_d \sum_{i=2}^{k+1} i^{-\frac{1}{d}} \leq \tilde{\alpha}_d k^{\frac{d-1}{d}}$$

for a new constant $\tilde{\alpha}_d$. We apply the last inequality to $k = v_d(x)$ and get:

$$x v_d \leq \sum_{\substack{e \in T \\ |e| > x}} |e| \leq \tilde{\alpha}_d v_d^{\frac{d-1}{d}}.$$

Thus rearranging the terms we finally get $\gamma_d = \tilde{\alpha}_d^d$.

□

Thanks to Lemma 2.1 we can now prove some very useful bounds for $M(x_1, \ldots x_n)$ when $\psi(x)$ is a power function:

Theorem 2.2. *For any $\psi(x) = |x|^a$ and any minimal spanning tree T of $\{x_1, \ldots, x_n\}$ we have:*

a) $\sum_{e \in T} |e|^a \leq \gamma'(a,d) n^{\frac{d-a}{d}}$ *for* $0 < a < d$

b) $\sum_{e \in T} |e|^d \leq \gamma'(d) logn$

c) $\sum_{|e| \geq y} |e|^a \leq \gamma'(a,d) y^{a-d}$ *for* $0 < y < \infty$ *and* $0 < a < d$.

Proof. We fix $\lambda > 0$ and have:

$$\sum_{e \in T} |e|^a = \sum_{|e| \leq n^{-\lambda}} |e|^a + \sum_{|e| > n^{-\lambda}} |e|^a.$$

The first sum can be bounded by $n^{1-\lambda a}$, as the number of edges whose length is $\leq n^{-\lambda}$ is $n - 1 - v_{n^\lambda}$. The second sum can be rewritten as an integral, getting us to:

$$\sum_{e \in T} |e|^a \leq n^{1-\lambda a} + \int_{n^{-\lambda}}^{\sqrt{d}} x^a d(n - 1 - v_d(x)).$$

Applying integration by parts:

$$\sum_{e \in T} |e|^a \leq n^{1-\lambda a} + n^{-\lambda a} v_d(n^{-\lambda}) + a \int_{n^{-\lambda}}^{\sqrt{d}} x^{a-1} v_d(x) dx.$$

Thanks to 2.1 we can bound $v_d(x)$ with $\gamma_d x^{-d}$, so we get:

$$a\int_{n^{-\lambda}}^{\sqrt{d}} x^{a-1}v_d(x)dx \le a\gamma_d \int_{n^{-\lambda}}^{\sqrt{d}} x^{a-1-d}dx = \frac{a}{d-a}\gamma_d\left(n^{-\lambda(a-d)} - \sqrt{d}^{a-d}\right) \le \frac{a}{d-a}\gamma_d n^{-\lambda(a-d)}$$

and

$$\sum_{e\in T}|e|^a \le n^{1-\lambda a} + \gamma_d n^{-\lambda(a-d)} + \frac{a}{d-a}\gamma_d n^{-\lambda(a-d)}.$$

The result then follows by setting $\lambda = \frac{1}{d}$.

The other two result derive from similar arguments:

$$\sum_{e\in T}|e|^d \le n^{1-\lambda d} + \gamma_d n^{-\lambda(d-d)} + d\gamma_d \int_{n^{-\lambda}}^{\sqrt{d}} x^{d-1-d}dx \le n^{1-\lambda d} + d\gamma_d\left(log(\sqrt{d}) + \lambda logn\right)$$

and $\lambda = \frac{1}{d}$,

$$\sum_{|e|>y}|e|^a \le \int_y^{\sqrt{d}} x^a d(n-1-v_d(x)) = y^a v_d(y) + a\int_y^{\sqrt{d}} x^{a-1}v_d(x)dx \le \gamma_d y^{a-d} + \frac{a}{d-a}\gamma_d y^{a-d}.$$

□

In the next proposition we will provide some bounds on $M(x_1,\dots,\hat{x}_i,\dots x_n)$ given $M(x_1,\dots x_n)$ and viceversa, where $M(x_1,\dots,\hat{x}_i,\dots x_n)$ is the minimum spanning tree of the graph made from the points $\{x_j\}_{j=1\dots n}$ except x_i.

Proposition 2.3. For any edge weight function ψ we have:

$$M(x_1,\dots,x_n) \le M(x_1,\dots\hat{x}_i,\dots x_n) + \min_{j\neq i}\psi(|x_i - x_j|).$$

Furthermore if ψ is not decreasing we have:

$$M(x_1,\dots\hat{x}_i,\dots x_n) \le M(x_1,\dots,x_n) + \sum_{j\in N(i)}\psi(2|x_i - x_j|)$$

where $N(i)$ is the set of all nodes adjacent to x_i in $M(x_1,\dots,x_n)$.

Proof. The first inequality comes from the fact that any spanning tree of $\{x_1,\dots\hat{x}_i\dots,x_n\}$ can be completed as a spanning tree on $\{x_1,\dots,x_n\}$ by simply adding an edge that connects any x_j to x_i.

For the second inequality we consider x', an element of $N(i)$ that satisfies:

$$|x' - x_i| = \min_{x_j \in N(i)} |x_j - x_i|$$

Then we build a connected graph spanning $\{x_1, \ldots, \hat{x}_i, \ldots x_n\}$ from T, a spanning tree of $\{x_1, \ldots x_n\}$: first we consider the edges of T, then we delete every edge of T that connects x_i to its adjacent nodes, finally for any $x_j \in N(i)$ apart from x', we add the edge that connects x_j to x'.

Clearly we have:

$$M(x_1, \ldots, \hat{x}_i, \ldots x_n) \leq M(x_1, \ldots x_n) - \sum_{j \in N(i)} \psi(|x_j - x_i|) + \sum_{j \in N(i)} \psi(|x' - x_j|).$$

Now using the triangle inequality and the fact that x' minimizes the distance from x_i:

$$|x' - x_j| \leq |x' - x_i| + |x_i - x_j| \leq 2|x_i - x_j|$$

for any $x_j \in N(i)$.

Since ψ is nondecreasing

$$\psi\left(|x' - x_j|\right) \leq \psi\left(2|x_i - x_j|\right)$$

and

$$\psi\left(|x_i - x_j|\right) \leq \psi\left(2|x_i - x_j|\right).$$

Providing:

$$\begin{aligned} M(x_1, \ldots \hat{x}_i, \ldots x_n) &\leq M(x_1, \ldots x_n) - \sum_{j \in N(i)} \psi(|x_j - x_i|) + \sum_{j \in N(i)} \psi(|x' - x_j| \\ &\leq M(x_1, \ldots, x_n) + \sum_{j \in N(i)} \psi(2|x_i - x_j|). \end{aligned}$$

□

The next lemma provides a bound on the degree of any node in the minimum spanning tree.

Lemma 2.4. If ψ is strictly increasing, for every MST of n points in $\mathbb{R}^d$ there exist a constant D_d, dependent only on the dimension d, such that the maximum degree of the minimum spanning tree bounded by D_d.

Proof. If $\psi(x) = x$, then the degree is limited by N_d, the number of spherical caps with angle 60° needed to cover a sphere in $\mathbb{R}^d$. If there was a node x with degree over N_d then there would be two points, y and z, connected to x whose edges form an angle less than 60°. But that would be absurd as it would hold

$$|x-y|+|x-z| > \min\{|x-y|,|x-z|\}+|y-z|$$

meaning that there would be a spanning tree of cost strictly less than the MST.

For a general ψ we recall the Kruskal algorithm for minimum spanning trees. Since the algorithm orders the edges according to their weight and then adds them one by one, and ψ is strictly increasing then the ordering would be the same as the Euclidean case, meaning that also the resulting MST would be the same. Then the lemma is true for every ψ with $D_d = N_d$. □

We can now prove a bound on the difference of the lengths of minimum spanning trees of two sets of points, that depends only on how many points they have in common.

Theorem 2.5. *Given the weight function $\psi = |x|^a$ with $0 < a < d$, for every two finite subsets of $[0,1]^d$, χ and χ', there exists a constant $\gamma''(a,d)$ such that*

$$\left|M(\chi) - M(\chi')\right| \leq \gamma''(a,d)\left|\chi \triangle \chi'\right|^{\frac{d-a}{d}} \tag{2.2}$$

where $\chi \triangle \chi' = (\chi \cup \chi') \setminus (\chi \cap \chi')$.

Proof. Using the second part of proposition 2.3 repeatedly from $M(\chi \cup \chi')$, considering every node in $\chi \setminus \chi'$ we get:

$$M(\chi') \leq M(\chi \cup \chi') + \sum_{x' \in \chi - \chi'} \sum_{x \in N(x')} \psi(2|x'-x|)$$

Now, let $E' = \{(x,x') \,|\, x \in N(x'), x' \in \chi \setminus \chi'\}$, and V' the set of the nodes incident to any edge of E'. Thanks to lemma 4 we know that $|V'| < (1+D_d)|\chi \setminus \chi'|$.

If e belongs to E' then it also belongs to $MST\{x_1,..,x_n\}$ and as a consequence of the cut property **a**), there exists $A, A^c \in x_1,\ldots,x_n$ such that:

$$|e| = \min\{e \in E' : \{e\} \cap A \cap V' \neq \emptyset, \{e\} \cap A^c \cap V' \neq \emptyset\}$$

which means that e must also be an edge of $MST(V')$ leading to:

$$\sum_{x' \in \chi - \chi'} \sum_{x \in N(x')} \psi(2|x'-x|) \leq \sum_{e \in MST(V')} \psi(2|e|).$$

We can then bound this equation with Theorem 2.2 a) together with the bound we have for $|V'|$ and write:

$$M(\chi') \leq M(\chi \cup \chi') + 2^a \gamma'(a,d)\,(1+D_d)^{\frac{d-a}{d}} \left|\chi \setminus \chi'\right|^{\frac{d-a}{d}}.$$

We now bound $M(\chi \cup \chi')$ using $M(\chi)$. If we consider also $M(\chi' \setminus \chi)$ we can build a spanning tree for $\chi \cup \chi'$ by joining the MST of $\chi \cup \chi'$ and $\chi' \setminus \chi$ with an edge. Therefore we have the inequality:

$$M(\chi \cup \chi') \leq M(\chi) + M(\chi \cup \chi') + \min_{x \in \chi, x' \in \chi'} \psi(|x - x'|)$$

We again use Theorem 2.2 a) to bound $M(\chi' \setminus \chi)$ with $\gamma'(a,d)\left|\chi' \setminus \chi\right|^{\frac{d-a}{d}}$ and also bound $\min_{x \in \chi, x' \in \chi'} \psi(|x - x'|)$ with $d^{\frac{a}{2}}$

Combining the inequalities we finally get:

$$M(\chi') - M(\chi) \leq \gamma'(a,d)\left(1 + 2^a\,(1+D_d)^{\frac{d-a}{d}}\right)\left|\chi' \setminus \chi\right|^{\frac{d-a}{d}} + d^{\frac{a}{2}}$$

and by symmetry we also have

$$M(\chi) - M(\chi') \leq \gamma'(a,d)\left(1 + 2^a\,(1+D_d)^{\frac{d-a}{d}}\right)\left|\chi \setminus \chi'\right|^{\frac{d-a}{d}} + d^{\frac{a}{2}}.$$

Since both $|\chi \setminus \chi'|$ and $|\chi' \setminus \chi|$ are less or equal than $|\chi \triangle \chi'|$ and we can assume that $|\chi \triangle \chi'| \geq 1$ (otherwise the result would be trivial), the thesis is verified for

$$\gamma''(a,d) = \gamma'(a,d)\left(1 + 2^a\,(1+D_d)^{\frac{d-a}{d}}\right) + d^{\frac{a}{2}}.$$

□

Chapter 3

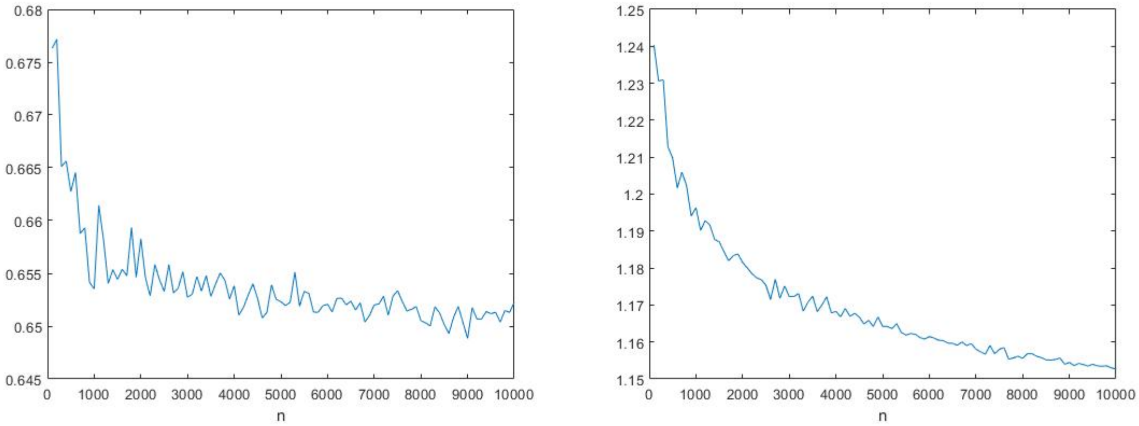
Asymptotic growth in mean

In this chapter we study the random Euclidean minimum spanning tree problem where the points are uniform and independent random variables. The main results we are presenting are given by the asymptotic behaviour of the mean weight of the minimum spanning tree as the number of points tends to infinity.

We show two proofs as presented in [5] and [6], and in [1] that work with slightly different hypothesis. The first proof considers the points to be in the unitary cube of $\mathbb{R}^d$ and exploits a standard technique of Poissonization and then de-Poissonization of the number of points. The second one requires the points to be in a toroidal space, in order to avoid possible changes of distribution near the boundary of the cube. The proof revolves around extracting the right properties from the model and then applying a general result through those properties. Indeed it is known, as shown in [4], that in both models the weighted mean of the cost converges to the same constant.

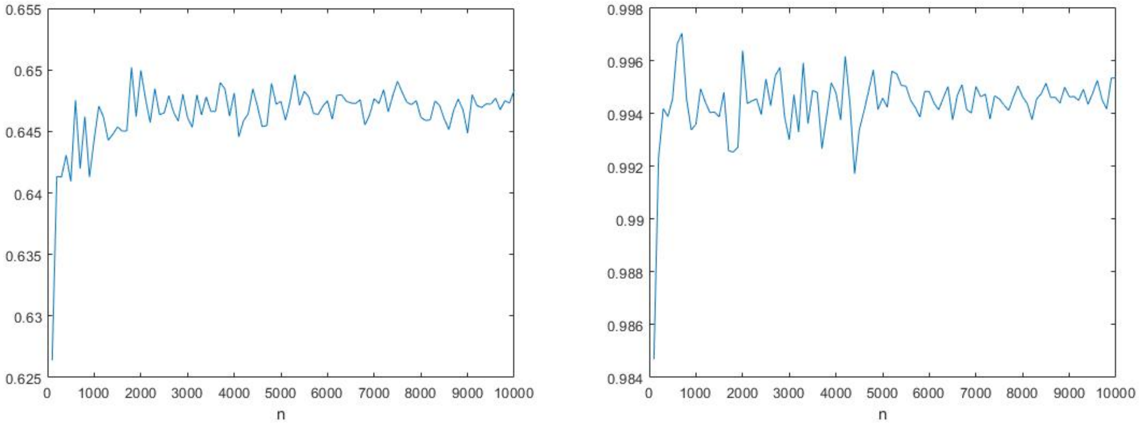
Both results are supported by experimental evidence as shown in the following charts, where we used Prim's algorithm to calculate the minimum spanning tree of n random uniform points in $\mathbb{R}^d$, and plot its cost divided by $n^{1-\frac{1}{d}}$.

Figure 3.1: $\frac{M(x_1,\dots,x_n)}{n^{1-\frac{1}{d}}}$ in $[0,1]^d$



d=2 d=13

Figure 3.2: $\frac{M(x_1,\dots,x_n)}{n^{1-\frac{1}{d}}}$ in the toroidal model



d=2 d=13

3.1 Cube model

In order to study the properties of the minimum spanning tree functional when the points are $(X_i)_1^n$, independent uniform random variables on $[0,1]^d$, we first consider a Poissonized version of the problem.

Definition 3.1. We consider N, a Poisson random variable with parameter t^d and set of random points $(X_i)_{i=0}^{\infty}$ with uniform distribution in $[0,t]^d$. We also suppose M and $(X_i)_{i=0}^{\infty}$

to be independent. We define the Poisson point process on $[0,t]^d$ with parameter t^d to be:

$$\mu : \Omega \to \mathcal{M}([0,t]^d)$$
$$\omega \to \mu^\omega = \sum_{i=1}^{M(\omega)} \delta_{X_i(\omega)}$$

where, Ω is a probability space where the variables are defined and $\mathcal{M}([0,t]^d)$ is the set of the measures of $[0,t]^d$.

We also define, given $A \subset [0,t]^d$ a Borel subset, $\pi^\omega(A) = \{X_i(\omega) \in A : i < M(\omega)\}$ and $\mu^\omega(A) = \#\pi^\omega(A)$.

Lemma 3.1. Let $A_1, A_2, \ldots, A_n$ denote a partition of $[0,t]^d$ where every A_i is a Borel set. Let also be $(X_i)_{i=1}^m$ random independent points in $[0,t]^d$ and $k_1, \ldots, k_n \in \mathbb{N}$ such that $k_1 + \ldots + k_n = m$. Then

$$\mathbb{P}\left(k_1 \text{ points are in } A_1, \ldots, k_n \text{ points are in} A_n\right) = \frac{m!}{k_1! \ldots k_n!} \frac{1}{t^{md}} \prod_{i=1}^n |A_i|^{k_i}$$

where $|A_i|$ is the Lebesgue measure of A_i.

Proof. We know, since the points are uniform on $[0,t]^d$ and independent that:

$$\mathbb{P}\left(X_1 \in A_1, \ldots, X_{k_1} \in A_1, X_{k_1+1} \in A_2, \ldots, X_{k_1+\ldots+k_{n-1}+1} \in A_{k_n}, \ldots, A_{k_n} \in A_n\right) = \frac{1}{t^{md}} \prod_{i=1}^n |A_i|^{k_i}$$

Since the probability is not affected by the order we choose the points, we just need to consider the number of ways which we can choose subsets of $k_1, \ldots, k_n$ elements from a set of m elements, which is

$$\frac{m!}{k_1! \ldots k_n!}$$

proving the thesis. □

Thus we get that, for any A, Borel subset of $[0,t]^d$:

$$\mathbb{P}(\mu(A) = k) = \sum_{m=k}^{\infty} \mathbb{P}\left(\mu(A) = k, \mu(A^c) = m - k \middle| N = m\right) \mathbb{P}(N = m) =$$
$$\frac{|A|^k}{k!} \sum_{m=k}^{\infty} \frac{|A^c|^{m-k}}{(m-k)!} e^{-t^d} = \frac{|A|^k}{k!} e^{-(t^d - |A^c|)} = \frac{|A|^k}{k!} e^{-|A|} \quad (3.1)$$

Since for any bounded Borel set A, the set $\pi^\omega(A)$ is almost surely a finite set of points, we can define the quantity $M(\pi^\omega(A))$ as the minimum spanning tree of the points of $\pi^\omega(A)$.

Now we define $\phi(t) = E(M(\pi([0,t]^d))$, that is the expected value of the cost of the minimum spanning tree on the Poisson point process.

We now prove that $\phi(t)$ is continuous.

First of all we write $\phi(t)$ as:

$$\phi(t) = \sum_{n=0}^{\infty} E\left[M(\pi([0,t]^d)\middle|N=n\right]\frac{t^{nd}}{n!}e^{-t^d}$$

Since $t^a M(x_1,\dots,x_n) = M(tx_1,\dots,tx_n)$, and by setting $m_n = E\left[M(X_1,\dots,X_n)\right]$, where $X_1,\dots,X_n$ are uniform independent points on $[0,1]^d$ we get:

$$\phi(t) = t^a \sum_{n=0}^{\infty} m_n \frac{t^{nd}}{n!}e^{-t^d} \tag{3.2}$$

Since m_n is bounded by $\sqrt{d}n$ ($\sqrt{d}$ is the maximum possible edge in $[0,1]^d$), we can derive $\phi(t)$'s continuity from the dominated convergence theorem, as, every term of the sum is a composition of continuous functions, and if $t \to t_0$ the n-th term of the sum is, when t is sufficiently close to t_0, bounded by $\sqrt{d}\frac{(t_0+1)^{nd}}{(n-1)!}e^{-(t_0+1)^d}$, whose sum over n is $\sqrt{d}(t_0+1)^{a+d}$.

□

The next step is to prove convergence of $\frac{\phi(t)}{t^d}$ as $t \to \infty$. In order to do that we first need to find a relationship that ties $\phi(t)$ and $\phi(t')$ when $t' > t$. The following lemma provides an almost linear one, with a remainder term that will turn out to be manageable.

Lemma 3.2. If $\psi(x) = x^a$, with $0 < a < d$, then, for every $0 < t < \infty$ and $m \in \mathbb{N}$, $m \geq 1$

$$\phi(t) \leq m^d \phi\left(\frac{m}{t}\right) + Ct^a m^{d-a} \tag{3.3}$$

where C is a constant only dependent on a and d.

Proof. Firstly we divide $[0,t]^d$ in m^d cubes Q_i everyone with length $\frac{t}{m}$. Then from every Q_i we choose a point Y_i. Now we can build a spanning tree of $\pi([0,t]^d)$ by considering the MST of every cube Q_i and then connect all the graphs through the edges of $MST(Y_1,\dots,Y_{m^d})$. Knowing from lemma 2, adjusted for $[0,t]^d$, that

$$M(Y_1,\dots,Y_{m^d}) \leq t^a \gamma'(a,d) m^{d-a},$$

we get

$$M(\pi[0,t]^d) \leq \sum_{i=1}^{m^d} M(\pi(Q_i)) + M(Y_1,\dots,Y_{m^d}) \leq \sum_{i=1}^{m^d} M(\pi(Q_i)) + t^a\gamma'(a,d)m^{d-a}. \tag{3.4}$$

From 3.1 and 3.2 we also have $EM(\pi(Q_i)) = \phi\left(\frac{t}{m}\right)$, thus we finally get:

$$\phi(t) \leq m^d\phi\left(\frac{t}{m}\right) + t^a\gamma'(a,d)m^{d-a}.$$

□

We can now prove convergence for $\phi(t)$:

Theorem 3.3. *There exists a constant $c(a,d)$ such that*

$$\lim_{t\to\infty}\frac{\phi(t)}{t^d}c(a,d).$$

Proof. We set, for the same C as Lemma 3.2, $\tilde{\phi}(t) = \phi(t) + 2Ct^a$. Then right from 3.3 we have:

$$\tilde{\phi}(mt) = \phi(mt) + 2C(mt)^a \leq m^d\tilde{\phi}(t) + Ct^am^d + 2Ct^am^a.$$

Thus there exists a m_0 such that for every $m \geq m_0$ we have $2m^a < m^d$ and get:

$$\tilde{\phi}(mt) \leq m^d\tilde{\phi}(t). \tag{3.5}$$

Let $\beta = \liminf_{t\to\infty}\frac{\tilde{\phi}}{t^d}$. For every ϵ there exists a t_0 such that:

$$\frac{\tilde{\phi}(t_0)}{t^d} \leq \beta + \epsilon.$$

Since $\phi(t)$ is continuous $\tilde{\phi}(t)$ is also continuous, meaning that there is an interval $[t_0, t_0 + \delta]$ where $\tilde{\phi}(t)/t^d \leq \beta + \epsilon$. As a consequence of 3.5 we have that the same inequality is also true for $\bigcup_{m=m_0}^{\infty}[mt_0, m(t_0+\delta)]$.
Since if $m > \frac{t_0}{\delta}$ the intervals overlap, we have for $t' = \max\{m_0, \frac{t_0}{\delta}\}$

$$[t', \infty) \subset \bigcup_{m=m_0}^{\infty}[mt, m(t+\delta)].$$

Hence

$$\limsup_{t\to\infty}\frac{\tilde{\phi}}{t^d} \leq \beta + \epsilon.$$

Since this holds for every ϵ and $a < d$ then $\phi(t) \sim \beta t^d$.

□

We finally extract the asymptotic behaviour of m_n from $\phi(t)$.

Theorem 3.4. *There exists a constant $\beta_M(a,d)$ such that:*

$$\lim_{n\to\infty}\frac{m_n}{n^{\frac{d-a}{d}}}=\beta_M(a,d)$$

Proof. We consider $\tilde{N}$ a Poisson variable of mean n and set:

$$E[m_{\tilde{N}}]=\phi(n^{\frac{1}{d}})=\sum_{k=0}^{\infty}m_k\frac{n^k}{k!}e^{-k}.$$

Then by taking expectations from 2.2 we can bound:

$$|m_n-E[m_{\tilde{N}}]|\leq\sum_{k=0}^{\infty}|m_n-m_k|\frac{n^k}{k!}e^{-k}\leq\gamma''(a,d)\sum_{k=0}^{\infty}|n-k|^{\frac{d-a}{d}}\frac{n^k}{k!}e^{-k}.$$

We apply $||X||_1\leq||X||_p$, that holds for every random variable and $p\geq1$, with $p=\frac{2d}{d-a}$:

$$|m_n-E[m_{\tilde{N}}]|\leq\beta''(a,d)\left(\sum_{k=0}^{\infty}|n-k|^2\frac{n^k}{k!}e^{-k}\right)^{\frac{d-a}{2d}}$$

But we just got the expression of $Var(\tilde{N})$ that is equal to n, thus:

$$|m_n-E[m_{\tilde{N}}]|\leq\beta''(a,d)n^{\frac{d-a}{2d}}.$$

We have just proven $|a_n-E[a_{\tilde{N}}]|=O(n^{\frac{d-a}{2d}})$ and, from Theorem 3.3, we have

$$n^{\frac{a}{d}}E[m_{\tilde{N}}]=\phi(n^{\frac{1}{d}})\sim c(a,d)n \quad \text{as} \quad n\to\infty$$

leading to:

$$\beta_M(a,d):=\lim_{n\to\infty}\frac{m_n}{n^{\frac{d-a}{d}}}=c(a,d).$$

□

3.2 Toroidal model

Now we are going to present another point of view, where we identify some essential properties of the model that allow us to prove a convergence result.

Let $G_n = (V, E, W)$ be a graph such that $V = \{v_1, \dots, v_n\}$ and $d_{i,j}$, the weight of the edge that connects v_i and v_j, is a random variable defined on some probability space for any $i, j = 1, \dots, n$. We denote by $G_n(z)$ the sub-graph of G_n made by all the edges (v_i, v_j) such that $d_{i,j} \le z$.

Then for c_d, the volume of the unitary ball in $\mathbb{R}^d$, we assume for any the following properties:

I for every $k, s = 1, \dots, n$ there exists a permutation π such that $\pi(k) = s$ and the matrix $\left(d_{i,j}\right)_{i,j=1}^n$ has the same distribution as $\left(d_{\pi(i),\pi(j)}\right)_{i,j=1}^n$. We will refer to this as the exchangeability property.

Thanks to property **I** it's well defined for any $k \in \mathbb{N}$, $k \neq 0$:

$$P_{k,n}(z) = \mathbb{P}\left[\text{a given point belongs to a component of } G_n(z) \text{ having exactly } k \text{ elements}\right]$$

and we can state the remaining properties.

II For any fixed k there exists a function $f_k(y)$ such that:

$$\lim_{n \to 0} P_{k,n}\left[\left(\frac{y}{nc_d}\right)^{\frac{1}{d}}\right] = f_k(y)$$

III For any $k < n$, there exists $l_k(y)$ such that

$$P_{k,n}\left[\left(\frac{y}{nc_d}\right)^{\frac{1}{d}}\right] \le l_k(y)$$

and

$$\int_0^\infty l_k(y) y^{\frac{1}{d}-1} dy \le \infty$$

IV For all K and n,

$$\left| n^{\frac{1}{d}} \int_0^\infty \left(\sum_{k=K}^{n} \frac{1}{k} P_{k,n}(z) - \frac{1}{n} \right) dz \right| = O\left(K^{1-\frac{1}{d}}\right),$$

the implicit constant being independent of n.

We note that the properties shift the focus of the problem from the vertices towards the edges, as originally in [1] they were developed as a way to unify the Euclidean model and the independent model, where it is the edges that are random independent variables.

We now prove, given the properties, a result of convergence:

Theorem 3.5. *Let $\{G_n\}_{n=1}^{\infty}$ be a sequence of graphs of a model that satisfies the properties **I**, **II**, **III**, **IV** and let $\{T_n\}_{n=1}^{\infty}$ be the sequence of the lengths of their minimum spanning trees. Then:*

$$\beta(d) = \lim_{n\to\infty} \frac{E[T_n]}{n^{\frac{d-1}{d}}} = \frac{1}{d(c_d)^{1/d}} \sum_{k=1}^{\infty} \frac{1}{k} \int_0^{\infty} f_k(y) y^{\frac{1}{d}-1} dy$$

Proof. Let $C_n(z)$ the number of components of the graph formed by the edges of length less or equal than z. We first show that:

$$T_n = \int_0^{\infty} [C_n(z) - 1]\, dz.$$

To prove it we consider $0 = z_0 < z_1 < z_2 \ldots < z_{n-1}$, where z_j is the smallest distance such that $C_n(z_j) \geq n - j$.

Then:

$$\int_0^{\infty} [C_n(z) - 1]\, dz = \int_0^{z_n} [C_n(z) - 1]\, dz = \sum_{j=1}^{n-1} \int_{z_{j-1}}^{z_j} [C_n(z) - 1]\, dz$$
$$= \sum_{j=1}^{n-1} (n - j)(z_j - z_{j-1}) = \sum_{j=1}^{n-1} z_j.$$

Now since during the j-th step of Prim's algorithm an edge of length exactly z_j is added to the minimum spanning tree, then $T_n = \sum_{j=1}^{n-1} z_j$ and the equality is proven.

Consequently, as $C_n(z) - 1 \geq 0$ we can use Fubini-Tonelli's theorem to get:

$$E[T_n] = \int_0^{\infty} E[C_n(z) - 1] dz. \tag{3.6}$$

Then we rewrite $C_n(z)$ as:

$$C_n(z) = \sum_{k=1}^{n} \sum_{i=1}^{n} \frac{X_{i,k}(z)}{k}$$

where $X_{i,k}(z)$ is a random variable such that:

$$X_{i,k} = \begin{cases} 1 & \text{if the node } i \text{ belongs to a component of } G_n(z) \text{ of } k \text{ nodes} \\ 0 & \text{otherwise.} \end{cases}$$

Taking the expectation, and using the exchangeability property **I**:

$$E\left[C_n(z)\right] = n \sum_{k=1}^{n} \frac{P_{k,n}(z)}{k}.$$

Then we can substitute this in equation 3.6 and divide by $n^{\frac{d-1}{d}}$:

$$\frac{E[T_n]}{n^{\frac{d-1}{d}}} = n^{\frac{1}{d}} \int_0^\infty \left(\sum_{k=1}^{n} \frac{P_{k,n}(z)}{k} - \frac{1}{n} \right) dz.$$

Then we can take the limit and for any given K we can split the sum:

$$\beta(d) = \lim_{n\to\infty} \frac{E[T_n]}{n^{\frac{d-1}{d}}} = \lim_{n\to\infty} n^{\frac{1}{d}} \int_0^\infty \left(\sum_{k=1}^{K-1} \frac{P_{k,n}(z)}{k} \right) dz + n^{\frac{1}{d}} \int_0^\infty \left(\sum_{k=K}^{n} \frac{P_{k,n}(z)}{k} - \frac{1}{n} \right) dz.$$

Now we apply property **IV**:

$$\beta(d) = \lim_{n\to\infty} n^{\frac{1}{d}} \int_0^\infty \left(\sum_{k=1}^{K-1} \frac{P_{k,n}(z)}{k} \right) dz + O(K^{-1+\frac{1}{d}})$$

By linearity of the integral and taking the limit inside the sum:

$$\beta(d) = \sum_{k=1}^{K-1} \frac{1}{k} \left(\lim_{n\to\infty} \int_0^\infty P_{k,n}(z) n^{\frac{1}{d}} dz \right) + O(K^{-1+\frac{1}{d}}).$$

In each integral we apply the change of variables $z = \left(\frac{y}{nc_d}\right)^{\frac{1}{d}}$ and thus:

$$\beta(d) = \frac{1}{(dc_d)^{\frac{1}{d}}} \sum_{k=1}^{K-1} \frac{1}{k} \left(\lim_{n\to\infty} \int_0^\infty P_{k,n} \left[\left(\frac{y}{nc_d} \right)^{\frac{1}{d}} \right] y^{\frac{1}{d}-1} dy \right) + O(K^{-1+\frac{1}{d}}).$$

Thanks to property **III** we can now use dominated convergence and take the limit inside of the integral that we know converges for property **II**:

$$\beta(d) = \frac{1}{(dc_d)^{\frac{1}{d}}} \sum_{k=1}^{K-1} \frac{1}{k} \left(\int_0^\infty f_k(y) y^{\frac{1}{d}-1} dy \right) + O(K^{-1+\frac{1}{d}}).$$

Since the equality holds for every K then the thesis is proven thanks to property **IV**.

□

We now consider the Euclidean toroidal model and show that all four properties hold in this setting. The model considers a graph whose nodes are random uniform independent points in the metric space $\mathbb{T}^d$ defined as $[-\frac{1}{2}, \frac{1}{2}]^d$ with the Lebesgue measure and with the equivalence relation that states that two points are the same if their coordinates are equal mod 1. The length an edge is defined as the minimum, among every element of the class of the points, of the distances between the points.

The main reason to use this model is that, unlike the one where the points are in the cube, the distribution of the points and the edges does not change towards the boundary.

Property **I** is then verified as it is a direct consequence from the fact that the nodes are i.i.d. vectors.

Now for properties **II**, **III**, **IV** a little more work is needed:

II. for $x_0 = 0$ we define $\Theta'_k(z)$ as the subset of $\mathbb{T}^k$ made by every $\{x_1, \dots x_{k-1}\}$ such that there exists a tree that connects the points with edges of length $\leq z$, or equivalently, such that the union of the closed spheres in $\mathbb{T}$, $B'(x_j, \frac{z}{2})$, is a connected set. If k=1 we define $\Theta'_1(z)$ as the torus.

Now we consider how to compute $P_{k,n}(z)$ using $\Theta'_k(z)$. We will consider the case $k \geq 2$ and treat $k = 1$ separately. Firstly, without losing generality, we can suppose that the node taken into consideration is X_0, as they all have the same distribution. Then since $P_{k,n}(z)$ does not really depend on where X_0 falls, but only on how the other nodes around it are placed, we can always translate every node so that $X_0 = 0$, without affecting the probability. By definition $P_{k,n}(z)$ is the probability that 0 is in a component of $G_n(z)$ having exactly k nodes, namely the probability that 0 is in one of the configurations of $\Theta'_k(z)$ where the other $k-1$ can be any subset of $k-1$ elements of all the nodes apart from x_0, and the other $n-k$ nodes are not connected with any node of the configuration in G_z. Thus we can write.

$$P_{k,z} = \binom{n-1}{k-1} \int'_{\Theta'_k(z)} \left[1 - g'_{k,z}(x_1 \dots x_{k-1})\right]^{(n-1)-(k-1)} dx_1 \dots dx_{k-1}$$

where $g'_{k,z}(x_1 \dots x_{k-1})$ is the volume of $\bigcup_{j=0}^{k-1} S'(x_j, z)$, and $\int'$ refers to the integral over $\mathbb{T}^{k-1}$.

Since sets not touching the torus boundary are the same on the torus and $\mathbb{R}^d$, for $z \leq \frac{1}{2k}$ we get:

$$P_{k,n}(z) = \binom{n-1}{k-1} \int_{\Theta_k(z)} \left[1 - g_{k,z}(x_1 \dots x_{k-1})\right]^{n-k} dx_1 \dots dx_{k-1}$$

where:

- $\Theta_k(z)$ is the subset of $\mathbb{R}^{d \times (k-1)}$ made by every $\{x_1, \dots x_{k-1}\}$ such that there exists a tree that connects the points with edges of length $\leq z$
- $g_{k,z}(x_1 \dots x_{k-1})$ is the the volume of $\bigcup_{j=0}^{k-1} B(x_j, z)$
- $\int$ is the integral over $(\mathbb{R}^d)^{k-1}$.

We now apply the change of variables $u_j = \frac{x_j}{z}$, and since $g_{k,z}(zu_1 \dots zu_{k-1}) = z^d g_{k,1}(u_1 \dots u_{k-1})$ we have:

$$P_{k,n}(z) = \binom{n-1}{k-1} z^{d(k-1)} \int_{\Theta_k(1)} \left[1 - z^d g_{k,1}(u_1 \dots u_{k-1})\right]^{n-k} du_1 \dots du_{k-1}$$

Let $z = \left(\frac{y}{nc_d}\right)^{\frac{1}{d}}$. We can still assume $z \leq \frac{1}{2k}$ as we are going to take the limit $n \to \infty$.

Then:

$$P_{k,n}\left(\left(\frac{y}{nc_d}\right)^{\frac{1}{d}}\right) = \frac{y^{k-1}}{c_d^{k-1}(k-1)!} \prod_{j=1}^{k-1}\left(1 - \frac{j}{n}\right) \int_{\Theta_k(1)} \left[1 - \frac{y}{nc_d} g_{k,1}(u_1 \dots u_{k-1})\right]^{n-k} du_1 \dots du_{k-1}.$$

Now we take the limit on n, and since $|\Theta_k(1)|$ is bounded and $\left[1 - \frac{y}{nc_d} g_{k,1}(u_1 \dots u_{k-1})\right]^{n-k} \leq 1$ we can apply dominated convergence to get:

$$f_k(y) = \lim_{n\to\infty} P_{k,n}\left(\left(\frac{y}{nc_d}\right)^{\frac{1}{d}}\right) = \frac{y^{k-1}}{c_d^{k-1}(k-1)!} \int_{\Theta_k(1)} \exp\left[-\frac{y}{c_d} g_{k,1}(u_1 \dots u_{k-1})\right] du_1 \dots du_{k-1}.$$

If $k = 1$ instead, $P_{1,n}(z)$ is the same as the probability that the other $n-1$ points do not belong to $B(0,z)$, and thus:

$$f_k(y) = \lim_{n\to\infty} P_{k,n}\left(\left(\frac{y}{nc_d}\right)^{\frac{1}{d}}\right) = \left[1 - c_d \frac{y}{nc_d}\right]^{n-1} = e^{-y}.$$

□

III. Since $1 - x \leq e^{-x}$, for $k = 1$ we have:

$$P_{1,n}\left(\left(\frac{y}{nc_d}\right)^{\frac{1}{d}}\right) = \left[1 - c_d \frac{y}{nc_d}\right]^{n-1} \leq e^{-y\frac{(n-1)}{n}} \leq e^{-\frac{y}{2}}.$$

If $k \geq 2$, for $z \geq \frac{\sqrt{d}}{2}$, $P_{k,n}(z) = 0$. So we can assume $z < \frac{\sqrt{d}}{2}$.

Since $g'_{k,z}(x_1, \dots x_{k-1}) \leq c_d z^d$ we have:

$$\begin{aligned}
P_{k,z} =& \binom{n-1}{k-1} \int'_{\Theta'_k(z)} \left[1 - g'_{k,z}(x_1 \dots x_{k-1})\right]^{(n-1)-(k-1)} dx_1 \dots dx_{k-1} \\
\leq& \binom{n-1}{k-1} \int'_{\Theta'_k(z)} \exp\left[-(n-k) g'_{k,z}(x_1 \dots x_{k-1})\right] dx_1 \dots dx_{k-1} \\
\leq& \frac{n^{k-1}}{(k-1)} e^{-(n-k)c_d z^d} \int'_{\Theta'_k(z)} dx_1 \dots dx_{k-1}.
\end{aligned}$$

Now we bound $|\Theta'_k(z)|$ with $k^{k-2}(c_d z)^d$, as we note that given x_i one can choose x_{i+1} only in the ball of radius z centred in x_i, and those nodes can be connected in at most k^{k-2} trees.

We also use $\frac{n}{n-k} \geq \frac{1}{k+1}$, and get:

$$P_{k,z} \leq \frac{n^{k-1}}{(k-1!)} e^{-\frac{n}{k+1} c_d z^d} k^{k-2} (c_d z^d)^{k-1}.$$

And for $z = \left(\frac{y}{nc_d}\right)^{\frac{1}{d}}$:

$$P_{k,n}\left(\left(\frac{y}{nc_d}\right)^{\frac{1}{d}}\right) \leq l_k(y) = \frac{y^{k-1}}{(k-1)!} k^{k-2} e^{-\frac{y}{k+1}}$$

and since $\int_0^\infty l_k(y) y^{\frac{1}{d}-1} dy < \infty$ property **III** is also verified.

□

IV. Let $C_{K,n}(z)$ be the number of components, formed by K nodes or more, of $G_n(z)$. Then for the same argument used in Theorem 3.5 we have:

$$E\left[C_{n,K}\right] = n \sum_{k=K}^{n} \frac{P_{k,n}(z)}{k}$$

Therefore, knowing that there a constant M that bounds the length of the edges, we can write:

$$\begin{aligned}
n^{\frac{1}{d}} \int_0^\infty \left[\sum_{k=K}^{n} \frac{P_{k,n}(z)}{k} - \frac{1}{n}\right] dz =& \frac{1}{n^{1-\frac{1}{d}}} \int_0^M \left(E\left[C_{K,n}(z)\right] - 1\right) dz \\
\geq& -\frac{M}{n^{1-\frac{1}{d}}} \geq -\frac{M}{K^{1-\frac{1}{d}}}.
\end{aligned} \tag{3.7}$$

We call J the set of the z_j, where $j \geq 1$ and $z_j \leq z_{j+1}$, such that there is an edge of such length that once added increases or decreases the value of $C_{K,n}(z)$. Thus we have:

$$\int_0^M \left(C_{K,n}(z)-1\right)dz = \sum_{j=1}^{\#J-1}(C_{K,n}(z_j)-1)(z_{j+1}-z_j)+z_1C_{K,n}(0)$$

and by applying summation by parts with $z_0 = 0$ we get

$$\int_0^M \left(C_{K,n}(z)-1\right)dz = \sum_{z_i\in J_-} z_i^- - \sum_{z_i\in J_+} z_i^+ \leq \sum_{z_i\in J_-} z_i^- .$$

Where $J^+ \subset J$ is the set of the lengths of the edges that, when added, increase $C_{K,n}$, and J^- is the same for the edges that make $C_{K,n}$ decrease. We notice that the edges that the elements of J^- represent form a tree on clusters of points. Since every time an edge of J^- is added $C_{n,K}(z)$ decreases it means that there are two connected components made by at least K nodes that were previously not connected. This means that those edges connects clusters of points that were not previously connected, forming a tree on the multigraph generated by the clusters as vertices and keeping the edges of the graph as edges of the multigraph. Now from a direct consequence of Theorem 2.2 we have that also trees on multigraphs of r elements in $\mathbb{R}^d$ have length bounded by $k_d r^{\frac{d-1}{d}}$. Knowing that $|J^-| \leq \frac{n}{K}$ we get

$$n^{\frac{1}{d}}\int_0^\infty \left[\sum_{k=K}^{n}\frac{P_{k,n}(z)}{k}-\frac{1}{n}\right]dz \leq \frac{1}{n^{1-\frac{1}{d}}}k_d\left(\frac{n}{K}\right)^{\frac{d-1}{d}} = \frac{k_d}{K^{1-\frac{1}{d}}}. \tag{3.8}$$

Thus equation 3.7 and 3.8 give us the thesis. □

Combining the general theorem with what we have just developed about the Euclidean MST:

Theorem 3.6. *In the Euclidean toroidal model:*

$$\beta_T(d) := \lim_{n\to\infty}\frac{E\left[T_n\right]}{n^{\frac{d-1}{d}}} = \frac{\Gamma\left(\frac{1}{d}\right)}{d(c_d)^{\frac{1}{d}}}+\sum_{k=2}^{\infty}\frac{\Gamma\left(k+\frac{1}{d}-1\right)}{d(k!)}\int_{\Theta_k(1)} g_{k,1}(u_1,\dots u_{k-1})^{-(k+\frac{1}{d}-1)}du_1\dots du_{k-1}$$

where

- $\Theta_k(z)$ *is the subset of* $\mathbb{R}^{d\times(k-1)}$ *made by every points* $\{x_1,\dots x_{k-1}\}$*, such that there exists a tree that connects the points with edges of length* ≤ 1
- $g_{k,1}(x_1\dots x_{k-1})$ *is the the volume of* $\bigcup_{j=0}^{k-1}B(x_j,1)$.

In order to prove the properties when the weight function is $\psi(x) = x^a$ one can just repeat the same steps as the Euclidean case with minor adjustments.

Chapter 4

The bipartite case

The main goal of this section is to discuss whether the same techniques used for the random Euclidean MST can be adapted to the bipartite case. We recall that a graph is bipartite if the vertices can be partitioned in two different subsets, the red ones and the blue ones, such that the only edges in the graph are those who connects a node of one subset to a node of the other one. We will consider then the problem of finding some result of mean convergence in the case where there is an even number of random uniform points $2n$, where n are red and the others are blue. To stress the difference between the two problems we will also refer to the problem on the complete graph as the monopartite case.

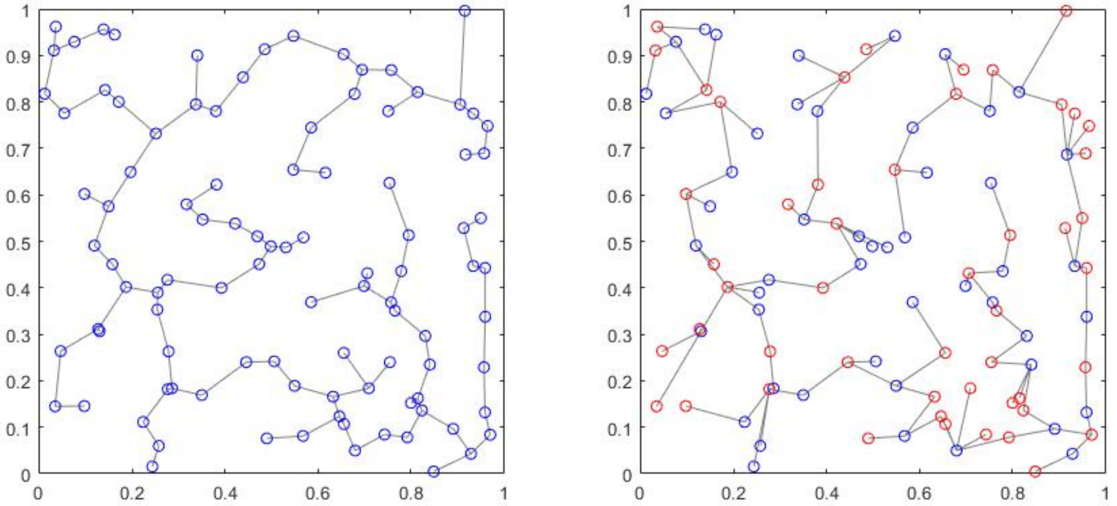


Figure 4.1: A Euclidean monopartite and bipartite minimum spanning tree of the same set of points in the plane

4.1 An algorithm

In this section we present an algorithm specific for the Euclidean bipartite graph. First a definition:

Definition 4.1 (Voronoi diagram)**.** Let S denote a set of n points (called sites) in $[0,1]^d$. Given $p \in S$, the set of points of $[0,1]^d$ for which p is the closest site is called the Voronoi region of p. The partition that the regions generate, $V(S)$, is called the Voronoi diagram of S.

The Voronoi diagram is known to be useful in many problems of computational geometry, as it can be calculated with relative ease and has a lot of interesting properties. As far as the Euclidean bipartite minimum spanning tree is concerned it is possible to use the Voronoi diagram to start building it.

Proposition 4.1. Given a bipartite graph $G = (R \cup B, E)$, and a red point $r \in G$, let $Vo(R)$ be the Voronoi diagram built using the red points as sites. Then if b is a blue point that is in the Voronoi region of r the edge (r,b) is in the minimum spanning tree of G.

Proof. If this wasn't true, one could try to add (r,b) to the tree, generating a cycle where there is an edge of length greater than the one (r,b), as b must be connected to at least another red point outside of the Voronoi region of r. □

The strategy to build the tree is then:

- generate the Voronoi diagram of the red points
- connect every blue node to the site of the Voronoi region they are in
- find the minimum spanning tree of the multigraph generated by the connected components built so far.

This strategy is simple enough in dimension 1 and has been used as a theoretical tool to prove the asymptotics of the mean cost of the MST as shown by [2].

The idea is, after connecting the red nodes to the site of their Voronoi region, to connect them by paying attention on the ticks that divides the region. For each tick, we select its closest red point, that could end up on its right or left, and connect it to the first blue point on the opposite side of the tick. This method generates exactly the minimum possible edge for every couple of consecutive nodes of the multigraph and since the resulting graph is connected, it generates its minimum spanning tree.

However, this strategy seems hard to be exploited in higher dimensions as one site may share an edge with any number of sites.

4.2 Properties

In this section we discuss the properties of the Euclidean bipartite MST, in particular making a parallelism with what we showed in the monopartite case.

Weight

A first major difference from the monopartite MST is that there is not a bound on its cost that is true in every instance of the problem as the one in Theorem 2.2. Indeed, the only true bound that depends only on the number of points, depends linearly from n as the points can be split according to their color and be on opposite corners of a square. As a consequence even Theorem 2.5 is not verified in the bipartite case.

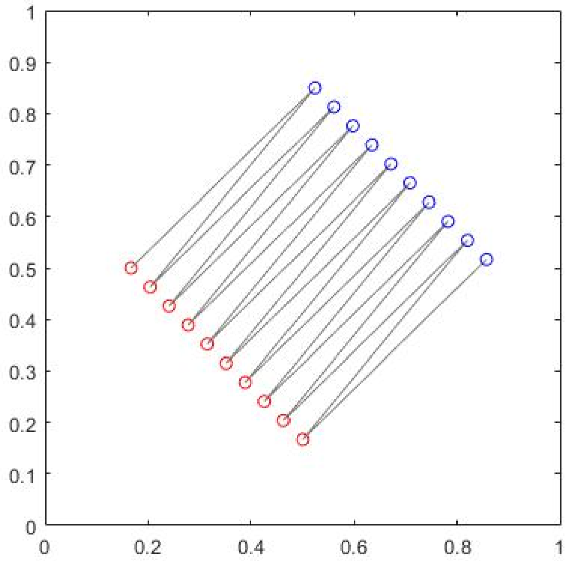


Figure 4.2: Example of bipartite MST whose cost scales linearly on the number of nodes

Even so, we can still prove a bound for m_n^B, the mean cost when the points are n independent uniform random variables on $[0,1]^d$, that points to the right scaling of the cost.

Lemma 4.2. Given a set of $2n$ points in $\mathbb{R}^d$ that form a bipartite graph made by $r_1,\ldots,r_n$ red points and $b_1,\ldots,b_n$ blue points call $M^B(r_1,\ldots,r_n,b_1,\ldots,b_n)$ the cost of the minimum spanning tree on the graph. Then

$$M(r_1,\ldots,r_n,b_1,..b_n) \leq M^B(r_1,\ldots,r_n,b_1,..b_n)$$

$$M^B(r_1,\dots,r_n,b_1,..b_n) \leq \sum_{i=1}^{n} \min_{j=1,\dots,n} |r_i - b_j| + 2\sum_{j=1}^{n} \min_{i=1,\dots,n} |b_j - r_i| + M(b_1,\dots b_n)$$

where $M(\cdot)$ is the cost of the monopartite minimum spanning tree on the points.

Proof. The first equation is consequence of the fact that the bipartite graph is a subgraph of the monopartite one.

We now prove the second equation. If we consider the Voronoi diagram made by the blue points as sites, we know that every edge connecting any red points to the site of the Voronoi region it is in, is an edge of the minimum spanning tree 4.1.

We now consider the monopartite minimum spanning tree on the blue points. We have built a spanning tree on the regular Euclidean case. From this we want to build a connected graph for the bipartite case (it will not be necessarily a tree).

We pick a blue node b and we know that we can visit every node of the blue MST. We consider a depth first search visit of the graph, where we explore as far as possible along the paths before backtracking, that allows us to associate to every other blue node a unique edge of the blue MST, the edge that connects it from its father, if we consider b to be the root of the tree.

Now we bound every edge by using its associated node: let (b_h, b_k) an edge of the tree. We remove and replace it with (b_h, r_{s_k}) and (r_{s_k}, b_k) where r_{s_k} is the closest red node to b_k. By repeating this process throughout every edge, we have a connected graph that is allowed in the bipartite problem whose weight is bounded by:

$$\sum_{i=1}^{n} \min_{j=1,\dots,n} |r_i - b_j| + \sum_{j=1}^{n} \min_{i=1,\dots,n} |b_j - r_i| + \sum_{\substack{(h,k) \text{ such that} \\ (b_h,b_k)\in MST(b_1,\dots b_n)}} |b_h - r_{s_k}|.$$

Then by using the triangular inequality we have $|b_h - r_{s_k}| \leq |b_h - b_k| + |b_k - r_{s_k}|$ leading to:

$$M^B(r_1,\dots,r_n,b_1,..b_n) \leq \sum_{i=1}^{n} \min_{j=1,\dots,n} |r_i - b_j| + 2\sum_{j=1}^{n} \min_{i=1,\dots,n} |b_j - r_i| + M(b_1,\dots b_n).$$

□

In order to complete the result we need to study the asymptotics of $E\left[\sum_{j=1}^{n} \min_{i=1,\dots,n} |b_j - r_i|\right]$. The sum on the red points will give the same result as the points have all the same distribution. We note that due to linearity of the expected value and the fact that the minimum is taken on the red points, we have to study only:

$$E\left[\min_{i=1,\dots,n} |b - r_i|\right]$$

where b is a fixed blue point. The following lemma is a particular case of Lemma 2.1.1 of [6] when $p = 1$.

Lemma 4.3. There exists a constant $\gamma_{min}(d)$ such that

$$E\left[\min_{i=1,\dots,n} |b - r_i|\right] \leq \gamma_{min}(d) n^{-\frac{1}{d}}$$

Proof. If we consider any blue point b we have, for $B(b,\lambda)$ the ball of radius λ and centre b and $0 < l < 1$, that the volume of $B(x,\lambda) \cap [0,1]^d$ is minimized when b is in a corner of $[0,1]^d$. We then have for any $b \in [0,1]^d$:

$$\left| B(b,\lambda) \cap [0,1]^d \right| \geq c_d \lambda^d 2^{-d}.$$

Thus

$$\mathbb{P}\left[\min_{j=1,\dots,n} |b - r_j| \geq \lambda\right] \leq \left(1 - c_d \lambda^d 2^{-d}\right)^n \leq \exp(-nc_d \lambda^d 2^{-d})$$

and

$$E\left[\min_{i=1,\dots,n} |b - r_i|\right] = \int_0^\infty \mathbb{P}\left[\min_{j=1,\dots,n} |b - r_j| \geq \lambda\right] d\lambda \leq \int_0^\infty \exp(-nc_d \lambda^d 2^{-d}) d_\lambda.$$

We then apply the change of variables $y = nc_d \lambda^d 2^{-d}$ and get:

$$E\left[\min_{i=1,\dots,n} |b - r_i|\right] \leq 2d(c_d n)^{-\frac{1}{d}} \int_0^\infty y^{\frac{1}{d}-1} e^{-y} d_y.$$

And the thesis is proven for $\gamma_{min}(d) = 2dc_d^{-\frac{1}{d}} \Gamma(\frac{1}{d})$.

□

Thus combining Lemma 4.2, taking expectations and using the linearity of the expected value together with Lemma 4.3 we get:

$$2^{1-\frac{1}{d}} \beta_M(1,d)) \leq \liminf_{n\to\infty} \frac{m_n^B}{n^{1-\frac{1}{d}}} \leq \limsup_{n\to\infty} \frac{m_n^B}{n^{1-\frac{1}{d}}} \leq \beta_M(1,d)) + 3\gamma_{min}(d).$$

We note that this result confirms that the growth rate of the mean cost of the Euclidean bipartite minimum spanning tree, even if there is not convergence, is the same as the monopartite Euclidean case.

Degree

Another difference is that there is no constant M, independent from the number of points, that bounds the degree of a node as in Lemma 2.4 since, if the vertices are divided according to their color in clusters that are far enough apart, except for one blue node that is in the cluster of red nodes, then every red node will be connected to that blue one.

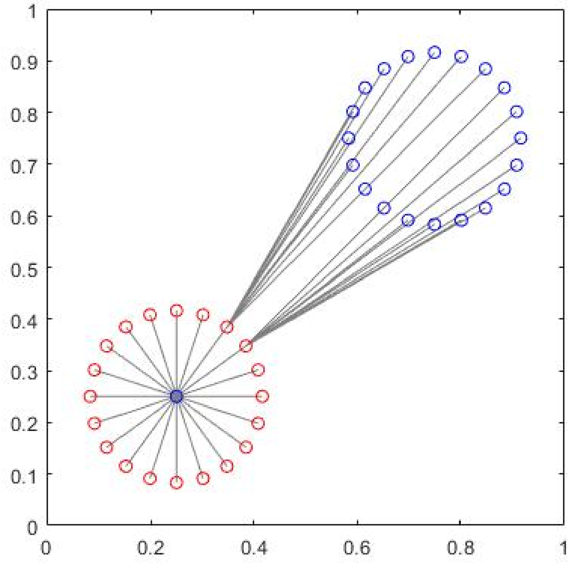


Figure 4.3: Example of bipartite MST where a node has degree n

However, we prove a result when M is dependent on n but does not scale linearly on it, and the red points are random independent and uniform on $[0,1]^d$, while the blue ones are arranged as grid in the unitary cube. For the sake of the argument we will assume that the points are $2n^d$ in order to have the region divided by the grid of the blue points in cubes of side exactly $\frac{1}{n}$.

Theorem 4.4. *We consider a distribution of points where there are n^d blue points arranged as a grid, and the red ones are n^d random uniform independent points in $[0,1]^d$. We call T_B the bipartite minimum spanning tree on those points. We then set:*

$$P_M = \mathbb{P}(\textit{there is at least one red node such that } \deg(r) \geq M \textit{ in } T_B).$$

Then there is a constant C dependent only on d such that if $M \sim C\log(n^d)$ then:

$$\lim_{n\to\infty} P_M = 0$$

Proof. We want to bound P_M with

$$\mathbb{P}(\text{there is a spherical region of area A that has no red points}).$$

Now if the minimum spanning tree connects a red node r with a blue node b with an edge of length L, then we consider S, the sector of 60^o of the circle around r of radius $L - \frac{\sqrt{d}}{2n}$, whose central radius is collinear to the edge that connects r and b. We will prove that S cannot have any red points.

To prove this we use the property **a**) and reach a contradiction. If there was a red point, r' in S we consider the MST of the graph and delete the edge (r, b). There are two cases:

- r' is in the same component as r: since $S \subset B(b, L)$ then the length of (r', b) is $< L$ but this means that there is an edge whose length is less than the one of (r, b) that connects the component of r and b .

- r' is in the same component as b. we call b_r the blue point that is in the same cell as r. then $(r', b_r) \leq (r', r) + (r, b_r) \leq L - \frac{\sqrt{d}}{2n} + \frac{\sqrt{d}}{2n} \leq L$.

We found then a region that must have no red points if there is an edge of length L.

Thus there must be a spherical region with no red points, within the sector associated with the edge of furthest length. Given that there are at least M nodes connected to r, it means that the minimum possible maximum edge, is less than $\tilde{\rho}$, the radius of the sphere of volume $\frac{M}{n^d}$. Thus there is always a sphere of radius $\rho = \frac{1}{3}\left(\tilde{\rho} - \frac{\sqrt{d}}{2n}\right)$ that contains no red points. Hence:

$$P_M \leq P[\text{there is a sphere of radius } \rho \text{ that has no red points}]$$

We now divide $[0,1]^d$ in squares Q_i, $i = 1, \dots \left\lceil \frac{\sqrt{d}}{\rho} \right\rceil^d$ of side length $\frac{1}{\left\lceil \frac{\sqrt{d}}{\rho} \right\rceil}$. Since we know there is a sphere of radius ρ with no red points, we know that at least one of the squares must have no red points.

Writing the down the expressions we get:

$$\tilde{\rho} = \left(\frac{M}{c_d n^d}\right)^{\frac{1}{d}}$$

$$\rho = \left(\frac{M}{3^d c_d n^d}\right)^{\frac{1}{d}} - \frac{\sqrt{d}}{6n^d}$$

where c_d is the volume of the d-sphere.

And thus

$$
\begin{aligned}
P_M \leq &\mathbb{P}[\text{one of the } Q_i \text{ has no red points}] = \left\lceil \frac{\sqrt{d}}{\rho} \right\rceil^d \left(1 - \frac{1}{\left\lceil \frac{\sqrt{d}}{\rho} \right\rceil^d}\right)^{n^d} \\
&\leq \left\lceil \frac{\sqrt{d}}{\rho} \right\rceil^d \exp\left(-n^d \frac{1}{\left\lceil \frac{\sqrt{d}}{\rho} \right\rceil^d}\right) \leq \left(\frac{\sqrt{d}+\rho}{\rho}\right)^d \exp\left(-n^d \frac{1}{\left(\frac{\sqrt{d}+\rho}{\rho}\right)^d}\right).
\end{aligned}
$$

Given the the dependence of ρ on n^d we can safely assume that $\rho < \sqrt{d}$, for n^d large enough, and write:

$$
\begin{aligned}
P_M \leq &\left(\frac{2\sqrt{d}}{\rho}\right)^d \exp\left(-n^d \left(\frac{\rho}{2\sqrt{d}}\right)^d\right) \\
= &\left(\frac{2\sqrt{d}}{\left(\frac{M}{3^d c_d n^d}\right)^{\frac{1}{d}} - \frac{\sqrt{d}}{6n^d}}\right)^d \exp\left(-\frac{M}{3^d c_d (2\sqrt{d})^d} + o\left(\frac{1}{n^d}\right)\right) \\
= &\left(\frac{(2\sqrt{d})^d}{\frac{M}{3^d c_d n^d} + o\left(\frac{1}{n^d}\right)}\right) \exp\left(-\frac{M}{3^d c_d (2\sqrt{d})^d} + o\left(\frac{1}{n^d}\right)\right).
\end{aligned}
$$

Thus, for $C = 3^d c_d (2\sqrt{d})^d$, if $M \sim C \log(n^d)$, we get that

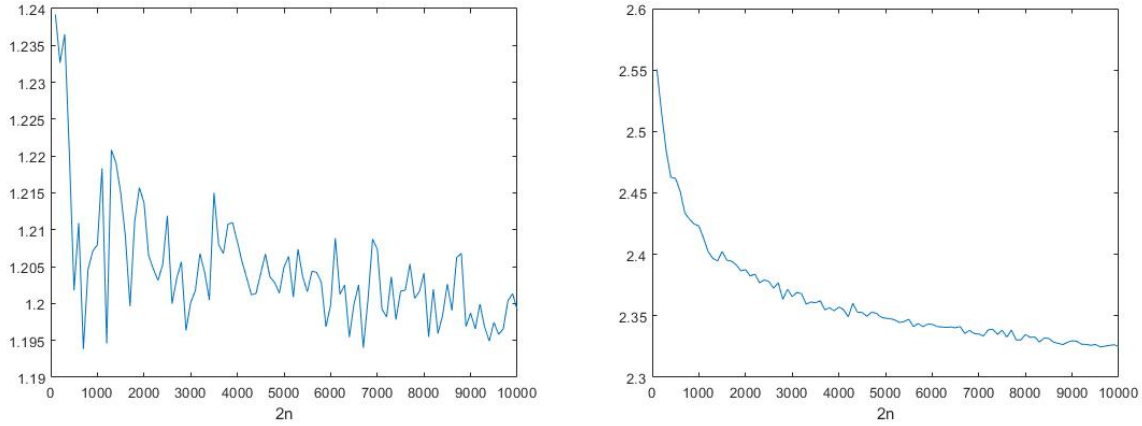
$$
P_M \leq \left(\frac{(2\sqrt{d})^d}{\frac{M}{3^d c_d n^d} + o\left(\frac{1}{n^d}\right)}\right) \exp\left(-\frac{M}{3^d c_d (2\sqrt{d})^d} + o\left(\frac{1}{n^d}\right)\right) \sim \frac{C}{\log(n^d)}.
$$

□

4.3 Asymptotics of the mean weight

In this section we discuss the asymptotics of the mean cost of the random Euclidean bipartite minimum spanning tree. Again convergence seems supported by experimental evidence in both the Euclidean model and the toroidal model. In the following charts we used again Prim's algorithm to calculate the bipartite minimum spanning tree of $2n$ random uniform points, n red and n blue, in $\mathbb{R}^d$, and plot its cost divided by $n^{1-\frac{1}{d}}$.

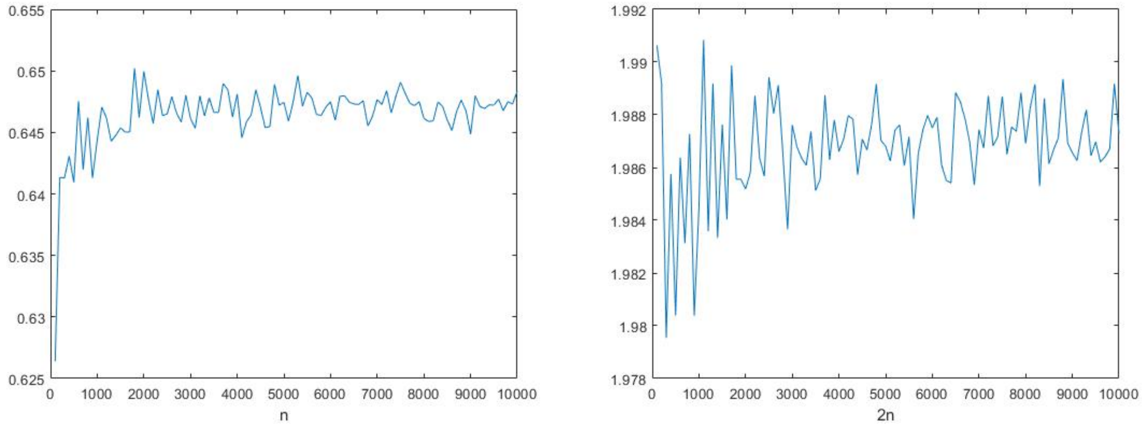
Figure 4.4: $\frac{M^B(x_1,\dots,x_n)}{n^{1-\frac{1}{d}}}$ in $[0,1]^d$



d=2

d=13

Figure 4.5: $\frac{M^B(x_1,\dots,x_n)}{n^{1-\frac{1}{d}}}$ in the toroidal model



d=2

d=13

We first look at 3.1 that works with uniform independent points in $[0,1]^d$. Even with the results of the previous section the proof does not look easily adaptable. The crucial point is given by the argument used to get 3.4, that provides a relation between Poisson point processes of different intensities. Essentially, the argument revolves around dividing the region in cubes and connecting them back, however, when dividing the region into cubes, it is not clear how many nodes will still need to be connected for each cube, as it may happen that only nodes of the same color are in there and they all still need to be connected. This property in particular does not look as easily provable with high probability as the previous ones and constitutes the biggest difficulty in adapting the first approach.

We prove convergence instead through the second approach, in the toroidal model, by showing that the four properties in **I**, **II**, **III**, **IV** can still be proven in the bipartite case.

I. in order to define the random vectors $\{d_{1,j}\}_{j=l}^{2n},..,\{d_{2n,j}\}_{j=1}^{2n}$ we note that the model is equivalent (in the sense that they generate the same minimum spanning tree) to the same graph where we add an edge of length $\sqrt{d}$ to connect every pair of vertices of the same colour. Then we have that if k and s are indexes of points of the same color, then permutation (k,s) gives a matrix of the same distribution as $\left(d_{i,j}\right)_{i,j=1}^{2n}$. On the other hand, if k and j represent points of different colors, then we just need to consider a permutation that exchange every index of a red point with an index of a blue point. □

Again, for properties **II**,**III** and **IV** a little more work is needed:

II. As we did in the monopartite case, for $x_0 = 0$, that we suppose to be a blue point, we define $\Theta'_k(z)$ as the subset of $\mathbb{T}^k$ made by every $\{x_1,\dots x_{k-1}\}$ such that there exists a tree that connects the points with edges of length $\le z$, or equivalently, the union out of every way to pick j blue points and $k-j-1$ red points such that there exists a tree that connects the points with edges of length $\le z$ for $j = 0\dots k-2$.

Now we consider how to compute $P_{k,n}(z)$ using $\Theta'_k(z)$. We will consider the case $k \ge 2$ and treat $k = 1$ separately.

Firstly, without losing generality, we can suppose that the node taken into consideration is X_0 and is blue , as they all have the same distribution.

Then since $P_{k,n}(z)$ does not really depend on where X_0 falls, but only on how the other nodes around it are placed, we can always translate every node so that $X_0 = 0$ without affecting the probability. By definition $P_{k,n}(z)$ is the probability that 0 is in a component of $G_(z)$ having exactly k nodes, namely the probability that 0 is in one of the configurations of $\Theta'_k(z)$ where the other $k-1$ points can be any subset of $k-1$ elements of all the nodes apart from 0, and the other $n-k$ nodes are not connected with any node of the configuration in G_z. Thus we can write:

$$P_{k,z} = \sum_{j=0}^{k-2}\binom{n-1}{j}\binom{n}{k-j-1}\int'_{\Theta'_{j,k}(z)}\left[1-g'_{j,z}(b_1,\dots,b_j)\right]^{(n-(k-j-1))} \\ \times\left[1-g'_{k-j-1,z}(r_1,\dots,r_{k-j-1})\right]^{(n-1)-j} db_1\dots db_j dr_1\dots dr_{k-j-1} \quad (4.1)$$

where $\Theta'_{j,k}(z)$ is the subset of $\Theta'_k(z)$ where every element has exactly j blue nodes, $g'_{s,z}(x_1\dots x_{s-1})$ is the volume of $\bigcup_{j=0}^{s-1} S'(x_j,z)$, and $\int'$ refers to the integral over $\mathbb{T}^{k-1}$.

As we noticed in the monopartite case, sets not touching the torus boundary are the same on the torus and $\mathbb{R}^d$, thus for $z \leq \frac{1}{2k}$ we get:

$$\begin{aligned} P_{k,z} = \sum_{j=0}^{k-2}\binom{n-1}{j}\binom{n}{k-j-1}\int_{\Theta_{j,k}(z)} \left[1-g_{j,z}(b_1,\dots,b_j)\right]^{(n-(k-j-1))} \\ \times\left[1-g_{k-j-1,z}(r_1,\dots,r_{k-j-1})\right]^{(n-1)-j} db_1\dots db_j dr_1\dots dr_{k-j-1} \end{aligned} \tag{4.2}$$

where:

- $\Theta_{j,k}(z)$ is the subset made by every element of $\Theta_k(z)$ having exactly j blue nodes, which in turn is the subset of $\mathbb{R}^{d\times(k-1)}$ made by every red and blue points, $\{x_1,\dots x_{k-1}\}$, such that there exists a tree that connects the points with edges of length $\leq z$;
- $g_{s,z}(x_1\dots x_{s-1})$ is the the volume of $\bigcup_{j=0}^{s-1} B(x_j,z)$;
- $\int$ is the integral over $(\mathbb{R}^d)^{k-1}$.

We now apply the change of variables $Ru_j = \frac{R_j}{z}$ and $Bu_j = \frac{B_j}{z}$ and, since $g_{k,z}(zx_1\dots zx_{k-1}) = z^d g_{k,1}(x_1\dots x_{k-1})$, we have:

$$\begin{aligned} P_{k,z} = \sum_{j=0}^{k-2}\binom{n-1}{j}\binom{n}{k-j-1} z^{d(k-1)}\int_{\Theta_{j,k}(1)} \left[1-z^d g_{j,1}(bu_1,\dots,bu_j)\right]^{(n-(k-j-1))} \\ \times\left[1-z^d g_{k-j-1,1}(ru_1,\dots,ru_{k-j-1})\right]^{(n-1)-j} dbu_1\dots dbu_j dru_1\dots dru_{k-j-1}. \end{aligned} \tag{4.3}$$

Let $z = \left(\frac{y}{nc_d}\right)^{\frac{1}{d}}$. We can still assume $z \leq \frac{1}{2k}$ as we are going to take the limit on n.

Then:

$$\begin{aligned} P_{k,n}\left(\left(\frac{y}{nc_d}\right)^{\frac{1}{d}}\right) = \sum_{j=0}^{k-2}\prod_{s=1}^{j}\left(1-\frac{s}{n}\right)\prod_{i=1}^{k-j-1}\left(1-\frac{i}{n}\right)\frac{y^{k-1}}{c_d^{k-1}j!(k-j-1)!}\int_{\Theta_{j,k}(1)}\left[1-\frac{y}{nc_d}g_{j,1}(bu_1,\dots,bu_j)\right]^{(n-(k-j-1))} \\ \times\left[1-\frac{y}{nc_d}g_{k-j-1,1}(ru_1,\dots,ru_{k-j-1})\right]^{(n-1)-j} dbu_1\dots dbu_j dru_1\dots dru_{k-j-1}. \end{aligned} \tag{4.4}$$

Now we take the limit on n, and since $|\Theta_{j,k}(1)|$ is bounded for every j and $\left[1-\frac{y}{nc_d}g_{k,z}(u_1\dots u_{s-1})\right]^{n-s} \leq 1$ we can apply dominated convergence to get:

$$f_k(y) = \lim_{n\to\infty} P_{k,n}\left(\left(\frac{y}{nc_d}\right)^{\frac{1}{d}}\right) = \sum_{j=0}^{k-2} \frac{y^{k-1}}{c_d^{k-1} j!(k-j-1)!} \int_{\Theta_{j,k}(1)} \exp\left[-\frac{y}{c_d} g_{j,1}(bu_1,\dots,bu_j)\right]$$
$$\times \exp\left[-\frac{y}{c_d} g_{k-j-1,1}(ru_1,\dots,ru_{k-j-1})\right] dbu_1\dots dbu_j dru_1\dots dru_{k-j-1}. \quad (4.5)$$

If $k=1$ instead, $P_{1,n}(z)$ is the same as the probability that the other n red points do not belong to $B(0,z)$, and thus:

$$f_k(y) = \lim_{n\to\infty} P_{k,n}\left(\left(\frac{y}{nc_d}\right)^{\frac{1}{d}}\right) = \lim_{n\to\infty}\left[1-c_d\frac{y}{nc_d}\right]^n = e^{-y}$$

□

III. Since $1-x\leq e^{-x}$, for $k=1$ we have:

$$P_{1,n}\left(\left(\frac{y}{nc_d}\right)^{\frac{1}{d}}\right) = \left[1-c_d\frac{y}{nc_d}\right]^n \leq e^{-y}$$

If $k\geq 2$, for $z\geq\frac{\sqrt{d}}{2}$, $P_{k,n}(z)=0$. So we can assume $z<\frac{\sqrt{d}}{2}$.
Since $g'_{k,z}(x_1,\dots x_{k-1})\geq c_d z^d$ we have:

$$\begin{aligned}
P_{k,z} &= \sum_{j=0}^{k-2}\binom{n-1}{j}\binom{n}{k-j-1}\int'_{\Theta'_{j,k}(z)}\left[1-g'_{j,z}(b_1,\dots,b_j)\right]^{(n-(k-j-1))} \\
&\quad\times\left[1-g'_{k-j-1,z}(r_1,\dots,r_{k-j-1})\right]^{(n-1)-j} db_1\dots db_j dr_1\dots dr_{k-j-1} \\
&\leq \sum_{j=0}^{k-2}\binom{n-1}{j}\binom{n}{k-j-1}\int'_{\Theta'_{j,k}(z)}\exp\Big(-(n-(k-j-1))g'_{j,z}(b_1,\dots,b_j)\Big) \\
&\quad\times\exp\Big(-(n-1-j)g'_{k-j-1,z}(r_1,\dots,r_{k-j-1})\Big) db_1\dots db_j dr_1\dots dr_{k-j-1} \\
&\leq \sum_{j=0}^{k-2}\frac{n^{k-1}}{j!(k-j-1)!}e^{-(n-k)c_d z^d}\int'_{\Theta'_{j,k}(z)} db_1\dots db_j dr_1\dots dr_{k-j-1}.
\end{aligned}$$

Again, for every j, we bound $|\Theta'_{j,k}(z)|$ with $k^{k-2}(c_d z)^d$, as we note that given x_i one can choose x_{i+1} only in the ball of radius z centred in x_i, and those nodes can be connected in less than k^{k-2} trees.

We also use $\frac{n}{n-k} \geq \frac{1}{k+1}$, and get:

$$P_{k,z} \leq \left(\sum_{j=0}^{k-2} \frac{n^{k-1}}{j!(k-j-1)!} \right) e^{-\frac{n}{k+1} c_d z^d} k^{k-1} (c_d z^d)^{k-1}.$$

Then for $z = \left(\frac{y}{nc_d} \right)^{\frac{1}{d}}$:

$$P_{k,n} \left(\left(\frac{y}{nc_d} \right)^{\frac{1}{d}} \right) \leq l_k(y) = \left(\sum_{j=0}^{k-2} \frac{1}{j!(k-j-1)!} \right) y^{k-1} k^{k-1} e^{-\frac{y}{k+1}}.$$

Thus, since $\int_0^\infty l_k(y) y^{\frac{1}{d}-1} dy < \infty$, property **III** is also verified. □

Before proving property **IV** we first need a lemma corresponding to Lemma 4.3 in the toroidal case.

Lemma 4.5. Given $b \in [0,1]^d$ and $r_1, \dots, r_n$ random uniform independent points in $[0,1]^d$, there exists a constant $\beta_{min}(d)$ such that

$$\lim_{n \to \infty} n^{\frac{1}{d}} E \left[\min_{i=1,\dots,n} |b - r_i| \right] = \beta_{min}(d).$$

Proof. Since we are working on the torus $\mathbb{P}\left[\min_{j=1,\dots,n} |b - r_j| \geq \lambda\right]$ does not depend on where b is, but only on how the other points around it are placed. Thus we can always translate every point so that $b = 0$ without affecting the probability. We then have for $\lambda \leq \frac{1}{2}$:

$$\mathbb{P}\left[\min_{j=1,\dots,n} |b - r_j| \geq \lambda \right] = \left(1 - c_d \lambda^d\right)^n.$$

We note that $\left(1 - c_d 2^{-d}\right)^n$ provides an upper bound for $\mathbb{P}\left[\min_{j=1,\dots,n} |b - r_j| \geq \lambda\right]$ when $\frac{1}{2} < \lambda < \sqrt{d}$ and there cannot be any point further from b than $\sqrt{d}$. Thus:

$$n^{\frac{1}{d}} \int_0^{\frac{1}{2}} \left(1 - c_d \lambda^d\right)^n d\lambda \leq n^{\frac{1}{d}} E \left[\min_{i=1,\dots,n} |r_i - b_j| \right] \leq n^{\frac{1}{d}} \int_0^{\frac{1}{2}} \left(1 - c_d \lambda^d\right)^n d\lambda + n^{\frac{1}{d}} \int_{\frac{1}{2}}^{\sqrt{d}} \left(1 - c_d 2^{-d}\right)^n d\lambda. \tag{4.6}$$

The last integral is bounded by

$$n^{\frac{1}{d}} (\sqrt{d} - \frac{1}{2}) \exp(-nc_d 2^{-d}).$$

On the first integral we apply the change of variables $y = nc_d\lambda^d$ and get:

$$n^{\frac{1}{d}}\int_0^{\frac{1}{2}}\left(1-c_d\lambda^d\right)^n d\lambda = d(c_d)^{-\frac{1}{d}}\int_0^{\infty} y^{\frac{1}{d}-1}\left(1-\frac{y}{n}\right)^n \mathbb{1}_{[0,c_d 2^{-d}n]}(y)dy.$$

Where $\mathbb{1}_A$is the indicator function of the set A.

Finally we take the limit on equation 4.6 and apply the dominated convergence theorem on $y^{\frac{1}{d}-1}\left(1-\frac{y}{n}\right)^n \mathbb{1}_{[0,c_d 2^{-d}n]}(y)$ getting:

$$\lim_{n\to\infty} n^{\frac{1}{d}}E\left[\min_{i=1,\dots,n}|r_i-b_j|\right] = d(c_d)^{-\frac{1}{d}}\Gamma\left(\frac{1}{d}\right) =: \beta_{min}(d).$$

□

We are now ready to prove **IV**.

IV. Let $C_{K,n}(z)$ be the number of components, formed by K nodes or more, of $G_n(z)$. Then for the same argument used in Theorem 3.5 we have:

$$E\left[C_{n,K}\right] = n\sum_{k=K}^{n}\frac{P_{k,n}(z)}{k}.$$

Therefore, knowing that there a constant M that bounds the length of the edges, we can write:

$$n^{\frac{1}{d}}\int_0^{\infty}\left[\sum_{k=K}^{n}\frac{P_{k,n}(z)}{k}-\frac{1}{n}\right]dz = \frac{1}{n^{1-\frac{1}{d}}}\int_0^{M}\left(E\left[C_{K,n}(z)\right]-1\right)dz \geq -\frac{M}{n^{1-\frac{1}{d}}} \geq -\frac{M}{K^{1-\frac{1}{d}}}. \qquad (4.7)$$

As we noticed in the proof of **IV** for the monopartite case:

$$\int_0^{M}\left(C_{K,n}(z)-1\right)dz = \sum_{z_i\in J_-} z_i^- - \sum_{z_i\in J_+} z_i^+ \leq \sum_{z_i\in J_-} z_i^-.$$

Where $J^+ \subset J$ is the set of the lengths of the edges that, when added, increase $C_{K,n}$, and J^- is the same for the edges that make $C_{K,n}$ decrease.

We call $m = |J_-|$ and try to bound $\sum_{z_i\in J_-} z_i^-$.

Since they represent the shortest nodes that connects cluster of points they from a tree on a multigraph, we can bound its cost by picking any representatives in the clusters, as the tree on the multigraphs will pick the edges that connects the closest nodes. We then choose $w_1,\dots w_m$ to be all blue and build the usual monopartite MST. Its length might actually be smaller than the minimum spanning tree in the bipartite case but it is possible to have similar edges by replacing every edge (w_i, w_j) with $w_i, r_{k_j})$, where r_{k_j} is the closest red node to w_j, and then connecting it back to w_j.

We note that since every $z_i{}^-$ connects only clusters of k points then, if $k \geq 2$, every w_j must be already connected to another red vertex through an edge that, by definition of $G_n(z)$, must have length less than z. But this means, again by definition of $G_n(z)$, that w_j must be already connected to their closest red point too and thus we just have to consider (w_i, r_{kj}) that are edges of the multigraph. We then got a graph that connects the representatives in the bipartite case and by the triangular inequality we get that: $|(w_i, r_{kj})| \leq |(w_i, w_j)| + |(w_j, r_{kj})|$ leading to:

$$M_B(w_1, \dots w_m) \leq M(w_1, \dots w_m) + \sum_{j=1}^{m} \min_{i=1,\dots,n} |r_i - w_j|.$$

We can now bound $M(w_1, \dots w_m)$ using Theorem 2.2 as we did in the monopartite case, while we still need a little more work for the second term.

By taking expectations we get:

$$\begin{aligned} n^{\frac{1}{d}} \int_0^\infty \left[\sum_{k=K}^{n} \frac{P_{k,n}(z)}{k} - \frac{1}{n} \right] dz \leq & \frac{1}{n^{1-\frac{1}{d}}} E\left[\sum_{z_i \in J_-} z_i^- \right] \\ \leq & \frac{1}{n^{1-\frac{1}{d}}} k_d \left(\frac{n}{K}\right)^{\frac{d-1}{d}} + \frac{1}{n^{1-\frac{1}{d}}} E\left[\sum_{j=1}^{m} \min_{i=1,\dots,n} |r_i - w_j| \right]. \end{aligned}$$

Now as a consequence of Lemma 4.5 we can choose $\epsilon > 0$ and $\tilde{n}$ such that for every $n > \tilde{n}$ and any $b \in [0,1]^d$

$$E\left[\min_{i=1,\dots,n} |r_i - b| \right] \leq n^{-\frac{1}{d}} (\beta_{min}(d) + \epsilon).$$

Since $m = |J^-| \leq \frac{n}{K}$, we have:

$$E\left[\sum_{j=1}^{m} \min_{i=1,\dots,n} |r_i - w_j| \right] \leq \begin{cases} \sqrt{d}\frac{\tilde{n}}{K} & \text{if } n \leq \tilde{n} \\ \frac{n^{1-\frac{1}{d}}}{K}(\gamma + \epsilon) & \text{if } n > \tilde{n} \end{cases}$$

and then:

$$E\left[\sum_{j=1}^{m} \min_{i=1,\dots,n} |r_i - w_j| \right] \leq \gamma'(d) \frac{n^{1-\frac{1}{d}}}{K} \leq \gamma'(d) \left(\frac{n}{K}\right)^{1-\frac{1}{d}}$$

where $\gamma'(d) = (\sqrt{d}\tilde{n} + \beta_{min}(d) + \epsilon)$.

Then,

$$n^{\frac{1}{d}} \int_0^\infty \left[\sum_{k=K}^{n} \frac{P_{k,n}(z)}{k} - \frac{1}{n} \right] dz \leq \frac{1}{n^{1-\frac{1}{d}}} C \left(\frac{n}{K} \right)^{\frac{d-1}{d}} = \frac{C}{K^{1-\frac{1}{d}}} \tag{4.8}$$

Where $C = k_d + \gamma'$

Thus equation 4.7 and 4.8 give us the thesis. □

Combining the general theorem with what we have just developed about the Euclidean bipartite MST:

Theorem 4.6. *In the Euclidean bipartite toroidal model:*

$$\beta_T^B(d) := \lim_{n\to\infty} \frac{E[T_n]}{n^{\frac{d-1}{d}}} = \frac{\Gamma\left(\frac{1}{d}\right)}{d(c_d)^{\frac{1}{d}}} + \sum_{k=2}^{\infty} \frac{\Gamma\left(k+\frac{1}{d}-1\right)}{dk} \sum_{j=0}^{k-2} \frac{\eta_{j,k}(d)}{j!(k-j-1)!} \tag{4.9}$$

for

$$\eta_{j,k}(d) = \int_{\Theta_{j,k}(1)} \left(g_{j,1}(bu_1,\ldots,bu_j) + g_{k-j-1,1}(ru_1,\ldots,ru_{k-j-1}) \right)^{-(k+\frac{1}{d}-1)} dbu_1 \ldots dbu_j dru_1 \ldots dru_{k-j-1}$$

where

- $\Theta_{j,k}(1)$ *is the subset made by every element of* $\Theta_k(z)$ *having exactly j blue nodes, which in turn is the subset of* $\mathbb{R}^{d\times(k-1)}$ *made by every red and blue points,* $\{x_1,\ldots x_{k-1}\}$*, such that there exists a tree that connects the points with edges of length* ≤ 1

- $g_{s,1}(x_1 \ldots x_{s-1})$ *is the the volume of* $\bigcup_{j=0}^{s-1} B(x_j, 1)$.

Conclusions and perspectives

In this thesis we have analyzed the random Euclidean MST problem, when the points are independent uniformly distributed random variables. While for the monopartite case the results were well established, we had to find some new arguments to adapt the techniques for the bipartite case. In chapter 2 we showed some deterministic properties for the monopartite problem and in chapter 3 we have proven through two different techniques that there is convergence for its mean cost.

It would be interesting to find when convergence is still verified, if the nodes are not uniform but have another distribution, even if the red and blue vertices have different distributions. Also another interesting variant of the problem is the case when the bipartite graph has also a constraint on the maximum degree of the nodes. Speaking of the maximum degree we also would like to know whether the bound that was proven on Theorem 4.4 is tight and it's possible to find a lower bound that is asymptotically similar. Finally another question that comes out naturally is if similar results can be generalised in the k-partite version of the problem, when the nodes are divided in k subsets of different colors.

Bibliography

[1] Florin Avram, Dimitris Bertsimas, et al. “The minimum spanning tree constant in geometrical probability and under the independent model: a unified approach”. In: *The Annals of Applied Probability* 2.1 (1992), pp. 113–130.

[2] Sergio Caracciolo, Dott Enrico Malatesta, and Andrea Riva. *The random minimum spanning tree problem*. URL: `http://pcteserver.mi.infn.it/~caraccio/Lauree/Riva.pdf`.

[3] P. Crescenzi. *Strutture di dati e algoritmi. Progettazione, analisi e visualizzazione*. Pearson, 2012. ISBN: 9788871927817.

[4] Patrick Jaillet et al. “Cube versus torus models and the Euclidean minimum spanning tree constant”. In: *Annals of Applied Probability* 3.2 (1993), pp. 582–592.

[5] J Michael Steele. “Growth rates of Euclidean minimal spanning trees with power weighted edges”. In: *The Annals of Probability* (1988), pp. 1767–1787.

[6] J.M. Steele. *Probability Theory and Combinatorial Optimization*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, 1997. ISBN: 9780898713800.