A NOTE ON THE NUMBER OF LEAVES OF A EUCLIDEAN MINIMAL SPANNING TREE

by

Patrick Jaillet

OR 235-90 November 1990

A Note on the Number of Leaves of a Euclidean Minimal Spanning Tree

Patrick Jaillet*

Laboratoire de Mathématiques et Modélisation

Ecole Nationale des Ponts et Chaussées, France

November 12, 1990

Abstract

We show that the number of vertices of degree k in the Euclidean minimal spanning tree through points drawn uniformly from either the d-dimensional torus or from the d-cube, $d \geq 2$, are asymptotically equivalent with probability one. Implications are discussed.

*Research carried out at MIT and supported in part by a CNRS/NSF Grant

1 Introduction

In Steele et al. [3], the authors prove that for any independent and uniform random variables $\{X_i : 1 \leq i < \infty\}$ in $[0,1]^d$, $d \geq 2$, the number of vertices of degree k in the Euclidean minimal spanning tree through $\{X_1, \ldots, X_n\}$ is asymptotic to a constant $\alpha_{k,d}$ times n with probability one. In the case $k = 1$ and $d = 2$ (i.e., for the number of leaves of the MST in the square), the authors have shown that the constant $\alpha = \alpha_{1,2}$ is positive and that Monte Carlo simulation results suggest that $\alpha = 2/9$ is a reasonable approximation. If one attempts to get any more information on this constant, one rapidly finds that the boundary effects of the square is a serious limitation on any analytical approach. This observation is the main motivation of this paper: we show that the results of Steele et al. [3] are still valid in the d-torus (with the same constants). Hence any attempts on characterizing these constants could now be made within the torus model, with no boundary problems. For example, it is clear, from the symmetry induced by the d-torus model, that $\alpha_{k,d}$ is now equal to $\lim_{n\to\infty} \mathbf{P}(H_d^{(n)} = k)$, where $H_d^{(n)}$ is the degree of any point, say X_1, in a minimal spanning tree through $\{X_1, \ldots, X_n\}$ in the d-torus. Before going into the details of our proofs, let us first give some notation.

Let $G = (V, E)$ be a connected graph with vertex set V and edge set E, together with a weight function $w : E \to \mathbf{R}$ which assigns a real number to each edge in E. A minimal spanning tree (MST) of G is a connected graph with vertex set V and edge set $T \subset E$ such that $\sum_{e \in T} w(e)$ is minimal. If no ambiguity can arise on the vertex set, a spanning tree is usually identified by its set of edges; we will adopt this convention hereafter. The two models of interest here are the following.

The d-cube model:

Let $\{x_i : 1 \leq i < \infty\}$ be an arbitrary infinite sequence of points in $[0,1]^d$ (the unit cube in $\mathbf{R}^d$, $d \geq 2$, considered as the d-dimentional space of real numbers, with its Euclidean metric and Lebesgue measure), and let $x^{(n)} = \{x_1, x_2, \ldots, x_n\}$ denote its first n points. For each finite n, $x^{(n)}$ will be the vertex set V, and $E = \{(x_i, x_j); 1 \leq i < j \leq n\}$ the edge set of our graph. The weight of an edge (x_i, x_j) will be the Euclidean distance $||x_i - x_j||$ from x_i to x_j.

The d-torus model:

In order to eliminate the boundary effects of the previous model, consider the previous sequence $x_1, x_2, \ldots, x_n, \ldots$ modulo 1 in each component. Alternatively, one can imagine a sequence on the d-torus $T^d = ([0,1] \bmod 1)^d$ (the metric space with its Lebesgue measure and Euclidean d-torus metric). Note that the weight of an edge (x_i, x_j) is now taken to be $||x_i - x_j(\text{mod}1)^d||$ (for $y \in [-1, 1]$, $y(\text{mod}1)$ is the minimum of $|y|$ and $1 - |y|$).

Other notations:

An optimal MST in the d-cube will be described by its set of edges $\mathcal{A}_{cube}(x^{(n)})$. With a slight abuse of notation, we will use $\mathcal{A}_{torus}(x^{(n)})$ for the corresponding problem in the d-torus. For a given k, we will write $VC_k(x^{(n)})$ and $VT_k(x^{(n)})$ for the number of vertices of degree k in the MST in the d-cube and d-torus, respectively. Finally, $L_{max}(x^{(n)})$ will be the length of the largest edge in an optimal MST in the d-torus. We can now state our main result.

Theorem 1 *Let $(X_i)_i$ is a sequence of points independently and uniformly distributed over $[0,1]^d$. Then there are constants $\alpha_{k,d}$ such that*

$$\lim_{n \to \infty} n^{-1} VT_k(X^{(n)}) = \lim_{n \to \infty} n^{-1} VC_k(X^{(n)}) = \alpha_{k,d} \text{ (a.s.)}. \tag{1.1}$$

The existence of the constants verifying the second equality was proved in Steele et al. [3]. For proving the remaining part of this theorem, we need two intermediate lemmas. Both of them have been proved in Jaillet [2], but, for completeness, the proofs are restated in the next section.

2 Intermediate lemmas

Lemma 1 *Let $(X_i)_i$ be a sequence of points independently and uniformly distributed over $[0,1]^d$. Then we have*

$$\mathbf{P}\left(L_{max}(X^{(n)}) > \lambda_d (\log n/n)^{1/d}\right) \leq 1/n^2 \log n, \tag{2.2}$$

where $\lambda_d = 12^{1/d}\sqrt{d+3}$.

Proof:

First let $(Q_j)_{1\leq j\leq m^d}$ be a partition of the d-cube $[0,1]^d$ into cubes with edges parallel to the axle and of length $1/m$. If for a sequence of points $\{x_i : 1 \leq i < \infty\}$, $x^{(n)} \cap Q_j$ is not empty for all j, then the MST in the d-torus is such that

$$L_{max}(x^{(n)}) \leq \sqrt{d+3}/m. \tag{2.3}$$

Indeed let e be an edge of $\mathcal{A}_{torus}(x^{(n)})$ so that its weight is $L_{max}(x^{(n)})$. By discarding e we end up with a forest with two components, with point sets, say V_e and W_e, such that for all $x_i \in V_e$ and all $x_j \in W_e$ we have $||x_i - x_j (\text{mod} 1)^d|| \geq L_{max}(x^{(n)})$ (by definition of an optimal MST). As a working hypothesis, let us then assume that $L_{max}(x^{(n)}) > \sqrt{d+3}/m$. Then $L_{max}(x^{(n)}) > \sqrt{d}/m$ and thus each Q_j either contains points from V_e or from W_e but not from both. Now, since all cubes are non-empty, we can always find a pair of adjacent (i.e, sharing a facet) cubes Q_i and Q_j such that Q_i contains points from V_e and Q_j contains points from W_e. But now, the largest possible edge connecting these two squares is bounded from above by $\sqrt{d+3}/m$ and thus, using our working hypothesis, by $L_{max}(x^{(n)})$. But this contradicts the definition of an optimal MST (see above). Hence (2.3) is valid.

Now let $(X_i)_i$ be a sequence of points independently and uniformly distributed over $[0,1]^d$, and consider the same partition as before. If N_j denotes the cardinality of $X^{(n)} \cap Q_j$, then, with $p = 1/m^d$, we have, for $h \geq 12$ and $n \geq 3$,

$$\mathbf{P}\left(\forall j : N_j > np - \sqrt{hnp\log n}\right) \geq 1 - 1/2pn^{h/4}. \tag{2.4}$$

Indeed for all j let $\mathcal{B}_{n,j}$ be the event $\left\{N_j \leq np - \sqrt{hnp\log n}\right\}$. We obviously have

$$\mathbf{P}\left(\exists j : \mathcal{B}_{n,j}\right) \leq \sum_{j=1}^{m^d} \mathbf{P}\left(\mathcal{B}_{n,j}\right) = m^d \mathbf{P}\left(\mathcal{B}_{n,1}\right) = \mathbf{P}\left(\mathcal{B}_{n,1}\right)/p. \tag{2.5}$$

Now N_1 is a binomial random variable with n trials and parameter p. Using classical bounds on the tail of a binomial distribution (see [1, Corollary 4, p.11]) we have, with $q = 1 - p$,

$$\mathbf{P}\left(N_1 \leq np - \sqrt{hnp\log n}\right) \leq \frac{1}{2}\exp\left\{-h\log n/3q + 1/q\right\} \leq 1/2n^{h/4}, \tag{2.6}$$

the last inequality being valid for $h \geq 12$, and $n \geq 3$. Now (2.4) follows from (2.5) and (2.6).

Finally let us take $m^d (= 1/p)$ to be $< n/(h\log n)$. Then, from (2.4), we have

$$\mathbf{P}\left(\forall j : N_j > 0\right) > 1 - 1/2hn^{h/4-1}\log n. \tag{2.7}$$

But, from (2.3), we also have, for any $\varepsilon > 0$,

$$\mathbf{P}\left(L_{max}(X^{(n)}) \leq (h+\varepsilon)^{1/d}\sqrt{d+3}\,(\log n/n)^{1/d}\right) \geq \mathbf{P}\left(\forall j : N_j > 0\right). \tag{2.8}$$

Hence the lemma follows from (2.7) and (2.8) by taking $h = 12$.

∎

Lemma 2 *Let $\{x_i : 1 \leq i < \infty\}$ be an arbitrary infinite sequence of points in $[0,1]^d$ and $\mathcal{A}_{torus}(x^{(n)})$ be an optimal MST in the d-torus through $x^{(n)}$. Let $\mathcal{B}_{torus}(x^{(n)})$ be the set of edges (x_i, x_j) of this solution such that $||x_i - x_j(\mathrm{mod} 1)^d|| = ||x_i - x_j||$ (i.e., that do not 'cross' the boundary of the d-cube). Then there exists an optimal solution for this problem in the d-cube, say $\mathcal{A}^*{}_{cube}(x^{(n)})$, such that:*

If (x_i, x_j) belongs to $\mathcal{B}_{torus}(x^{(n)})$, then (x_i, x_j) belongs to $\mathcal{A}^*{}_{cube}(x^{(n)})$ (2.9)

Proof:

Take any edge $(x_i, x_j) \in \mathcal{B}_{torus}(x^{(n)})$ and consider a given solution $\mathcal{A}_{cube}(x^{(n)})$ in the

d-cube. Suppose that (x_i, x_j) does not belong to this solution. Then $\mathcal{A}_{cube}(x^{(n)}) \cup \{(x_i, x_j)\}$ contains a unique cycle, say $\mathcal{C}_{cube}$, such that for all edges $(x_k, x_l) \in \mathcal{C}_{cube} \setminus \{(x_i, x_j)\}$ we have

$$||x_k - x_l|| < ||x_i - x_j||. \tag{2.10}$$

(Note that we can discard the easy case for which there is an edge (x_k, x_l) such that $||x_k - x_l|| = ||x_i - x_j||$. Indeed we would then exchange the two edges and obtain an optimal solution in the d-cube that verifies (2.9) for the edge (x_i, x_j) under consideration). Now among the edges of $\mathcal{C}_{cube} \setminus \{\mathcal{A}_{torus}(x^{(n)}) \cap \mathcal{C}_{cube}\}$ there is at least one edge, say (x_k, x_l) such that $\mathcal{A}_{torus}(x^{(n)}) \cup \{(x_k, x_l)\}$ has a cycle containing (x_i, x_j). The proof of this key result goes as follows. Let $Z = (z_1, \ldots, z_m)$ be the points (other than x_i and x_j) along the cycle $\mathcal{C}_{cube}$, numbered as they appear from x_i to x_j. By definition of a spanning tree there is a unique path in $\mathcal{A}_{torus}(x^{(n)})$ going from x_i to each of these points. Color a point of Z red if this path does not go through x_j and blue otherwise. Note that the blue points can alternatively be defined as the points reached from x_j without going through x_i. Also color x_i red and x_j blue. Now any edge of $\mathcal{C}_{cube} \setminus \{(x_i, x_j)\}$ with adjacent points of opposite color, if added to $\mathcal{A}_{torus}(x^{(n)})$, would form a cycle (in $\mathcal{A}_{torus}(x^{(n)})$) containing (x_i, x_j). Now, by going along the cycle, starting from x_i and in the opposite direction of x_j, we must find such an edge, since x_i and x_j are of opposite color. It is now easy to conclude. Indeed if we remove (x_i, x_j) from $\mathcal{A}_{torus}(x^{(n)})$ and replace it by (x_k, x_l), we end up (from (2.10)) with a spanning tree of weight less than $L_{torus}(x^{(n)})$: A contradiction.

■

3 Proof of the main result

From Lemma 2, we see that $\mathcal{A}_{torus}(x^{(n)})$ and $\mathcal{A}_{cube}(x^{(n)})$ have a strong related structure. In fact, for the case of random points, $\mathcal{A}_{torus}(X^{(n)})$ and $\mathcal{A}_{cube}(X^{(n)})$ are unique with probability one and thus, if $\mathcal{E}_{torus}(X^{(n)})$ denotes the set of edges of $\mathcal{A}_{torus}(X^{(n)})$ that 'crosses' the boundary of the d-cube, this lemma tells us that, with probability one, we have

$$\mathcal{A}_{torus}(X^{(n)}) \setminus \mathcal{E}_{torus}(X^{(n)}) \subset \mathcal{A}_{cube}(X^{(n)}). \tag{3.11}$$

Now consider any spanning tree T on an arbitrary connected graph $G = (V, E)$, and any pair of edges $e \in T$ and $e' \in E \setminus T$ such that $T' = T \setminus \{e\} \cup \{e'\}$ is still a spanning tree. For any given k, let N_k and N'_k be the number of vertices of degree k in T and T', respectively. Then it is easy to see that

$$|N_k - N'_k| \leq 4. \tag{3.12}$$

(in fact, for leaves, this can be improved to $|N_1 - N'_1| \leq 2$). From (3.11) and (3.12) we then have

$$\left| VC_k(X^{(n)}) - VT_k(X^{(n)}) \right| \leq 4\text{card}(\mathcal{E}_{torus}(X^{(n)})). \tag{3.13}$$

Lemma 1 will now play a role. Consider $Q(r) = [0,1]^d \setminus [r, 1-r]^d$ a layer of width r on the inside of the d-cube, and partition $\mathcal{E}_{torus}(X^{(n)})$ in two sets: the set $\mathcal{E}_1^{(r)}(X^{(n)})$ of 'crossing' edges having at least one endpoints in $Q(r)$, and the set $\mathcal{E}_2^{(r)}(X^{(n)})$ of the remaining 'crossing' edges. Now it is easy to see that

$$\text{card}(\mathcal{E}_1^{(r)}(X^{(n)})) \leq D_d \text{card}(\{X_i : X_i \in Q(r)\}), \tag{3.14}$$

where D_d is the number of spherical caps with angle $\pi/3$ which are needed to cover the unit sphere in $\mathbf{R}^d$ (an upper bound on the degree of any vertex in an optimal MST). From the the strong law of large numbers we have

$$\lim_{n\to\infty} n^{-1}\text{card}(\{X_i : X_i \in Q(r)\}) = 1 - (1-2r)^d \text{ (a.s.)}. \tag{3.15}$$

Also from Lemma 1 we have

$$\sum_{n=1}^{\infty} \mathbf{P}(\text{card}(\mathcal{E}_2^{(r_n)}(X^{(n)})) > 0) \leq \sum_{n=1}^{\infty} \mathbf{P}(L_{max}(X^{(n)}) > r_n) \leq \sum_{n=1}^{\infty} 1/n^2 \log n < \infty, \tag{3.16}$$

where $r_n = \lambda_d(\log n/n)^{1/d}$. From Borel-Cantelli this implies that

$$\lim_{n\to\infty} \mathrm{card}(\mathcal{E}_2^{(r_n)}(X^{(n)})) = 0 \text{ (a.s.)}. \tag{3.17}$$

One finally concludes from (3.13), (3.15) (with $r = r_n$), and (3.17).

References

[1] Bollobás, B. 1985. *Random Graphs.* Academic Press. London.

[2] Jaillet, P. 1990. Cube versus Torus Models for Combinatorial Optimization Problems and the Euclidean Minimum Spanning Tree Constant. Working Paper ORC/MIT October 1990.

[3] Steele, J., L. Shepp, and W. Eddy. 1986. On the Number of Leaves of a Euclidean Minimal Spanning Tree. *J. Appl. Proba.*, **24**, 809–826.