Probab. Theory Relat. Fields 92, 247–258 (1992)

Probability Theory and Related Fields

Asymptotics for Euclidean minimal spanning trees on random points

David Aldous[1,*] **and J. Michael Steele**[2,**]

[1] Department of Statistics, University of California, Berkeley, CA 94720, USA
[2] Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA

Received March 1, 1991; in revised form September 3, 1991

Summary. Asymptotic results for the Euclidean minimal spanning tree on n random vertices in R^d can be obtained from consideration of a limiting infinite forest whose vertices form a Poisson process in all R^d. In particular we prove a conjecture of Robert Bland: the sum of the d'th powers of the edge-lengths of the minimal spanning tree of a random sample of n points from the uniform distribution in the unit cube of R^d tends to a constant as $n \to \infty$.

Whether the limit forest is in fact a single tree is a hard open problem, relating to continuum percolation.

1 Introduction

Let $\mathbf{x} = \{x_1, \ldots, x_n\}$ be a finite set of points in R^d, $d \geqq 2$. The minimal spanning tree (MST) t of $\mathbf{x}$ is a connected graph with vertex-set $\mathbf{x}$ such that the sum of the edge-lengths of t is minimal, i.e.

$$\sum_{e \in t} |e| = \min_{G} \sum_{e \in G} |e|$$

where $|e| = |x_i - x_j|$ is the Euclidean length of the edge $e = (x_i, x_j)$ and the minimum is over all connected graphs G. Minimal spanning trees are one of the most studied objects in combinatorial optimization, and even the probability theory of MST is rather well developed. See Steele [13] for an overview. In that paper it was shown that, if $(x_i : 1 \leqq i \leqq n)$ are i.i.d. with compactly supported density f and if $0 < \alpha < d$, then

$$n^{-(d-\alpha)/d} \sum_{e \in t} |e|^{\alpha} \to c(\alpha, d) \int_{R^d} (f(x))^{(d-\alpha)/d} \, dx \quad \text{a.s.} \tag{1}$$

* Research supported by N.S.F. Grants MCS87-01426 and MCS 90-01710
** Research supported in part by N.S.F. Grant DMS88-12868, A.F.O.S.R. Grant 89-0301, ARO Grant DAAL03-89-G-0092 and NSA Grant MDA-904-H-2034

where $c(\alpha, d)$ depends only on α and d. The most concrete goal of this paper is to extend partially this result to the extreme case $\alpha = d$ (see Proposition 3a). The proof of (1) used subadditivity, a technique going back to the original probabilistic analysis of the traveling salesman problem by Beardwood et al. [4], who stated that the $\alpha = 1$, $d = 2$ case of (1) could be proved by their method. This paper uses a completely different approach. We study the analog of the MST for an *infinite* set of points, the points of a Poisson process on all R^d. Since empirical distributions of i.i.d. sequences look locally like Poisson processes, we can relate their MST to the MST of the unbounded Poisson process. This approach also gives another proof of the existence of a limit distribution for degrees of vertices in the MST (Proposition 3b).

To formalize this approach, let $\mathbf{x} = (x_i)$ be a finite or countably infinite subset of R^d, $d \geqq 2$. Call $\mathbf{x}$ *nice* if

(i) $\mathbf{x}$ is *locally finite*, i.e. has only finitely many elements in bounded subsets of R^d; and

(ii) the interpoint distances $(|x_j - x_i|, i < j)$ are all distinct.

Given a pair $(x, \mathbf{x})$ with $\mathbf{x}$ nice and $x \in \mathbf{x}$ we can define trees $t_m(x, \mathbf{x})$ with vertices from $\mathbf{x}$ as follows. Let $\xi_1 = x$, and let t_1 be the single vertex ξ_1. Let t_2 be the tree consisting of the vertex ξ_1 and the vertex $\xi_2 \in \mathbf{x} \backslash \{\xi_1\}$ which is closest to ξ_1 in Euclidean distance, together with the edge (straight line segment) connecting ξ_1 and ξ_2. Inductively define $t_m = t_m(x, \mathbf{x})$ to be t_{m-1} together with a new edge (ξ_{j_m}, ξ_m), where $j_m \leqq m - 1$ and $\xi_m \in \mathbf{x} \backslash \{\xi_1, \ldots, \xi_{m-1}\}$ are chosen so that the edge-length $|\xi_m - \xi_{j_m}|$ is minimal (over all possible edges connecting t_{m-1} to $\mathbf{x} \backslash t_{m-1}$.) In the finite case where $n = |\mathbf{x}| < \infty$, this procedure terminates with the tree $t_n(x, \mathbf{x})$. Here we have described a variant of the well-known greedy algorithm for constructing the Euclidean MST (see Sect. 4). That is, in the finite case $t(x, \mathbf{x})$ is the *unique* Euclidean MST on vertices $\mathbf{x}$, and so in particular it does not depend on the choice of the starting vertex x.

What happens when $\mathbf{x}$ is *infinite* is less simple. In the infinite case, write $t_\infty(x, \mathbf{x}) = \cup_n t_n(x, \mathbf{x})$.

Lemma 1 *Let $g = g(\mathbf{x})$ be the graph on an infinite nice vertex-set $\mathbf{x}$ defined by taking (x_1, x_2) as an edge in g if it is an edge in either $t_\infty(x_1, \mathbf{x})$ or $t_\infty(x_2, \mathbf{x})$. Then the graph g is a forest and each component of g is an infinite tree.*

This lemma is proved in Sect. 2. An equivalent definition of g is given in Lemma 12.

The central character of the paper is the random tree $\mathscr{T}$ defined as follows. Take a Poisson point process $\mathscr{N} = \{\eta_i\}$ of rate 1 in R^d, $d \geqq 2$. Let $\mathscr{N}^0 = \mathscr{N} \cup \{0\}$. Throwing away a null set, $\mathscr{N}^0$ is a random nice subset of R^d. Let $\mathscr{G} = g(\mathscr{N}^0)$ be the forest constructed via Lemma 1.

It is natural to conjecture that $\mathscr{G}$ is in fact a.s. a *tree*, but this seems to be related to deep issues in continuum percolation (Sect. 5). Here we finesse the issue by writing $\mathscr{T}$ for the component of $\mathscr{G}$ containing the vertex 0. The next lemma records some simple facts about $\mathscr{T}$. This lemma and Proposition 3 will be proved in Sects. 2 and 3.

Lemma 2 *Let D be the degree of vertex 0 in $\mathscr{T}$. Let $L_1, \ldots, L_D$ be the lengths of the edges of $\mathscr{T}$ incident at 0. Then*

(a) $D \leqq b_d$, *a constant.*

(b) $ED = 2$.

(c) $l_d \equiv \frac{1}{2} \sum_i E L_i^d < \infty$.

Now let $\mathcal{N}_n$ denote the point process consisting of n points $(\eta_i: 1 \leqq i \leqq n)$ that are independent and uniformly distributed on the unit cube $[0, 1]^d$. Throwing away a null set, $\mathcal{N}_n$ is a nice subset of R^d. Let $\mathcal{S}_n = t_n(\eta_1, \mathcal{N}_n)$ be the MST on these n vertices. It is intuitive that $\mathcal{S}_n$, suitably rescaled, converges locally to $\mathcal{G}$. This idea is formalized in Proposition 9 and leads to a proof of our main result.

Proposition 3 (a) *Let $(|e_i|)$ be the lengths of the edges of $\mathcal{S}_n$. Then*

$$\sum_1^{n-1} |e_i|^d \xrightarrow{L^2} l_d \quad as \quad n \to \infty .$$

(b) *Let $\Delta_{n,i}$ be the proportion of vertices of $\mathcal{S}_n$ with degree i. Then for each i*

$$E\Delta_{n,i} \to P(D = i) \quad as \quad n \to \infty .$$

Discussion 1. This idea of getting asymptotics by considering a kind of MST on the Poisson limit process is rather natural to modern theoretical probabilists. The purpose of this paper is to provide the details, which are slightly less obvious than the first author thought initially. Additional motivation was provided by the emergence of Conjecture 13 below, which seems both interesting and difficult. Our results in themselves do not provide useful numerical bounds for the limiting constants. In work complementary to the present paper, Avram and Bertsimas [3] give a unified treatment of the Euclidean model and the i.i.d. model (see below) and discuss numerical bounds; they do not explicitly use the Poisson limit process. As remarked in Steele [13], for $d = 2$ the existence of a limit in (a) was conjectured by Robert Bland on the basis of simulations. Indeed, in this case there is an absolute bound on $\sum_{e \in t} |e|^2$ regardless of n or the positions of the n points in $[0, 1]^2$.

2. In the setting of (1), where the MST is built from i.i.d. points with some density with compact support K, it seems intuitively clear that

$$\sum_1^{n-1} |e_i|^d \xrightarrow{L^2} l_d |K| \quad as \quad n \to \infty$$

where $|K|$ is the d-dimensional Lebesgue measure of K. But we have not attempted to write out the details.

3. A simpler "i.i.d. model" for MSTs discards the geometry of R^d and supposes that the inter-point distances are i.i.d. with some distribution ξ. The analog of part (a) of Proposition 3 has been well-studied by combinatorial methods in the case where ξ has uniform distribution. See Timofeev [15], Avram and Bertsimas [3] and Aldous [2] for recent results on more general distributions.

4. The existence of limits in (b) was proved by different methods in Steele et al. [14], who also proved a.s. convergence. The same question in the i.i.d. model was solved explicitly in Aldous [2], using "limit process" arguments in the spirit of this paper.

5. Proving central limit theorems in this area seems technically difficult. Ramey [12] shows that the CLT for total length of the Euclidean MST can be reduced to a technical "conditional independence" conjecture.

2 Technicalities

Let us first record one simple fact.

Lemma 4 *There is a bound b_d on the degree of any vertex in any Euclidean MST in R^d.*

Proof. An easy application of the triangle inequality shows that (in any dimension) two edges of a MST meeting at a vertex cannot make an angle of less than 60°.

We now work toward the proof of Lemma 1. Let c_l be the graph with vertex-set $\mathbf{x}$ and such that (x_i, x_j) is an edge of c_l iff $|x_j - x_i| < l$. Write $t_\infty(x)$ for $t_\infty(x, \mathbf{x})$.

Lemma 5 *If (y_1, y_2) is an edge of $t_\infty(x)$ for some $x \in \mathbf{x}$, then it is also an edge of either $t_\infty(y_1)$ or $t_\infty(y_2)$.*

Proof. In the construction of $t_\infty(x)$ with vertices $(x = \xi_1, \xi_2, \ldots)$, we may suppose that the edge (y_1, y_2) occurs as (ξ_i, ξ_j), say, with $i < j$ and has length l. If the edge is not in $t_\infty(\xi_i)$ then the component of c_l containing ξ_i must be infinite; but then the edge cannot appear in $t_\infty(x)$.

Proof of Lemma 1. By Lemma 5 and the definition of g, a component of g which contains a vertex x must contain all edges of $t_\infty(x)$. Thus all components of g are infinite. Next, suppose g contains a circuit $(y_1, y_2, \ldots, y_k, y_1)$ with the maximal edge-length attained by the edge $e = (y_k, y_1)$, say. Consider $t_\infty(y_1)$. This cannot contain the edge e, because the algorithm would first have added the vertices $\{y_2, \ldots, y_k\}$ which can be connected with y_1 using shorter edges. Similarly, $t_\infty(y_k)$ cannot contain edge e, and so g does not contain e.

There is a natural notion of convergence for locally finite sets $\mathbf{x}_n$, $\mathbf{x}$. We define $\mathbf{x}_n \to \mathbf{x}$ to mean: we can label $\mathbf{x}$ as $x_1, x_2, \ldots$ and $\mathbf{x}_n$ as $x_{n,1}, x_{n,2}, \ldots$ such that as $n \to \infty$

$$x_{n,i} \to x_i, \quad 0 \leqq i < \infty; \tag{2}$$

$$|\mathbf{x}_n \cap C_L| \to |\mathbf{x} \cap C_L| \tag{3}$$

for each L such that $\mathbf{x}$ has no point on the boundary of $C_L \equiv [-L, L]^d$.

Now suppose h_n and h are graphs with respective vertex-sets $\mathbf{x}_n$ and $\mathbf{x}$. Define convergence of graphs $h_n \to h$ to mean: $\mathbf{x}_n \to \mathbf{x}$, and, for each L as above, for all $n \geqq$ some $n_0(L)$, we have the two properties

$$\text{if } (x_{n,i}, x_{n,j}) \text{ is an edge of } h_n \text{ with } x_{n,i} \in C_L \text{ then } (x_i, x_j) \text{ is an edge of } h \tag{4}$$

and

$$\text{if } (x_i, x_j) \text{ is an edge of } h \text{ with } x_i \in C_L \text{ then } (x_{n,i}, x_{n,j}) \text{ is an edge of } h_n\,. \tag{5}$$

Note from this definition that convergence $h_n \to h$ and $x_{n,i} \to x_i$ implies that the degree $d(x_{n,i})$ of $x_{n,i}$ in h_n converges to the degree $d(x_i)$ of x_i in h.

Recall the trees $t_k(x, \mathbf{x})$ defined in the introduction. Define a graph $g_k(\mathbf{x})$ by: (x_i, x_j) is an edge in g_k iff it is an edge in either $t_k(x_i, \mathbf{x})$ or $t_k(x_j, \mathbf{x})$. It is clear that, for nice sets $\mathbf{x}_n$ and $\mathbf{x}$ such that $\mathbf{x}$ is infinite,

$$\text{if } \mathbf{x}_n \to \mathbf{x},\ x_n \to x, \text{ then } t_k(x_n, \mathbf{x}_n) \to t_k(x, \mathbf{x}) \text{ for each } k$$

and hence, for fixed k,

$$\text{if } \mathbf{x}_n \to \mathbf{x} \text{ then } g_k(\mathbf{x}_n) \text{ and } g_k(\mathbf{x}) \text{ satisfy (5)}\,. \tag{6}$$

(Of course if $\mathbf{x}_n$ is finite then $t_k(x_n, \mathbf{x}_n)$ and $g_k(\mathbf{x}_n)$ are defined only for $k \leqq |\mathbf{x}_n|$, but the convergence assertions above still make sense.)

Recall the definition in Lemma 1 of $g(\mathbf{x})$ for an infinite nice set $\mathbf{x}$. To make our notation consistent, write $g(\mathbf{x})$ for the MST in the finite case. Returning to the infinite case, by definition $g(\mathbf{x}) = \cup_k g_k(\mathbf{x})$, and so from (6) we obtain a *semicontinuity* result for the map $\mathbf{x} \to g(\mathbf{x})$:

$$\text{if } \mathbf{x}_n \to \mathbf{x} \quad \text{and} \quad \mathbf{x} \text{ is infinite then } g(\mathbf{x}_n) \text{ and } g(\mathbf{x}) \text{ satisfy (5)} .$$

In general we do not have continuity. As an example , write $s_n = \sum_{i=1}^{n} 1/i$ and consider the 1-dimensional set

$$\mathbf{x}_n = \{-\gamma s_n, \ldots, -\gamma s_1, s_1, \ldots, s_n\} . \tag{7}$$

For a suitable irrational $\gamma > 0$ this defines a nice set. Then $(-\gamma, 1)$ is an edge in each $g(\mathbf{x}_n)$ but not in $g(\mathbf{x}_\infty)$.

However, if $\mathbf{x}_n \to \mathbf{x}$ and (5) holds, then to establish (4) it clearly suffices to check that the number of edges with some end in C_L converges to the correct limit as $n \to \infty$. So we may summarize the discussion as follows.

Lemma 6 *Suppose $\mathbf{x}_n \to \mathbf{x}$, where $\mathbf{x}_n$ and $\mathbf{x}$ are nice and $\mathbf{x}$ is infinite.*

(a) *If $x_n \in \mathbf{x}_n$ and $x \in \mathbf{x}$ are such that $x_n \to x$ then*

$$\liminf_n d(x_n) \geqq d(x)$$

where $d(\cdot)$ denotes the degree of the vertex in $g(\mathbf{x}_n)$ or $g(\mathbf{x})$.

(b) *For each L such that $\mathbf{x}$ has no point on the boundary of C_L,*

$$\liminf_n \sum_{x_{n,i} \in C_L} d(x_{n,i}) \geqq \sum_{x_i \in C_L} d(x_i) .$$

(c) *Suppose that, for each L such that $\mathbf{x}$ has no point on the boundary of C_L,*

$$\sum_{x_{n,i} \in C_L} d(x_{n,i}) \to \sum_{x_i \in C_L} d(x_i) \quad \text{as } n \to \infty .$$

Then $g(\mathbf{x}_n) \to g(\mathbf{x})$.

A (simple) *point process* is just a random locally finite set of points in R^d. The notion of convergence in distribution (i.e. weak convergence) of point processes is discussed in e.g. Daley and Vere-Jones [6] Sect. 9.1. It is straightforward to show (c.f. [6] exercise 9.1.6) that their notion of weak convergence is just the general notion of weak convergence on a metric space, applied to a metrization of the convergence of locally finite sets defined at (2, 3). We are mostly concerned with consequences of convergence in distribution, and these are most easily obtained by using the Skorohod representation theorem ([7] Theorem 3.1.8). That theorem says

$$\mathcal{N}_n \xrightarrow{d} \mathcal{N}_\infty \text{ iff we can choose versions } \mathcal{N}'_n \text{ of } \mathcal{N}_n \quad (n = 1, 2, \ldots, \infty)$$

$$\text{such that } \mathcal{N}'_n(\omega) \to \mathcal{N}'_\infty(\omega) \quad \text{a.s.} \tag{8}$$

Saying $\mathcal{N}'_n$ *is a version of* $\mathcal{N}_n$ means they have the same distribution. And the almost sure convergence is understood in terms of convergence of deterministic locally finite sets in (2, 3). Similarly to (8), the previous definition of convergence of deterministic graphs can be used to define convergence in distribution of random graphs.

For our final technical lemma, let $\mathcal{N}$ be a nice stationary point process, with intensity $0 < \rho < \infty$. For $\xi \in \mathcal{N}$ let $d(\xi)$ be the degree of ξ in $g(\mathcal{N})$. For each $x \in R^d$ let $\mathcal{N}^x$ have the *Palm distribution*, i.e. the conditional distribution of $\mathcal{N}$ given there exists a point of $\mathcal{N}$ at x. Let D have the distribution of $d(x)$, the degree of x in $g(\mathcal{N}^x)$, and note that by stationarity this distribution does not depend on x.

Lemma 7 *If $\mathcal{N}$ is a nice stationary ergodic point process and D is defined as above, then $ED = 2$.*

Proof. Recall $C_L = [-L, L]^d$. For fixed L an elementary argument gives the following deterministic identity, for realizations with no point on the boundary of C_L.

$$\sum_{\xi \in C_L \cap \mathcal{N}} (d(\xi) - 2) = B_L - 2F_L \tag{9}$$

where B_L is the number of edges of $g(\mathcal{N})$ with one endpoint inside C_L and the other endpoint outside C_L, and where F_L is the number of components of the forest on $\mathcal{N} \cap C_L$ obtained by taking as edges those edges of $g(\mathcal{N})$ which join two points in C_L.

Now

$$E|C_L \cap \mathcal{N}| = \rho(2L)^d$$

and, using the fundamental property of Palm distributions,

$$E \sum_{\xi \in \mathcal{N} \cap C_L} d(\xi) = (2L)^d \rho ED \, .$$

Since $F_L \leqq B_L$, (9) will imply the lemma if we show that $EB_L = o(L^d)$. An elementary argument (suggested by the referee) goes as follows. Let $D_r(x)$ be the number of edges in $g(\mathcal{N}^x)$ of the form (x, x_i) with $|x_i - x| \geqq r$. Then $ED_r(x)$ does not depend on x. Now B_L counts edges crossing into C_L, so by considering separately those edge with endpoint in $C_L \setminus C_{L-r}$ and those with endpoint in C_{L-r}, and applying Lemma 4 to the former,

$$EB_L \leqq \rho b_d |C_L \setminus C_{L-r}| + \rho(2(L-r))^d ED_r(0) \, .$$

So $\limsup_{L \to \infty} EB_L/(2L)^d \leqq \rho ED_r(0)$, and letting $r \to \infty$ the bound tends to 0.

3 Convergence of MSTs

As in Sect. 1, let $\mathcal{N}_n$ denote the point process consisting of n points $(\eta_i : 1 \leqq i \leqq n)$ which are independent and uniform on the unit cube $[0, 1]^d$. For each n let

$$\mathcal{N}_n^* = \{n^{1/d}(\eta_i - \eta_1) : 1 \leqq i \leqq n\} \, .$$

Thus, $\mathcal{N}_n^*$ is the scatter of n points, centered at one of those points and rescaled to have intensity 1 on the rescaled cube. Let $\mathcal{N}^0 = \mathcal{N} \cup \{0\}$, where $\mathcal{N}$ is the Poisson process of intensity 1. Equivalently, one could say that $\mathcal{N}^0$ has the Palm distribution of $\mathcal{N}$, given that it contains a point at 0.

Lemma 8 $\mathcal{N}_n^* \xrightarrow{d} \mathcal{N}^0$ *as* $n \to \infty$.

Proof. This is one of those "obvious" facts for which it is hard to cite a reference. In brief, given that η_1 is not within $Ln^{-1/d}$ of the boundary of the unit cube, $|\mathcal{N}_n^* \cap C_L \backslash \{0\}|$ has Binomial $(n-1, (2L)^d/n)$ distribution, and the points of $\mathcal{N}_n^* \cap C_L \backslash \{0\}$ are i.i.d. uniform on C_L. And of course $|\mathcal{N}^0 \cap C_L \backslash \{0\}|$ has Poisson $((2L)^d)$ distribution, and the points of $\mathcal{N}^0 \cap C_L \backslash \{0\}$ are i.i.d. uniform on C_L. Then the assertion of the Lemma follows from Binomial convergence to Poisson.

Write $g(\mathcal{N}_n^*)$ for the MST on $\mathcal{N}_n^*$, so $g(\mathcal{N}_n^*)$ is just a rescaling of the MST $\mathcal{S}_n$ on $\mathcal{N}_n$.

Proposition 9 (a) $g(\mathcal{N}_n^*) \xrightarrow{d} g(\mathcal{N}^0)$.

(b) *Let* $D_{n,n}$ *be the degree of* η_1 *in* $\mathcal{S}_n$, *and let* D *be the degree of* 0 *in* $g(\mathcal{N}^0)$. *Then* $D_{n,n} \xrightarrow{d} D$.

Proof. We shall first show

$$\limsup ED_{n,n} \leqq ED = 2 . \tag{10}$$

We know $ED = 2$ by Lemma 7. Any tree on n vertices has exactly $n-1$ edges, so the average degree of the vertices is $2 - 2/n$. It follows that $ED_{n,n} = 2 - 2/n$, by exchangeability of $(\eta_1, \ldots, \eta_n)$. This establishes (10).

The reader with a lot of intuition about weak convergence and point processes will now see that the Proposition follows from Lemma 8 and the deterministic convergence facts in Lemma 6. To give the details, we start with an integration fact.

Lemma 10 *If non-negative r.v.'s* $Y_1, Y_2, \ldots, Y$ *satisfy*

$$\liminf_n Y_n \geqq Y \text{ a.s.}, \quad EY_n \to EY < \infty$$

then $Y_n \xrightarrow{L_1} Y$ *and hence in distribution.*

Proof. Write

$$Y_n - Y = (Y_n - Y)^+ - (Y_n - Y)^- . \tag{11}$$

By hypotheses and dominated convergence $E(Y_n - Y)^- \to 0$. Then using the second hypothesis and (11) we see $E(Y_n - Y)^+ \to 0$. So $E|Y_n - Y| = E(Y_n - Y)^+ + E(Y_n - Y)^- \to 0$.

The point is that Lemma 8 and the Skorokhod representation theorem (8) allow us to use versions of $\mathcal{N}_n^*$ and $\mathcal{N}^0$ such that $\mathcal{N}_n^* \to \mathcal{N}^0$ a.s., where we suppress the 's, and it suffices to prove the assertions of the Proposition for these versions. Now Lemma 6(a) implies that for these versions $\liminf_n D_{n,n} \geqq D$ a.s.. In view of (10) we

can apply Lemma 10 to show $D_{n,n} \xrightarrow{L_1} D$ and hence $D_{n,n} \xrightarrow{d} D$, establishing assertion (b) of the Proposition.

To work toward part (a), fix L and define

$$Y_n = \sum_{\xi \in \mathcal{N}_n^* \cap C_L} d(\xi) = \sum_{i=1}^{n} d_n(\eta_i) 1_{(\eta_i - \eta_1 \in C_{Ln^{-1/d}})}$$

where $d_n(\cdot)$ means "degree in $g(\mathcal{N}_n^*)$". Using exchangeability of $(\eta_1, \ldots, \eta_n)$,

$$EY_n = E \sum_{i=1}^{n} d_n(\eta_1) 1_{(\eta_1 - \eta_i \in C_{Ln^{-1/d}})} = E|\mathcal{N}_n^* \cap C_L| D_{n,n} .$$

Defining

$$Y = \sum_{\xi \in \mathcal{N}^0 \cap C_L} d(\xi)$$

a similar argument using independence properties of the Poisson process gives

$$EY = E|\mathcal{N}^0 \cap C_L| D .$$

As before, take versions of the point processes which converge a.s., so

$$|\mathcal{N}_n^* \cap C_L| \to |\mathcal{N}^0 \cap C_L| \quad \text{a.s.}$$

With these versions we already proved that $D_{n,n} \xrightarrow{L_1} D$, so using the degree bound in Lemma 4 we see

$$|\mathcal{N}_n^* \cap C_L| D_{n,n} \xrightarrow{L_1} |\mathcal{N}^0 \cap C_L| D$$

and hence $EY_n \to EY$. But Lemma 6(b) implies that these versions satisfy $\liminf_n Y_n \geqq Y$ a.s., so Lemma 10 shows these versions satisfy $Y_n \xrightarrow{L_1} Y$. So far we have assumed that L is fixed. But now we can argue that every increasing sequence of integers n contains a subsequence (n_m) for which

$$\sum_{\xi \in \mathcal{N}_{n_m}^* \cap C_L} d(\xi) \to \sum_{\xi \in \mathcal{N}^0 \cap C_L} d(\xi) \quad \text{a.s., for each rational } L . \tag{12}$$

A realization sequence satisfying (12) must satisfy the hypothesis of Lemma 6(c), and hence satisfy its conclusion $g(\mathcal{N}_{n_m}^*) \to g(\mathcal{N}^0)$. This "subsequence argument" implies that the versions satisfy $g(\mathcal{N}_n^*) \xrightarrow{d} g(\mathcal{N}^0)$ as $n \to \infty$ through all integers. This completes the proof of Proposition 9.

Note that part (b) of Proposition 3 is the assertion $P(D_{n,n} = i) \to P(D = i)$, so this is established by (b) above. Of the results stated in Sect. 1, it remains to prove Lemma 2(c) and Proposition 3(a). Write $(L_{n,j}(\eta_i); j = 1, 2, \ldots)$ for the lengths of the edges of $\mathcal{S}_n$ at η_i. Then

$$Q_n \equiv \sum_{1}^{n-1} |e_i|^d = \sum_{i=1}^{n} \frac{1}{2} \sum_j L_{n,j}^d(\eta_i)$$

and so

$$EQ_n = \frac{n}{2} E \sum_j L_{n,j}^d(\eta_1)$$

Proposition 9 implies

$$(n^{1/d} L_{n,j}(\eta_1);\ j \geq 1) \xrightarrow{d} (L_j;\ j \geq 1)$$

and then the fact that $EQ_n \to l_d < \infty$ will follow from part (a) of the Lemma below.

Lemma 11 (a) $\{n \sum_j L_{n,j}^d(\eta_1);\ n \geq 1\}$ *is uniformly integrable.*

(b) $\{nE \sum_i |e_i|^{2d};\ n \geq 1\}$ *is bounded.*

Proof. Fix $x, y \in [0,1]^d$ and define

$$r = r(x,y) = |x - y|$$

$$A(x,y) = \{z \in [0,1]^d : |z - x| < r, |z - y| < r\} .$$

It is easy to see

$$\text{volume}(A(x,y)) \geq a_d r^d$$

where a_d is a constant. If (x, y) is an edge in the MST on some subset $\mathbf{x} \subset [0,1]^d$, then $A(x, y)$ cannot contain any points of $\mathbf{x}$, by considering the greedy algorithm. Applied to $\mathcal{N}_n$ we obtain

$$P((x,y) \text{ is an edge of } \mathcal{S}_n | x, y \in \mathcal{N}_n) \leq (1 - a_d r^d)^{n-2}$$
$$\leq \exp(-(n-2)a_d r^d) .$$

Now let x and r be given. Conditioning on $\eta_1 = x$ and integrating over possible locations y of other points with $|x - y| \in [r, r + dr]$,

$$\sum_j P(L_{n,j}(\eta_1) \in [r, r + dr]) \leq \exp(-(n-2)a_d r^d)(n-1)s_d r^{d-1}\, dr \qquad (13)$$

where s_d is the surface area of the unit sphere in R^d. Setting $l = nr^d$, a little algebra gives

$$\sum_j E n L_{n,j}^d(\eta_1) I(n L_{n,j}^d(\eta_1) \in [l, l + dl]) \leq \exp(-(1 - 2/n)a_d l)(1 - 1/n)d^{-1} s_d l\, dl$$
$$\leq \exp(-(1/3)a_d l) d^{-1} s_d l\, dl, \quad n \geq 3 .$$

The bound has finite integral over $0 < l < \infty$, and so for fixed j the sequence $(nL_{n,j}^d(\eta_1);\ n \geq 1)$ is uniformly integrable. Then (a) follows using the degree bound in Lemma 4. Similarly, setting $u = n^2 r^{2d}$ in (13) leads to

$$\sum_j E n^2 L_{n,j}^{2d}(\eta_1) I(n^2 L_{n,j}^{2d}(\eta_1) \in [u, u + du])$$
$$\leq \exp(-(1 - 2/n)a_d u^{1/2})(1 - 1/n)(2d)^{-1} s_d u^{1/2}\, du$$
$$\leq \exp(-(1/3)a_d u^{1/2})(2d)^{-1} s_d u^{1/2}\, du, \quad n \geq 3 .$$

Again the bound has finite integral over $0 < u < \infty$, and so the sequence $n^2 E \sum_j L_{n,j}^{2d}(\eta_1)$ is bounded in n. But by exchangeability of $(\eta_1, \ldots, \eta_n)$, $n^2 E \sum_j L_{n,j}^{2d}(\eta_1)$ is twice the quantity in part (b) of the Lemma.

To complete the proof of Proposition 3(a), it is enough to show that $\mathrm{var}(Q_n) \to 0$. By ([13], p. 1778 top),

$$\mathrm{var}(Q_{n-1}) \leqq E \sum_{i=1}^{n} \min_j L_{n,j}^{2d}(\eta_i) + 2^{2d} E \sum_{i=1}^{n} \left(\sum_j L_{n,j}^{d}(\eta_i) \right)^2 .$$

Using Schwarz's inequality and Lemma 4, the final $(\cdot)^2$ term is bounded by $b_d \sum_j L_{n,j}^{2d}(\eta_i)$, so

$$\mathrm{var}(Q_{n-1}) \leqq (1 + b_d 2^{2d}) E \sum_i \sum_j L_{n,j}^{2d}(\eta_i)$$

$$= 2(1 + b_d 2^{2d}) E \sum_i |e_i|^{2d} .$$

This last expression is $O(1/n)$ by Lemma 11(b).

4 Greedy algorithms and MST

Here is an alternate description of the graph $g(\mathbf{x})$ defined in Lemma 1.

Lemma 12 *Let* $\mathbf{x}$ *be a nice infinite set. Let* c_l *be the graph on vertices* $\mathbf{x}$ *containing as edges all pairs* (x_i, x_j) *with* $|x_j - x_i| < l$. *Define* $g^*(\mathbf{x})$ *by:* (x_i, x_j) *is an edge of* $g^*(\mathbf{x})$ *if* x_i *and* x_j *are in different components of* $c_{|x_j - x_i|}$ *and at least one of these components is finite.*

Then $g^*(\mathbf{x}) = g(\mathbf{x})$.

In example (7) the edge $(-\gamma, 1)$ is not in $g(\mathbf{x})$, although the endpoints $-\gamma$ and 1 are in different (infinite) components of $c_{1+\gamma}$. This shows that the condition "at least one of these components is finite" is needed.

Proof. If (x_1, x_2) is an edge of $g(\mathbf{x})$ then by definition it is an edge of $t_\infty(x_1, \mathbf{x})$, say, and so is an edge of $t_n(x_1, \mathbf{x})$ for some minimal n. So x_1 and x_2 are in different components of $c_{|x_1 - x_2|}$, and the component containing x_1 has exactly $n-1$ vertices. The converse argument is identical.

For completeness, let us discuss briefly the connection between our definitions, aimed at the infinite case, and the usual greedy algorithm for constructing a Euclidean MST on a finite set $\mathbf{x}$. This is a specialization of Kruskal's algorithm (e.g. Chartrand and Lesniak [5] p. 71) for constructing a minimum-weight spanning tree in a finite graph with positive edge-weights (thus the geometry of R^d, and even the triangle inequality for distances, are irrelevant).

The greedy algorithm. Construct graphs f_k, $1 \leqq k \leqq n = |\mathbf{x}|$ on vertex-set $\mathbf{x}$ by

f_1 has no edges

f_k is f_{k-1} plus an extra edge (x_i, x_j) chosen such that $|x_j - x_i|$ is minimal over all possible edges with end-points in different components of f_{k-1}.

Then $f_n = \hat{t}(\mathbf{x})$, say, is the MST.

For finite $\mathbf{x}$, it is easy to see that $\hat{t}(x)$ is the same as the tree $t_n(x, \mathbf{x})$ defined in the introduction, and is the same as the graph $g^*(\mathbf{x})$ defined in Lemma 12. But where $\mathbf{x}$ is a nice *infinite* set, the greedy algorithm will typically not make sense, because the interpoint distances $|x_j - x_i|$ will typically form a dense set. The construction of Lemma 1 seems the simplest "algorithmic" way to define the analog of the MST on an infinite set.

5 A conjecture

For the Poisson process $\mathcal{N}^0$ it seems natural to make

Conjecture 13 (a) $\mathcal{G} = g(\mathcal{N}^0)$ *is a.s. a single tree.*

(b) $\mathcal{G}$ *contains no doubly-infinite path* $\ldots, x_{-1}, x_0, x_1, \ldots$ *with distinct vertices.*

Remarks 1. Studying this conjecture soon leads one to considerations of continuum percolation. Indeed, the connection between minimal spanning trees on random points and continuum percolation has been used in a statistics context by Hartigan [9] and in an unpublished thesis of Ramey [12], and has been noted by percolation theorists (Kesten [10], Sect. 2.4). See Grimmett [8] Sect. 10.5 for a recent account of continuum percolation. The "uniqueness of infinite clusters" property implies that for Poisson processes the condition "at least one of these components is finite" in Lemma 12 is irrelevant: in other words, the behavior of Example (7) cannot occur. But another possible type of "undesirable behavior" is as follows. Consider the example in R^2

$$\mathbf{x} = \{(0, s_j), (v_j, s_j); j \geqq 1\}$$

where $s_j = \sum_{i=1}^{j} 1/i$, and $v_j \downarrow 2$ is chosen so that $\mathbf{x}$ is nice. Here $g(\mathbf{x})$ consists of two trees t_1 and t_2, where t_1 has vertices $\{(0, s_j); j \geqq 1\}$. The reason $g(\mathbf{x})$ is a forest rather than a single tree is that the infimum $\inf_{x_1 \in t_1, x_2 \in t_2} |x_2 - x_1| = 2$ is *not* attained. Proving Conjecture 13(a) is tantamount to showing this behavior cannot happen with a Poisson process.

2. If the Conjecture is true, it makes sense to define L_* as the length of the first edge in the unique infinite path in $\mathcal{G}$ starting at 0. It is then easy to see that Proposition 3(a) holds with $l_d = EL_*^d$.

3. For many classes of combinatorial trees on n vertices, there is a result: as $n \to \infty$ the tree, viewed from a uniform random vertex, converges to a limit infinite tree satisfying (b) above. See Aldous [1] for a survey of such results. Conjecture 13 would imply that the same type of result holds for Euclidean MSTs on n random points.

4. Pemantle [11] gives interesting results on the analog of Conjecture 13 for a quite different model of d-dimensional random trees (uniform random spanning trees of the lattice). In that model, the answers are different for $d \leqq 4$ and for $d \geqq 5$.

Acknowledgement. We thank the referee for carefully reading and pointing out some obscurities in the original draft.

References

1. Aldous, D.J.: Asymptotic fringe distributions for general families of random trees. Ann. Appl. Probab. **1**, 228–266 (1991)
2. Aldous, D.J.: A random tree model associated with random graphs. Random Structures and Algorithms **1**, 383–402 (1990)
3. Avram, F., Bertsimas, D.: The minimum spanning tree constant in geometric probability and under the independent model: a unified approach. Technical Report, M.I.T., 1990. Ann. Appl. Probab. (to appear)
4. Beardwood, J., Halton, H.J., Hammersley, J.M.: The shortest path through many points. Proc. Cambridge Philos. Soc. **55**, 299–327 (1959)
5. Chartrand, G., Lesniak, L.: Graphs and diagraphs. 2nd edn. Belmont, CA: Wadsworth 1986
6. Daley, D.J. Vere-Jones, D.: An introduction to the theory of point processes. Berlin Heidelberg New York: Springer 1988
7. Ethier, S.N., Kurtz, T.G.: Markov processes: characterization and convergence. New York: Wiley 1986
8. Grimmett, G.R.: Percolation. Berlin Heidelberg New York: Springer 1989
9. Hartigan, J.A.: Consistency of single linkage for high-density clusters. J. Am. Stat. Assoc. **76**, 388–394 (1981)
10. Kesten, H.: Percolation theory and first-passage percolation. Ann. Probab. **15**, 1231–1271 (1987)
11. Pemantle, R.: Choosing a spanning tree for the integer lattice uniformly. Technical Report, Cornell University, 1989. Ann. Probab. **19**, 1559–1574 (1991)
12. Ramey, D.B.: A non-parametric test of bimodality with applications to cluster analysis. PhD thesis, Yale University, 1982
13. Steele, J.M.: Growth rates of Euclidean minimal spanning trees with power weighted edges. Ann. Probab. **16**, 1767–1787 (1988)
14. Steele, J.M., Shepp, L.A., Eddy, W.F.: On the number of leaves of a Euclidean spanning tree. J. Appl. Probab. **24**, 809–826 (1987)
15. Timofeev, E.A.: On finding the expected length of a random minimal tree. Theory Probab. Appl. **33**, 361–365 (1988)