Aspects of a randomly growing cluster in $\mathbb{R}^d, d \geq 2$

Alan Frieze* Ravi Kannan† Wesley Pegden‡

Abstract

We consider a simple model of a growing cluster of points in $\mathbb{R}^d, d \geq 2$. Beginning with a point X_1 located at the origin, we generate a random sequence of points $X_1, X_2, \ldots, X_i, \ldots,$. To generate $X_i, i \geq 2$ we choose a uniform integer j in $[i-1] = \{1, 2, \ldots, i-1\}$ and then let $X_i = X_j + D_i$ where $D_i = (\delta_1, \ldots, \delta_d)$. Here the δ_j are independent copies of the Normal distribution $N(0, \sigma_i)$, where $\sigma_i = i^{-\alpha}$ for some $\alpha > 0$. We prove that for any $\alpha > 0$ the resulting point set is bounded a.s., and moreover, that the points generated look like samples from a β-dimensional subset of $\mathbb{R}^d$ from the standpoint of the minimum lengths of combinatorial structures on the point-sets, where $\beta = \min(d, 1/\alpha)$.

1 Introduction

In this short note, we study the following process: beginning with a point X_1 located at the origin, we generate a random sequence of points $X_1, X_2, \ldots, X_i, \ldots,$ in $\mathbb{R}^d$. To generate $X_i, i \geq 2$ we choose a uniform integer j in $[i-1] = \{1, 2, \ldots, i-1\}$ and then let $X_i = X_j + D_i$ where $D_i = (\delta_1, \ldots, \delta_d)$. Here the δ_j are independent copies of the Normal distribution $N(0, \sigma_i)$, where $\sigma_i = i^{-\alpha}$ for some $\alpha > 0$. Thus as more and more points are added, new points are likely to cluster around old points.

We denote the set points $\{X_1, X_2, \ldots, X_n\}$ by $\mathcal{X}_n$ and by $\mathcal{X}_\infty = \bigcup_{n=1}^\infty \mathcal{X}_n$ the set of all points generated by the process. Our first result shows that there is an exponential tail on the diameter $\rho(\mathcal{X}) = \max\{|X| : X \in \mathcal{X}\}$ of the resulting infinite cluster:

Theorem 1. *$\mathbb{P}(\rho(\mathcal{X}_\infty) \geq L) \leq e^{-L^2/600d}$ for large L.*

As a consequence, the convex hull of $\mathcal{X}_\infty$ is bounded a.s., by Borel-Cantelli applied to the events $\{\rho(\mathcal{X}_\infty) \geq L\}$ for $L = 1, 2, \ldots$. This stands in contrast to the case of a the set $\mathcal{Y}_\infty = \{Y_1, Y_2, \ldots\}$ where each Y_i is an independent standard Gaussian in $\mathbb{R}^d$, which is unbounded a.s. (and everywhere dense).

Our next theorem concerns the length L_n of the minimum spanning tree on this collection of points under the Euclidean distance. We note that the length of such minimum Euclidean structures are not just relevant from an optimization standpoint but can also be seen as a way of capturing the dimensionality of a set

*Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh PA15213, USA. Research supported in part by NSF grant DMS1952285 Email: frieze@cmu.edu

†Simons Institute for the Theory of Computing, UC Berkeley, Berkeley CA94720-2190. Email: kannan100@gmail.com

‡Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh PA15213, USA. Research supported in part by NSF grant DMS1700365.Email: wes@math.cmu.edu

or distribution: For n points chosen uniformly from a compact subset $\Omega \subseteq \mathbb{R}^d$ of dimension β (e.g., a β-dimensional manifold for integer β, or a suitably regular fractal of dimension β for non-integer β), the expected length of a spanning tree through the points is grows like $n^{1-1/\beta}$ [2]. We show that from the standpoint of the length of a minimum spanning tree, the points generated by the process we study here look like uniform samples from a subset of $\mathbb{R}^d$ of dimension $\min(d, \frac{1}{\alpha})$. $a_1, \ldots, a_6$ are absolute constants.

Theorem 2.

$$L_n \text{ satisfies } \begin{cases} a_1 n^{1-1/d} \leq \mathbb{E}(L_n) \leq a_2 n^{1-1/d} & \alpha < 1/d. \\ a_3 n^{1-\alpha} \leq \mathbb{E}(L_n) \leq a_4 n^{1-\alpha} \log^3 n & \alpha > 1/d. \\ a_5 (n/\log n)^{1-1/d} \leq \mathbb{E}(L_n) \leq a_6 n^{1-1/d} & \alpha = 1/d. \end{cases}$$

Note that the particular choice of *spanning tree* as our combinatorial structure is not so important here. Indeed if T and H are the lengths of the minimum spanning tree and Hamilton cycle on the point-set, respectively, then we have $T \leq H \leq 2T$ and so the statement of Theorem 2 holds immediately for Hamilton cycles in place of trees here. For other spanning structues like 2-factors or perfect matchings (for even n), the upper bounds in the theorem translate immediately, and the proofs of our lower bounds translate as well; in particular our proofs show not just that the lower bounds in Theorem 2 apply to the length of a spanning tree on the points in $\mathcal{X}_n$, but to the total length of any collection of edges of linear size among the points $\mathcal{X}_n$.

2 Maximum distance: proof of Theorem 1

We define a tree T_n with vertex set $\mathcal{X}_n$ and edges of the form $X_i X_{\pi(i)}$ for $i \in [n]$. Thus if X_i chooses to be "close" to X_j then we add the edge $X_i X_j$ to to T_n.

It is important to note that T_n has the structure of a *random recursive tree*, see for example Chapter 14.2 of Frieze and Karoński [1].

Let $\lambda(i) = \lambda_n(i)$ denote the level of X_i in the tree T_n, i.e. the number of edges from X_i to X_1 in T_n. Let $\mathcal{E}(i, \ell, L)$ be the event that $\lambda(i) = \ell$ and that the length of the edge from X_i to its parent in T_n is at least $\frac{L}{\ell^2 \zeta(2)}$, where $\zeta(2) = \sum_{k-1}^{\infty} k^{-2} = \pi^2/6$. If none of these events occur then every i is at distance at most $\sum_{\ell=1}^{\infty} \frac{L}{\ell^2 \zeta(2)} = L$ from the origin X_1.

In general when $i \leq m$ we have that for integers $t \leq m$,

$$\mathbb{P}(\lambda_m(i) > t) \leq \sum_{\substack{S \subseteq [m] \\ |S| = t}} \prod_{j \in S} \frac{1}{j} \leq \frac{1}{t!} \left(\sum_{j=1}^{m} \frac{1}{j} \right)^t \leq \left(\frac{e(1 + \log m)}{t} \right)^t.$$

It follows that

$$\mathbb{P}(\lambda(i) \geq 10(1 + \log m)) \leq m^{-4}, \quad \text{for } m \text{ large.} \tag{1}$$

Now we have the following inequality for $N(0, \sigma)$:

$$\mathbb{P}(N(0, \sigma) \geq x) \leq \frac{\sigma e^{-x^2/2\sigma^2}}{x(2\pi)^{1/2}}. \tag{2}$$

We see from (2) that

$$\mathbb{P}(\mathcal{E}(i, \ell, L)) \leq d\mathbb{P}\left(N(0, i^{-\alpha}) \geq \frac{L}{d^{1/2} \ell^2 \zeta(2)} \right) \leq \frac{d \ell^2 \zeta(2)}{(2\pi)^{1/2} L i^{\alpha}} \exp\left\{ -\frac{L^2 i^{2\alpha}}{2d \ell^4 \zeta(2)^2} \right\}. \tag{3}$$

(If $(\delta_1^2+\delta_2^2+\cdots+\delta_d^2)^{1/2}\geq u=L/(\ell^2\zeta(2))$ then there exists i such that $\delta_i\geq u/d^{1/2}$. We can make a small improvement by using the bound on the upper tail of the χ^2-distribution in Laurent and Massart [4].)

So for $L\leq k_1<k_2\leq n$ we have, using (1) and (3),

$$\mathbb{P}\left(\exists i\in[k_1,k_2]:\bigcup_{\ell\leq 10\log i}\mathcal{E}(i,\ell,L)\text{ occurs}\right)\leq\sum_{i=k_1}^{k_2}\left(i^{-4}+\sum_{\ell=1}^{10(1+\log i)}\exp\left\{-\frac{L^2i^{2\alpha}}{3d\ell^4\zeta(2)^2}\right\}\right)$$
$$\leq(k_2-k_1)\exp\left\{-\frac{L^2k_1^{2\alpha}}{4d(10\log k_2)^4\zeta(2)^2}\right\}+k_1^{-3}.$$

So, let $m_0=n$ and $m_t=\log^{4/\alpha}m_{t-1}$ for $t=1,2,\ldots,t_0=\min\{t:m_t\leq M\}$ where $M=e^{L^2/1000d}$. It follows that

$$\sum_{t=1}^{t_0}\mathbb{P}(\exists i\in[m_t,m_{t-1}]:\ \mathcal{E}(i,\ell,L)\text{ occurs for some }\ell)\leq\sum_{t=1}^{t_0}\left((m_{t-1}-m_t)\exp\left\{-\frac{L^2m_t^{2\alpha}}{4d(10m_t^{\alpha/4})^4\zeta(2)^2}\right\}+m_t^{-3}\right)$$
$$\leq 2\sum_{t=1}^{t_0}m_t^{-3}\leq\frac{1}{M^2}.$$

It follows that
$$\mathbb{P}(\exists i:dist(i)\geq L)\leq\frac{1}{M^2}+\frac{1}{M^3}+e^{-L^2/500d}\leq e^{-L^2/600d}.$$
The term
$$\frac{1}{M^3}+\frac{Md(10(1+\log M))^2\zeta(2)e^{-L^2/300d}}{(2\pi)^{1/2}L}\leq\frac{1}{M^3}+e^{-L^2/500d}$$
arises from applying (2) (with $\sigma=1$) and (1) to bound the probability that $\mathcal{E}(i,\ell,L)$ occurs for some $i,\ell\leq M$. This completes the proof of Theorem 1.

3 Minimum spanning tree

3.1 Upper bound

We bound the length of the recursive tree T_n. Very crudely, the cost of the first $\log n$ edges is $O(\log^2 n)$ q.s.[1] Next let $L_i=i^{-\alpha}\log^3 n$. Suppose that $\mathcal{E}(i,\ell,L_i)$ does not occur for $i\geq\log n$. Then the length of the tree produced is at most
$$O(\log^2 n)+\sum_{i=\log n}^{n}\sum_{\ell=1}^{10(1+\log i)}\frac{L_i}{\ell^2\zeta(2)}\leq O(\log^2 n)+\sum_{i=\log n}^{n}L_i.$$
We see from (3) that the probability we fail to produce a tree of the claimed size is at most

$$o(1)+\sum_{i=\log n}^{n}\sum_{\ell=1}^{10(1+\log i)}\frac{d\ell^2\zeta(2)}{\log^3 n}\exp\left\{-\frac{\log^6 n}{d\ell^4\zeta(2)^2}\right\}=o(1).$$

[1] A sequence of events $\mathcal{E}_n,n\geq 1$ occurs *quite surely* (q.s.) if $\mathbb{P}(\neg\mathcal{E}_n)=O(n^{-K})$ for any constant $K>0$.

Thus w.h.p. there is a tree of length at most

$$O(\log^2 n) + \sum_{i=\log n}^{n} \frac{\log^3 n}{i^\alpha} \leq n^{1-\alpha} \log^3 n.$$

This gives the upper bound in Theorem 2 for $\alpha \geq 1/d$. For $\alpha < 1/d$ we appeal to the fact the claimed upper bound holds for all sets of n points, in a bounded region, see for example Steele and Snyder [5]. So from Theorem 1 we can claim that the expected length of the minimum spanning tree is at most

$$c_3 n^{(d-1)/d} \int_{L=0}^{\infty} L e^{-L^2/600d} dL = O(n^{(d-1)/d}).$$

This proves the upper bound for $\alpha < 1/d$.

3.2 Lower bounds

Consider two vertices i, j whose common ancestor in the recursive tree is m. Then we have

$$\mathbb{P}(|X_i - X_j| \leq \delta) \leq \mathbb{P}(|N(0, m^{-\alpha}) - N(0, m^{-\alpha})| \leq \delta)^d = \mathbb{P}(|N(0, 2m^{-\alpha})| \leq \delta)^d = O((\delta m^\alpha)^d).$$

Now in general, the expected number of pairs i, j with common ancestor m is at most $2n^2/m^2$, see equation (7) in Section 3.2.1. So, if Z_δ denotes the number of pairs of vertices at distance at most δ, then for some constants C_1, C_2,

$$\mathbb{E}(Z_\delta) \leq C_1 n^2 \delta^d \sum_{m=1}^{n} m^{\alpha d - 2} \leq C_2 n^2 \delta^d \times \begin{cases} 1 & \alpha < 1/d \\ n^{\alpha d - 1} & \alpha > 1/d \\ \log n & \alpha = 1/d. \end{cases}$$

If $\alpha < 1/d$ then we can put $\delta = \varepsilon n^{-1/d}$ for small $\varepsilon > 0$ and see that the expected number of pairs i, j at distance at most δ is at most $C_2 \varepsilon^d n$. In which case the expected length of the minimum spanning tree is at least

$$((n-1) - C_2 \varepsilon^d n)\delta \geq c_1 n^{1-1/d} \tag{4}$$

for constant c_1.

If $\alpha > 1/d$ then we can put $\delta = \varepsilon n^{-\alpha}$ and see that the expected number of pairs i, j at distance at most δ is at most $C_2 \varepsilon^d n$. In which case the expected length of the minimum spanning tree is at least

$$((n-1) - C_2 \varepsilon^d n)\delta \geq c_2 n^{1-\alpha} \tag{5}$$

for constant c_2.

If $\alpha = 1/d$ we put $\delta = \varepsilon (n \log n)^{-1/d}$ and see that the expected number of pairs i, j at distance at most δ is at most $2C_2 \varepsilon^d n$. In which case the expected length of the minimum spanning tree is at least

$$((n-1) - 2C_2 \varepsilon^d n)\delta \geq c_3 (n/\log n)^{1-1/d} \tag{6}$$

for constant c_3.

This completes the proof of Theorem 2. (Note that (4), (5), (6) show that the lower bounds in Theorem 2 apply to any set of $\Omega(n)$ edges.)

3.2.1 Polya-Eggenburger Urn

In the Polya-Eggenburger Urn with parameters $W_0 = 1, B_0 = m - 1, \tau_0 = W_0 + B_0, s = 1$ we start with an urn containing W_0 white balls, B_0 blue balls. At each round we choose a ball at random and replace it and then add s balls of the same color to the urn. For us, the balls are the vertices of the tree. The white balls are the descendants of vertex m. Let W_n denote the number of white balls in the urn after n rounds. Then Corollary 5.1.1 of [3] states

$$\mathbb{E}(W_n) = \frac{W_0}{\tau_0} sn + W_0 \text{ and } \mathbb{VAR}(W_n) = \frac{W_0 B_0 s^2 n(sn + \tau_0)}{\tau_0^2(\tau_0 + s)}.$$

Plugging in our values, we get $\mathbb{E}(W_{n-m}) = (n - m)/m + 1 = n/m$ and

$$\mathbb{E}(W_{n-m}^2) = \mathbb{VAR}(W_{n-m}) + \mathbb{E}(W_{n-m})^2 = \frac{(m-1)n(n-m)}{m^2(m+1)} + \left(\frac{n}{m}\right)^2 \leq \frac{2n^2}{m^2}. \tag{7}$$

4 Summary

We have introduced a new model of a point process and have proved bounds on its spread and the cost of the minimum spanning tree through the points. We could have considered starting the process with $k > 1$ points placed arbitrarily. This would involve k trees with sizes determined by the Polya-Eggenburger model and it is not hard to see that our two theorems are still valid. It might be of some interest to try and remove the polylog factors from Theorem 2. Maybe also, one could try other sequences of standard deviation, other than $i^{-\alpha}$.

One natural question to ask, is as to what happens when $\alpha = 0$, i.e. when the δ_j in the definition of the X_i are $N(0, 1)$. In this case Theorem 1 fails. We know that w.h.p. the depth of T_n is $\Omega(\log n)$. In which case, the distance of leaves in T_n from the root X_1 are bounded below by the sum of $\Omega(\log n)$ standard normals and so they will w.h.p. be $\Omega(\log n)$ from X_1.

References

[1] A.M. Frieze and M. Karoński, Introduction to random Graphs, Cambridge University Press, 2015.

[2] A.M. Frieze and Wesley Pegden, The bright side of simple heuristics for the TSP, *The Electronic Journal of Combinatorics* (2024).

[3] H. Mahmoud, Polya Urn Models and Connections to Random Trees: A Review, *Journal of the Iranian Statistical Society*, 2 (2003) 53-114.

[4] B. Laurent, P. Massart, Adaptive estimation of a quadratic functional by model selection, *Annals of Statistics* 28 (200) 1302-1338.

[5] J. Steele and T. Snyder(1989), Worst-case growth rates of some classical problems of combinatorial optimization, *SIAM Journal on Computing* 18 (1989) 278-287.