FRACTIONAL BROWNIAN MOTION, HEAVY TAILS, AND COALESCENCE: A RANDOM WALK APPROACH

A Dissertation
Presented to
The Academic Faculty

By

Joshua Marsh

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
College of Sciences
School of Mathematics

Georgia Institute of Technology

May 2025

FRACTIONAL BROWNIAN MOTION, HEAVY TAILS, AND COALESCENCE: A RANDOM WALK APPROACH

Thesis committee:

Dr. Michael Damron
School of Mathematics
Georgia Institute of Technology

Dr. Zongchen Chen
School of Computer Science
Georgia Institute of Technology

Dr. Christian Houdré
School of Mathematics
Georgia Institute of Technology

Dr. Wenjing Liao
School of Mathematics
Georgia Institute of Technology

Dr. Mayya Zhilova
School of Mathematics
Georgia Institute of Technology

Date approved: April 7, 2025

Life is a journey, not a destination.

Ralph Waldo Emerson

What Emerson says about life applies to education—a PhD is a journey, not a degree. For her steady support on this journey, I dedicate this thesis to my wife.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Random walks are ubiquitous. They are of interest in and of themselves and as the driving randomness controlling other processes. Vanilla random walks with finite variance and independent increments have Brownian motion scaling limits, but more exotic random walks can give fractional Brownian motion, which has dependent increments. Nonetheless, these random walks with dependent increments can be generated using ancillary random walks which do have independent increments, provided that the ancillary random walks have increments following certain heavy-tailed distributions with infinite variance.

In this thesis, we extend works [9, 14, 17] which generate fractional Brownian motion using one-dimensional random walks to higher dimensions, generating a multidimensional family of fractional Brownian motions.

Key to the constructions is the coalescence of transient random walks. The Green's function of random walks describes how the coalescence probability decays as the starting points of the random walks move further apart. The self-similarity of random sums in the domain of attraction of a stable law, combined with a local limit theorem to give quantitative bounds, gives another perspective on the large-scale coalescence behavior.

In another chapter, we investigate a percolation-style model in which each vertex of $\mathbb{Z}^d$ is connected to its k nearest neighbors under translation-invariant edge weight distributions. In the case where the edge weights are independent and identically distributed, coalescing branching random walks naturally arise in the analysis of infinite connected components.

Using these coalescing random walks, we show that when $k \geq d + 1$, there is almost surely a unique infinite connected component, and this component contains every vertex. Along the way, we resolve a question about the recurrence of coalescing

branching random walks with a branching factor of 2 on the lattice $\mathbb{Z}^d$: almost surely, the origin is occupied infinitely often.

Finally, we step away from random walks and turn to physics-informed neural networks (PINNs). Neural networks are parametric functions with strong approximation capabilities. Using a technique called automatic differentiation, arbitrary derivatives of neural networks can be computed. A PDE can be viewed as a constraint on the derivatives of a function. If we minimize the degree to which this constraint is not satisfied, we expect that the function converges to a solution of the PDE.

For a PDE defined on a domain Ω, the PINN approach is to use automatic differentiation to give arbitrary derivatives of a neural network evaluated at points $x \in \Omega$, at which we minimize the degree to which the PDE is not satisfied. However, the choice of which points to evaluate the derivative at has significant impact on convergence speed of the neural network and on the quality of the final trained neural network.

We present an adversarial method for selection of points. It has the advantage of requiring low computational overhead, in contrast to many established methods. Numerically, we demonstrate its effectiveness in high dimensions, where classical numerical methods often struggle.

CHAPTER 1

FRACTIONAL BROWNIAN MOTION

The contents of this chapter include joint work with Michael Damron, Paul Jung, and Wooyoung Chin.

1.1 Introduction

1.1.1 Background

Fractional Brownian motion is a generalization of standard Brownian motion which, instead of having independent increments, has correlated increments. A natural way to obtain Brownian motion is as an appropriately rescaled limit of a random walk. Hammond and Sheffield [14] provided a construction for fractional Brownian motion as a scaling limit of a discrete random walk with correlated increments. In this chapter, we provide an alternate construction, which takes place on a higher dimensional lattice.

Fractional Brownian motion is characterized by its Hurst parameter $H \in (0, 1)$, with $H = 1/2$ corresponding to standard Brownian motion. For $H > 1/2$, the increments are positively correlated, while when $H < 1/2$, they are negatively correlated. We briefly show in subsection 1.2.3 that the correlation of increments which are t apart in time decays like t^{2H-2}. In this work, only the positive correlation case is considered, i.e., $H > 1/2$. It is therefore natural to ask that the i-th and the j-th increments of a random sum have correlation which asymptotically behaves like $|j - i|^s$ for some $s \in (-1, 0)$.

Hammond and Sheffield [14] achieve this by defining ± 1-valued random variables $(X_i)_{i\in\mathbb{Z}}$ such that, after the random variables $\{X_i : i < k\}$ have been defined, X_k is obtained by choosing a random positive integer R_k and setting $X_k = X_{k-R_k}$, copying the variable R_k back. We call $k-R_k$ the ancestor of k. Then the random sum $\sum_{i=1}^n X_i$ has correlated increments, where the correlation decay depends on the distribution of R_k. Notably, correlation between X_i and X_j does not just arise because one may have copied the increment of the other. Rather, they may have chains of ancestors which intersect. So $\mathrm{Cov}[X_i, X_j]$ decays slower, as a function of $|i-j|$, than the tail of R_k decays.

One may view this process as a collection of coalescing random walks, with one random walk started at each vertex of $\mathbb{Z}$, where the increment distribution of the random walk is the distribution of $-R_k$. If the random walks started at i and j intersect (and thus coalesce), then $X_i = X_j$. Otherwise, X_i is independent of X_j.

While the correlation structure is perhaps the key feature of the model, there are many random processes with a given covariance function. Only when the process is jointly Gaussian is it a fractional Brownian motion. It is our experience that proving asymptotic Gaussianity presents more difficulty than showing that the covariance structure is asymptotically that of a fractional Brownian motion. Hammond and Sheffield [14] use a martingale central limit theorem to prove asymptotic Gaussianity.

The Hammond-Sheffield model can be generalized. Finding a model which can achieve negative correlation ($H < 1/2$) would be particularly interesting. We do not do this, although we do discuss how the obvious modifications to generate negative correlation within our proposed model fail to give $H < 1/2$ (instead, it appears that the modifications would generate standard Brownian motion).

Igelbrink and Wakolbinger [17] observe that the Hammond-Sheffield model is, at its core, a model to generate a random partition of $\mathbb{Z}$. Each block of the partition is assigned an i.i.d. random variable, and every point within the block shares this random variable. The random variables assigned to each point in $\mathbb{Z}$ are used as increments of a sum. When rescaled, this random sum may converge to fractional

Brownian motion.

The authors present a framework for analyzing sequences of random partitions, the n-th partition being a partition of $\{1, 2, \ldots, n\}$. While Hammond and Sheffield [14] use a martingale central limit theorem to prove asymptotic Gaussianity, Igelbrink and Wakolbinger [17] make use of Stein's method. They show that it is sufficient to obtain certain asymptotic bounds on both the probability that three random points in $\{1, 2, \ldots, n\}$ are in the same partition and the corresponding probability for four points.

Another work, by Drogin [9], takes the perspective of a random voter model with long-range interactions. The (± 1-valued) voters live on $\mathbb{Z}$, and taking sums of voters again gives a fractional Brownian motion, now evolving in time with the dynamics of the voter model. Lindeberg swapping is the tool of choice for asymptotic Gaussianity.

In this chapter, we present a generalization to higher dimensions. The model is the most similar to the voter model, but with voters living on $\mathbb{Z}^d$. In contrast to the work of Drogin [9], we take sums along the time dimension instead of taking sums along the voters. To prove asymptotic Gaussianity, we use an appropriate martingale central limit theorem. A local limit theorem for convergence of random walks to a stable law arises as a key tool. We make an effort to keep the proofs intuitive, with a preference for random walk arguments over, say, Fourier arguments.

1.1.2 Model description

Let $d \geq 1$ be an integer and set $D = d + 1$. Let μ be a probability measure on $\mathbb{Z}^d$. We assume that μ is symmetric, in that $\mu(A) = \mu(-A)$, and that it is in the domain of attraction of a nondegenerate stable law on $\mathbb{R}^d$ with parameter $\alpha \in (0, 2]$. Nondegenerate means that the law is not supported on a proper hyperplane. We also assume that μ is not supported on a translation of a proper subgroup of the additive group $\mathbb{Z}^d$.

Being in the domain of attraction of a stable law with parameter α is equivalent to the existence of $b(n) = n^{1/\alpha}\ell(n)$ such that as $n \to \infty$,

$$\frac{Z_n}{b(n)} \to U$$

in distribution, where ℓ is a slowly varying function (defined below), Z_n is a sum of n independent random variables each with distribution μ, and U is a nondegenerate stable law of parameter α. We will make extensive use of $\ell(n)$. A function L from $(0, \infty)$ to $(0, \infty)$ is slowly varying if for any $c > 0$,

$$\lim_{n\to\infty} \frac{L(cn)}{L(n)} = 1.$$

The book *Regular Variation* [1] has details about the theory of slowly varying functions. For convenience, we capture the assumptions on μ and some associated functions and random variables in one place.

Assumption A. Assume that μ is a symmetric probability measure on $\mathbb{Z}^d$ which is in the domain of attraction of a nondegenerate stable law on $\mathbb{R}^d$ with parameter $\alpha \in (0, 2]$. And assume that μ is not supported on a translation of a proper subgroup of the additive group $\mathbb{Z}^d$. Call the stable law U and let g be its density. Let $(X_i)_{i\geq 1}$ be a sequence of independent random variables each with distribution μ, and set $Z_n = X_1 + \cdots + X_n$. Let $b(n) = n^{1/\alpha}\ell(n)$ be a positive function such that ℓ is slowly varying and as $n \to \infty$,

$$\frac{Z_n}{b(n)} \to U$$

in distribution.

We can define a random directed graph G_μ with vertex set $\mathbb{Z}^D$ as follows. Let $(R_v)_{v\in\mathbb{Z}^D}$ be i.i.d. $\mathbb{Z}^D$-valued random variables, with first component equal to -1 and remaining d components jointly distributed according to μ. The graph is defined by its edge set $\{\langle v, v+R_v\rangle : v \in \mathbb{Z}^D\}$, and we use the notation $x \to y$ to indicate that there is directed path from x to y. The graph can be split into its connected components, so we write $u \sim v$ if u and v are in the same component. Let Y be a fixed mean zero real-valued random variable which is bounded (most parts of the proof only require a fourth moment). Assign independently to each component a random variable whose law is that of Y. We call the number associated with each component the *color* of that component. Let Y_v be the color of the component containing the vertex v. These colors serve as the increments of some stochastic process. We are interested in sequences of partial sums such as

$$\sum_{i=1}^{n} Y_{y+ix}$$

for some fixed vectors x and y, as n varies. This is a sum along a ray of vertices started at y going in the x direction. Of particular interest is when y is 0 or x is e_1. Note that the summands are not independent, as for any two vertices, there is a positive probability that they are in the same component of the random graph. With suitable parameters, the sum can produce a fractional Brownian motion when taking an appropriately scaled limit.

The main contribution of this chapter is to show that the random graph model gives correlation decay on the order of $|i-j|^{1-d/\alpha}$, that the random sum has variance which grows like $n^{3-d/\alpha}$, and that the sum is asymptotically Gaussian. Thus, the appropriately scaled random sum converges to a fractional Brownian motion with Hurst parameter $H = \frac{1}{2}(3 - d/\alpha)$.

The boundedness assumption on the color distribution is only used for applying Hoeffding's inequality in the proof of Lemma 1.18. All other parts of the proof require only a fourth moment assumption. It seems likely that the boundedness assumption can be relaxed. However, the fourth moment assumption is used to prove concentration of the conditional variance using a second moment method, and it is not clear how hard it would be to do better than a fourth moment assumption. The work by Igelbrink and Wakolbinger [17], which shows asymptotic fractional Brownian motion for sums defined from general random partitions, also requires a finite fourth moment for the color distribution.

1.1.3 Main results

In this chapter, we treat sums along rays in the e_1 direction. Since the joint distribution of the random graph is translation invariant, there is no loss of generality in assuming that the sum is started at the vertex 0. Concretely, we work with the sum

$$S_n = \sum_{i=1}^{n} Y_{ie_1}. \tag{1.1.1}$$

Let $\sigma_n^2 = \mathrm{Var}[S_n]$. For $t \geq 0$, define the n-th rescaling of this sum as

$$G^{(n)}(t) = \frac{1}{\sigma_n} S_{\lfloor nt \rfloor}. \tag{1.1.2}$$

This is S_n rescaled so that $G^{(n)}(1)$ uses n terms of the random sum and has unit variance. We may easily modify the definition of $G^{(n)}(\cdot)$ to have continuous sample paths by, say, using linear interpolation.

The main result is that $G^{(n)}(\cdot)$ converges in distribution to a fractional Brownian motion with Hurst parameter $H = \frac{1}{2}(3 - \frac{d}{\alpha})$. This follows from the asymptotic Gaussianity and asymptotic covariance structure of $G^{(n)}(\cdot)$.

Theorem 1.1. *Assume $\alpha \in (d/2, d)$. With $H = \frac{1}{2}(3 - \frac{d}{\alpha})$, for $t \geq s \geq 0$,*

$$\lim_{n\to\infty} \mathrm{Cov}\left[G^{(n)}(t), G^{(n)}(s)\right] = \frac{1}{2}\left(t^{2H} + s^{2H} - (t-s)^{2H}\right).$$

This is the covariance of a fractional Brownian motion with Hurst parameter H.

Centered Gaussian processes are uniquely determined by their covariance structure. Since each Y_v has mean 0, $G^{(n)}(t)$ does as well, showing it's centered. The next theorem says that it is asymptotically a Gaussian process.

Theorem 1.2. *Assume $\alpha \in (d/2, d)$. For $k \geq 1$, given $\alpha_i \in \mathbb{R}$ and $t_i \geq 0$ for $1 \leq i \leq k$, the random variable*

$$\sum_{i=1}^{k} \alpha_i \, G^{(n)}(t_i)$$

converges in distribution to a Gaussian random variable as $n \to \infty$.

As a corollary of Theorem 1.1 and Theorem 1.2, we get that the processes $G^{(n)}(\cdot)$ converge to fractional Brownian motion. Theorem 1.1 is proved in section 1.3, and Theorem 1.2 is proved in section 1.4.

Since stable laws exist only for $\alpha \in (0, 2]$, the requirement $\alpha \in (d/2, d)$ imposes some restrictions on the dimension. For $d = 1$ and $d = 2$, the full range of $\alpha \in (d/2, d)$ is possible. When $d = 3$, we get the restriction $\alpha \in (3/2, 2]$. And for $d \geq 4$, there is no valid α.

1.2 Heavy tails, coalescence, and fractional Brownian motion

In this section, we summarize background results about heavy-tailed random walks in the domain of attraction of a stable law, discuss how coalescence generates a slowly decaying covariance structure, and note some properties of fractional Brownian motion. We also establish an important equation relating the probability that two vertices are in the same component to the probability that some independent random walks collide.

The notation $f(n) \sim g(n)$ will be used to indicate that two functions are asymptotically equivalent:

$$\lim_{n \to \infty} \frac{f(n)}{g(n)} = 1.$$

We also use standard big O notation for positive functions f and g: $f(n) = O(g(n))$ if $\limsup_n \frac{f(n)}{g(n)} < \infty$; $f(n) = o(g(n))$ if $\lim_n \frac{f(n)}{g(n)} = 0$; $f(n) = \Omega(g(n))$ if $g(n) = O(f(n))$; and $f(n) = \omega(g(n))$ if $g(n) = o(f(n))$. In particular, $f(n) = o(1)$ if $\lim_n f(n) = 0$ and $f(n) = O(1)$ if $f(n)$ is bounded.

1.2.1 Heavy tails

The following facts about a $\mathbb{Z}^d$-valued random walk $(Z_n)_{n\geq 0}$ started at $Z_0 = 0$ with i.i.d. μ-distributed increments in the domain of attraction of a nondegenerate stable law with parameter α are drawn from Le Gall and Rosen [22]. Assume that μ satisfies Assumption A. As stated in Assumption A, when μ is in the domain of attraction of a stable law, there is a function $b(n) = n^{1/\alpha}\ell(n)$ such that $Z_n b^{-1}(n)$ converges in distribution to a stable law U of parameter α, and $\ell(n)$ is a slowly varying function. Denote characteristic function of Z_1 by φ, and that of U by Φ. It is known that

$$\Phi(\xi) = \mathbb{E}\exp(i\xi \cdot U) = \exp\left(-|\xi|^\alpha S\left(\frac{\xi}{|\xi|}\right)\right),$$

for some continuous complex-valued function S on the sphere S^{d-1} with real part $\Re S(x) \geq a > 0$ for some real constant a and any x in S^{d-1}. And since μ is assumed to be symmetric, S is as well.

If $\alpha < d$, the random walk $(Z_n)_{n\geq 0}$ is transient. The two dimensional simple random walk corresponds to $d = 2, \alpha = 2$. It is recurrent and, in this sense, it is only barely recurrent.

To get good asymptotics for the random walks, we use a local limit theorem for stable laws. A local limit theorem says that, at a high level, the probability of the random walk occupying each vertex is asymptotically what one would expect, and the convergence happens uniformly, in that there are no spikes in the distribution. Our running assumption is that a rescaled $(Z_n)_{n\geq 0}$ converge in distribution to a stable law, but convergence in distribution does not preclude large spikes in the distribution. Convergence in distribution only requires that these spikes have $o(1)$ mass.

The following local limit theorem for sums of independent $\mathbb{Z}^d$-valued random vectors in the domain of attraction of a stable law can be found in §50 of Gnedenko and Kolmogorov [11] for the univariate case, generalized to the multivariate case in Theorem 6.1 of Rvačeva [29].

Theorem 1.3. *Suppose $\alpha \in (0,2]$ and that μ satisfies Assumption A. Let ℓ, $(Z_n)_{n\geq 1}$, and g be defined as in Assumption A. Then*

$$\lim_{n\to\infty} \sup_{x\in\mathbb{Z}^d} \left| n^{d/\alpha}\ell(n)^d\, \mathbb{P}\left(Z_n = x\right) - g\left(\frac{x}{n^{1/\alpha}\ell(n)}\right)\right| = 0.$$

In Appendix A, we prove in Proposition A.1 that the density of a stable law exists and is bounded when the stable law is not supported on a hyperplane.

1.2.2 Coalescence

We have a random walk started at each vertex of the graph. The random walks are not purely independent random walks. Instead, the random walks are such that if they collide, they coalesce. Prior to coalescence, they are independent. The probability of coalescence is critical to the computations. See Figure 1.1 for an illustration of the random walks started at 0 and ne_1.

Let's consider what happens if we sum along the ray started at 0 in the e_1 direction: $S_n = \sum_{i=1}^n Y_{ie_1}$. We would like to know the asymptotics of $\mathrm{Var}[S_n]$. To compute this, it is useful to know how $\mathbb{P}(ie_1 \sim je_1)$ depends on i and j. By translation invariance of the distribution of the random graph, $\mathbb{P}(ie_1 \sim je_1) = \mathbb{P}(|i-j|\, e_1 \sim 0)$, making $\mathbb{P}(ne_1 \sim 0)$ a key quantity of interest.

Since every vertex in the graph has out-degree 1, there is a $\mathbb{Z}^D$-valued random walk associated with each vertex given by following the directed edges. Call these the

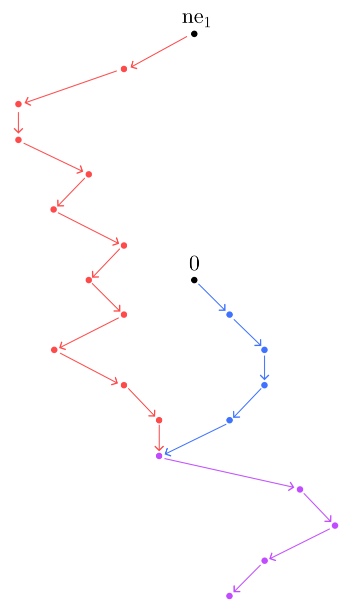

Figure 1.1: The random walks started at two vertices are independent unless they collide, at which point they coalesce. Only the random walks started at 0 and ne_1 are shown here.

ancestral random walks. The ancestral random walk always has increment -1 in the first coordinate and random increment with law μ in the remaining coordinates. We see that two vertices are in the same component exactly when the two ancestral random walks started at the vertices intersect (and thus coalesce). Prior to intersection, the ancestral random walks are independent.

Let $(Z_i)_{i\geq 0}$ and $(\tilde{Z}_i)_{i\geq 0}$ be i.i.d. random walks on $\mathbb{Z}^d$ (not $\mathbb{Z}^D$), started at 0, whose (independent) increments are distributed according to μ. The i-th step of the ancestral random walk started at $0 \in \mathbb{Z}^D$ has the same distribution as $(-i, \tilde{Z}_i)$ (i.e., the point in $\mathbb{R}^d$ whose first component is $-i$ and whose last d components are $\tilde{Z}_i$), and the i-th step of the ancestral random walk started at ne_1 has the same distribution as $(n-i, Z_i)$. Since the ancestral random walks are independent prior to intersection, the probability the ancestral random walks intersect is the same as the probability that $Z_{n+i} = \tilde{Z}_i$ for some $i \geq 0$. Another perspective is that, prior to intersection, $\tilde{Z}_i$ and Z_{n+i} have the same joint distribution as projections of the two ancestral random walks onto the last d coordinates. The ancestral random walks intersect if the projected random walks collide when the indices are n apart. This shows that

$$\mathbb{P}(ne_1 \sim 0) = \mathbb{P}(Z_{n+i} = \tilde{Z}_i \text{ for some } i \geq 0). \tag{1.2.1}$$

1.2.3 Fractional Brownian motion

Fractional Brownian motion with Hurst parameter $H \in (0,1)$ is characterized by its variance function. If B_t is a fractional Brownian motion such that B_1 has unit variance, then $\mathrm{Var}[B_t] = t^{2H}$. Because it has stationary increments, this determines the covariance structure as well: $\mathrm{Cov}[B_t, B_s] = \frac{1}{2}(t^{2H} + s^{2H} - |t-s|^{2H})$. If $H = 1/2$, this is standard Brownian motion with independent increments. We can see the correlation decay of the increments. For example, consider the correlation between $B_1 = B_1 - B_0$ and $B_{t+1} - B_t$ for $t > 1$:

$$\begin{aligned} 2\,\mathrm{Cov}[B_1, B_{t+1} - B_t] &= (1 + (t+1)^{2H} - t^{2H}) - (1 + t^{2H} - (t-1)^{2H}) \\ &= (t+1)^{2H} + (t-1)^{2H} - 2t^{2H} \\ &\sim 2H(2H-1)t^{2H-2}. \end{aligned}$$

This is positive for $H > 1/2$, negative for $H < 1/2$, and 0 when $H = 1/2$. The constructions in this chapter only give $H > 1/2$. The correlation decay suggests that we should want the increments i and j of our random sum to have correlation that decays like $|i-j|^a$ for $a \in (-1, 0)$. Indeed, as we will shortly see, this is what the constructions give.

Since fractional Brownian motion has a stationary increment distribution, it is characterized just by $\mathrm{Var}[B_t]$. Indeed, this is true in general. For this reason, Equation 1.2.2 in the following remark and the covariance of fractional Brownian motion are quite similar.

Remark 1.4. If a one-dimensional centered process $(B_t)_{t\geq 0}$ with $B_0 = 0$ has stationary increments, then its variance function $V_t = \mathrm{Var}[B_t]$ determines its covariance as long as $V_t < \infty$:

$$\begin{aligned}
\mathbb{E}B_tB_s &= \mathbb{E}B_t(B_s - B_t + B_t) \\
&= \mathbb{E}B_t^2 + \mathbb{E}B_t(B_s - B_t) \\
&= \mathbb{E}B_t^2 + \mathbb{E}(B_t - B_s + B_s)(B_s - B_t) \\
&= \mathbb{E}B_t^2 - \mathbb{E}(B_t - B_s)^2 + \mathbb{E}B_s^2 - \mathbb{E}B_sB_t,
\end{aligned}$$

and we see that

$$\mathbb{E}B_tB_s = \frac{1}{2}(V_t + V_s - V_{|t-s|}). \tag{1.2.2}$$

The stationary increment assumption is only used for the identity $\mathbb{E}(B_t - B_s)^2 = \mathbb{E}(B_{|t-s|} - B_0)^2 = V_{|t-s|}$.

1.3 Asymptotic covariance structure

In this section, we argue that, when $\alpha \in (d/2, d)$, the variance of the random sum S_n defined in Equation 1.1.1 grows like $n^{3-d/\alpha}$. Since $3 - d/\alpha \in (1, 2)$, this matches the variance of a fractional Brownian motion with Hurst parameter greater than 1/2.

1.3.1 Coalescence probabilities

By Equation 1.2.1, we can characterize the probability that two points n apart on the e_1 axis are in the same component as the probability that two independent random walks collide at indices which are n apart. The following proposition gives precise asymptotics of this probability.

Proposition 1.5. *Suppose $\alpha < d$ and that μ satisfies Assumption A. Let $(Z_i)_{i\geq 0}$ and $(\tilde{Z}_i)_{i\geq 0}$ be i.i.d. $\mathbb{Z}^d$-valued random walks started at 0 with independent μ-distributed* *increments. Let ℓ and g be as in Assumption A. Then there is a constant $C > 0$ such that*

$$\mathbb{P}(Z_{n+i} = \tilde{Z}_i \textit{ for some } i) \sim Cn^{1-d/\alpha}\ell^{-d}(n).$$

If G is the Green's function of the random walk whose increments are independent and have law equal to that of $Z_1 + \tilde{Z}_1$, we have

$$C = \frac{g(0)}{2\,G(0,0)(d/\alpha - 1)}.$$

Proof. The assumption that μ is symmetric is for convenience, so that the local limit theorem can be applied to the difference of the random walks. However, heuristics suggest that this assumption can removed.

Let $G(x, y)$ be the Green's function of the $\mathbb{Z}^d$-valued random walk with independent increments distributed like $Z_1 - \tilde{Z}_1$. Since μ is symmetric, this is the same as the distribution of $Z_1 + \tilde{Z}_1$. Then

$$G(0,0)\,\mathbb{P}(Z_{n+i} = \tilde{Z}_i \text{ for some } i \geq 0) = \sum_{i=0}^{\infty} \mathbb{P}(Z_{n+i} = \tilde{Z}_i),$$

since $\mathbb{P}(Z_{n+i} = \tilde{Z}_i \text{ for some } i)$ is the probability the random walks $(Z_{n+i})_{i\geq 0}$ and $(\tilde{Z}_i)_{i\geq 0}$ collide and $G(0,0)$ is the expected number of times the random walks collide given that they do collide (recall that two random walks collide if they occupy the same point in space at the same point in time). When $\alpha < d$, the random walks are transient, so $G(0,0)$ and the expected number of collisions are finite.

By the assumed symmetry of the random walk and the independence of the random walks, $Z_{n+i} - \tilde{Z}_i$ has the same distribution as Z_{n+2i}. Fix $\varepsilon > 0$. By Theorem 1.3, there is some N such that if $k \geq N$, then

$$k^{-d/\alpha}\ell^{-d}(k)\,(g(0) - \varepsilon) \leq \mathbb{P}(Z_k = 0) \leq k^{-d/\alpha}\ell^{-d}(k)\,(g(0) + \varepsilon)\,.$$

Since $\mathbb{P}(Z_{n+i} = \tilde{Z}_i) = \mathbb{P}(Z_{n+2i} = 0)$, for $n \geq N$, we get upper and lower bounds

$$\sum_{i=0}^{\infty} \mathbb{P}(Z_{n+i} = \tilde{Z}_i) \leq (g(0) + \varepsilon) \sum_{i=0}^{\infty} (n+2i)^{-d/\alpha} \ell^{-d}(n+2i),$$
$$\sum_{i=0}^{\infty} \mathbb{P}(Z_{n+i} = \tilde{Z}_i) \geq (g(0) - \varepsilon) \sum_{i=0}^{\infty} (n+2i)^{-d/\alpha} \ell^{-d}(n+2i).$$

It should be quite believable that

$$\sum_{i=0}^{\infty} (n+2i)^{-d/\alpha} \ell^{-d}(n+2i) \sim \int_0^{\infty} (n+2x)^{-d/\alpha} \ell^{-d}(n+2x)\, dx, \tag{1.3.1}$$

so we delay the proof of Equation 1.3.1. It is a fact (Bingham, Goldie, and Teugels [1], Proposition 1.5.10) that for $\beta < -1$, if L is slowly varying, then

$$\frac{n^{\beta+1} L(n)}{\int_n^{\infty} x^{\beta} L(x)\, dx} \to -\beta - 1 \text{ as } n \to \infty.$$

We observe that ℓ^{-d} is slowly varying:

$$\lim_{x \to \infty} \frac{\ell^{-d}(cx)}{\ell^{-d}(x)} = \left(\lim_{x \to \infty} \frac{\ell(cx)}{\ell(x)} \right)^{-d} = 1.$$

Because of the assumption $\alpha < d$, we get $-d/\alpha < -1$, so

$$\int_0^{\infty} (n+2x)^{-d/\alpha} \ell^{-d}(n+2x)\, dx \sim \frac{n^{1-d/\alpha}}{2(d/\alpha - 1)} \ell^{-d}(n).$$

Thus,

$$g(0) - \varepsilon \leq \lim_{n \to \infty} \left(\sum_{i=0}^{\infty} \mathbb{P}(Z_{n+i} = \tilde{Z}_i) \right) \left(\frac{n^{1-d/\alpha}}{2(d/\alpha - 1)} \ell^{-d}(n) \right)^{-1} \leq g(0) + \varepsilon.$$

But ε was arbitrary, so $\sum_{i=0}^{\infty} \mathbb{P}(Z_{n+i} = \tilde{Z}_i) \sim \frac{g(0)}{2(d/\alpha-1)} n^{1-d/\alpha} \ell^{-d}(n)$, and we get the desired relation

$$\mathbb{P}(Z_{n+i} = \tilde{Z}_i \text{ for some } i) \sim \frac{g(0)}{2\,G(0,0)(d/\alpha - 1)} n^{1-d/\alpha} \ell^{-d}(n).$$

To complete the proof, we show Equation 1.3.1. It is a theorem (Bingham, Goldie, and Teugels [1], Theorem 1.5.4) that if L is slowly varying and $a > 0$, then there is a non-increasing function ψ such that $x^{-a}L(x) \sim \psi(x)$. We apply this to the slowly varying function ℓ^{-d} and $a = d/\alpha$. So for each $\varepsilon > 0$, there is some n_0 such that for $x > n_0$,

$$(1-\varepsilon)\psi(x) \le x^{-d/\alpha}\ell^{-d}(x) \le (1+\varepsilon)\psi(x).$$

For $n > n_0$, we get

$$(1-\varepsilon)\sum_{i=0}^{\infty} \psi(n+2i) \le \sum_{i=0}^{\infty} (n+2i)^{-d/\alpha}\ell^{-d}(n+2i) \le (1+\varepsilon)\sum_{i=0}^{\infty} \psi(n+2i).$$

And ε was arbitrary, so we see that

$$\sum_{i=0}^{\infty} (n+2i)^{-d/\alpha}\ell^{-d}(n+2i) \sim \sum_{i=0}^{\infty} \psi(n+2i).$$

A similar argument shows that

$$\int_0^{\infty} (n+2x)^{-d/\alpha}\ell^{-d}(n+2x)\,dx \sim \int_0^{\infty} \psi(n+2x)\,dx. \tag{1.3.2}$$

Since ψ is monotone, the error in approximating $\sum_{i=0}^{\infty} \psi(n+2i)$ by the corresponding integral can be bounded by the first term of the sum. Equation 1.3.2 says that $\int_0^{\infty} \psi(n+2x)\,dx$ behaves asymptotically like a slowly varying function times $n^{1-d/\alpha}$. The first term of $\sum_{i=0}^{\infty} \psi(n+2i)$ is $\psi(n)$, and it behaves like $\psi(n) \sim n^{-d/\alpha}\ell^{-d}(n)$. This is asymptotically smaller than the growth rate of $\int_0^{\infty} \psi(n+2x)\,dx$, so the approximation error from approximating the sum $\sum_{i=1}^{\infty} \psi(n+2i)$ by the corresponding integral is asymptotically negligible, giving

$$\sum_{i=0}^{\infty} \psi(n+2i) \sim \int_0^{\infty} \psi(n+2x)\,dx.$$

Lastly, since $\sum_{i=0}^{\infty}(n+2i)^{-d/\alpha}\ell^{-d}(n+2i) \sim \sum_{i=0}^{\infty} \psi(n+2i)$ and $\int_0^{\infty}(n+2x)^{-d/\alpha}\ell^{-d}(n+2x)\,dx \sim \int_0^{\infty} \psi(n+2x)\,dx$, we conclude that Equation 1.3.1 holds, finishing the proof.

□

From this proposition, along with Equation 1.2.1, we see that $\mathbb{P}(0 \sim ne_1) \sim n^{1-d/\alpha}L(n)$ for some slowly varying function L. We should compare this situation to the one in the original Hammond-Sheffield model [14] (with extra details available in Igelbrink-Wakolbinger [17]). In the Hammond-Sheffield model, the coalescence probabilities $\mathbb{P}(0 \sim n)$ decay as a slowly varying function times $n^{2\alpha-1}$, and this gives rise to a fractional Brownian motion process with Hurst parameter $\alpha + 1/2$. The parameter range is $\alpha \in (0, 1/2)$. That is, if $\mathbb{P}(0 \sim n)$ is a slowly varying function times n^{-s} for some $s \in (0,1)$, we get a Hurst parameter of $1 - s/2$. In our case, we end up with $s = d/\alpha - 1$, suggesting a Hurst parameter of $H = \frac{1}{2}(3 - \frac{d}{\alpha})$. Recall that $d/2 < \alpha < d$, allowing the Hurst parameter to be any value strictly between 1/2 and 1. We confirm this by computing the variance. The covariance function can be inferred from the variance due to stationary increments (see Remark 1.4).

1.3.2 Process variance

Proposition 1.6. *Let $(v_n)_{n\geq 1}$ be a sequence of vertices. Suppose that $\alpha \in (d/2, d)$ and the color distribution satisfies $\mathbb{E}Y_{v_i} = 0$ and $\mathrm{Var}[Y_{v_i}] = \sigma^2$. If, for $i \neq j$, $\mathbb{P}(v_i \sim v_j) = |i-j|^{1-d/\alpha}L(|i-j|)$ for a positive slowly varying function L, then the variance of the sum $S_n = \sum_{i=1}^{n} Y_{v_i}$ is asymptotically*

$$\mathrm{Var}[S_n] \sim n^{3-d/\alpha}L(n)\frac{2\sigma^2}{(2-d/\alpha)(1-d/\alpha)}.$$

Proof. We can decompose

$$\mathrm{Var}[S_n] = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}[Y_{v_i}, Y_{v_j}]. \tag{1.3.3}$$

Since the colors Y_{v_i} and Y_{v_j} are independent if $v_i \not\sim v_j$ and are identical if $v_i \sim v_j$,

$$\begin{aligned}\mathbb{E}[Y_{v_i} Y_{v_j}] &= \mathbb{E}[Y_{v_i} Y_{v_j} | v_i \sim v_j]\mathbb{P}(v_i \sim v_j) + \mathbb{E}[Y_{v_i} Y_{v_j} | v_i \not\sim v_j]\mathbb{P}(v_i \not\sim v_j) \\ &= \sigma^2 \mathbb{P}(v_i \sim v_j).\end{aligned}$$

Then for $i \neq j$, $\mathrm{Cov}[Y_{v_i}, Y_{v_j}] = \sigma^2 \mathbb{P}(v_i \sim v_j) = \sigma^2 |i-j|^{1-d/\alpha} L(|i-j|)$. Decomposing the double sum in Equation 1.3.3 by $|i-j|$,

$$\mathrm{Var}[S_n] = \sigma^2 \left(n + 2 \sum_{i=1}^{n-1} (n-i) i^{1-d/\alpha} L(i) \right). \tag{1.3.4}$$

The sum in Equation 1.3.4 can be rewritten

$$\sum_{i=1}^{n-1} (n-i) i^{1-d/\alpha} L(i) = n^{3-d/\alpha} \sum_{i=1}^{n-1} \left(1 - \frac{i}{n}\right) \left(\frac{i}{n}\right)^{1-d/\alpha} \left(\frac{1}{n}\right) L(i),$$

which we recognize as a Riemann sum, except with a slowly varying function in the sum. Set $f_n(i) = (1 - i/n)(i/n)^{1-d/\alpha}(1/n)$. We claim that it is asymptotically equivalent to take the slowly varying function out:

$$\sum_{i=1}^{n-1} L(i) f_n(i) \sim L(n) \sum_{i=1}^{n-1} f_n(i). \tag{1.3.5}$$

Assuming this, the Riemann sum converges to an integral, and we can conclude

$$\mathrm{Var}[S_n] \sim 2\sigma^2 n^{3-d/\alpha} L(n) \int_0^1 (1-x) x^{1-d/\alpha}\, dx,$$

where the $O(n)$ term in Equation 1.3.4 disappeared because it is dominated by $n^{3-d/\alpha}$

(we assumed $\alpha > d/2$). Computing the integral gives the constant in the result.

Now, we show that the claim. First, for any constant $c \in (0,1)$, we show that

$$\sum_{i=\lceil cn \rceil}^{n-1} L(i) f_n(i) \sim L(n) \sum_{i=\lceil cn \rceil}^{n-1} f_n(i) \tag{1.3.6}$$

as $n \to \infty$. The uniform convergence theorem for slowly varying functions (Bingham, Goldie, and Teugels [1], Theorem 1.2.1) tells us that, as $x \to \infty$, $L(\lambda x)/L(x) \to 1$ uniformly for $\lambda \in [c,1]$. Concretely, with c fixed, for $\varepsilon > 0$, there is an N such that if $n > N$, then $|L(\lambda n)/L(n) - 1| < \varepsilon$ for all $\lambda \in [c,1]$. We apply this to $L(i) = L(\frac{i}{n} n)$. When $cn \le i \le n$, we have $\frac{i}{n} \in [c,1]$. Thus, when summing from $\lceil cn \rceil$ to n, we have $|L(i)/L(n) - 1| < \varepsilon$ for n sufficiently large. This justifies Equation 1.3.6.

We want to send $c \to 0$. We only care about the behavior of L on integer points, and it is always nonzero, so we can use Potter's theorem (Bingham, Goldie, and Teugels [1], Theorem 1.5.6) to see that for every $\varepsilon > 0$, there is a constant C with $L(i)/L(n) \le C(n/i)^{\varepsilon}$ whenever $1 \le i \le n$. Then, independently of c, we can bound

$$\sum_{i=1}^{\lceil cn \rceil} f_n L(i) \le L(n)\, C \sum_{i=1}^{\lceil cn \rceil} \left(1 - \frac{i}{n}\right) \left(\frac{i}{n}\right)^{1-d/\alpha} \left(\frac{1}{n}\right) \left(\frac{n}{i}\right)^{\varepsilon}. \tag{1.3.7}$$

The sum in the right hand side of Equation 1.3.7 is a Riemann sum for the integral $\int_0^c (1-x) x^{1-d/\alpha-\varepsilon}\, dx$. Taking ε small enough that $1 - d/\alpha - \varepsilon > -1$, the integral is finite. For n sufficiently large,

$$\sum_{i=1}^{\lceil cn \rceil} f_n(i) L(i) \le 2CL(n) \int_0^c (1-x) x^{1-d/\alpha-\varepsilon}\, dx,$$

where the factor of 2 is to allow for the error in approximating the Riemann sum by the integral. Set $g(c) = 2C \int_0^c (1-x) x^{1-d/\alpha-\varepsilon}\, dx$. Then, suppressing the ceiling function for notational clarity, we bound

$$
\begin{aligned}
\frac{\sum_{i=1}^{n} f_n(i)L(i)}{L(n)\sum_{i=1}^{n} f_n(i)} &= \frac{\sum_{i=1}^{cn} f_n(i)L(i) + \sum_{i=cn}^{n} f_n(i)L(i)}{L(n)\sum_{i=1}^{n} f_n(i)} \\
&\le \frac{L(n)g(c) + \sum_{i=cn}^{n} f_n(i)L(i)}{L(n)\sum_{i=1}^{n} f_n(i)} \\
&\sim \frac{L(n)g(c) + L(n)\int_c^1 (1-x)x^{1-d/\alpha}\,dx}{L(n)\int_0^1 (1-x)x^{1-d/\alpha}\,dx}.
\end{aligned}
$$

Thus, the limit of $\frac{\sum_{i=1}^{n} f_n(i)L(i)}{L(n)\sum_{i=1}^{n} f_n(i)}$ as $n \to \infty$ is at most

$$\frac{g(c) + \int_c^1 (1-x)x^{1-d/\alpha}\,dx}{\int_0^1 (1-x)x^{1-d/\alpha}\,dx}.$$

Since c was arbitrary, we can send it to 0. Noting that $g(c) \to 0$ as $c \to 0^+$, we get that the limit as $n \to \infty$ of $\frac{\sum_{i=1}^{n} f_n(i)L(i)}{L(n)\sum_{i=1}^{n} f_n(i)}$ is at most 1. We can repeat the same computation except using the trivial lower bound $\sum_{i=1}^{cn} f_n(i)L(i) \ge 0$ to get that the limit is at least

$$\frac{\int_c^1 (1-x)x^{1-d/\alpha}\,dx}{\int_0^1 (1-x)x^{1-d/\alpha}\,dx}.$$

Taking $c \to 0^+$ gives that the limit is at least 1. So $\frac{\sum_{i=1}^{n} f_n(i)L(i)}{L(n)\sum_{i=1}^{n} f_n(i)} \to 1$. We conclude that Equation 1.3.5 holds. This completes the proof. □

As a corollary of Proposition 1.5, Proposition 1.6, and Equation 1.2.1, we get the asymptotic variance of S_n.

Corollary 1.7. *Suppose that $\alpha \in (d/2, d)$ and that μ satisfies Assumption A. Let ℓ be as in Assumption A and $C > 0$ be as in Proposition 1.5. Then*

$$\mathrm{Var}[S_n] \sim Cn^{1-d/\alpha}\ell^{-d}(n).$$

At this point, we have shown that the proposed process S_n has the correct asymptotic variance. The key ingredient is that the probability that 0 is in the same component as ne_1 decays at an appropriate rate. This correlation decay is exactly what is needed to give the (asymptotic) covariance structure of fractional Brownian motion. We show in the next section that the process is asymptotically Gaussian. It is always interesting to ask whether there are modifications we can make to the model which give other useful covariance structures.

1.4 Asymptotic Gaussianity

For a process to converge to fractional Brownian motion, it must have the correct asymptotic covariance, and it must be asymptotically a Gaussian process. The previous section establishes the correct asymptotic covariance structure of the process. To show asymptotic Gaussianity, we use a martingale central limit theorem.

1.4.1 CLT overview

The goal is to show that $G^{(n)}(\cdot)$ is asymptotically a Gaussian process. Concretely, this means that for any $k \geq 1$, any choice of $0 \leq t_1 \leq \cdots \leq t_k$, and any real coefficients $a_1, \ldots, a_k$, the linear combination $\sum_{i=1}^{k} a_i G^{(n)}(t_i)$ converges in distribution to a Gaussian distribution. For notational simplicity, we only write down the proof for the random variable $S_n/\sqrt{\operatorname{Var} S_n}$ and indicate the changes needed for full joint Gaussianity. In places, it is also simpler for notational convenience to discuss the behavior of S_n directly, with the understanding that it must be rescaled appropriately.

The core tool is a martingale central limit theorem. There are many variations, but most require two core ingredients: asymptotic negligibility of individual increments and some control over the conditional variances or realized variance. The asymptotic negligibility of increments requirement is intuitive. No individual increment should contribute an asymptotically positive fraction of the variance; otherwise, this increment will have too great of influence to be averaged out. The conditional variance requirement can be thought of as asking that the process progresses at the correct speed. In continuous time martingales, the quadratic variation provides a natural measure of time in the martingale. Indeed, any continuous local martingale with unbounded quadratic variation is a time change of a Brownian motion, where the time change is such that the quadratic variation grows at the correct rate (Le Gall [21] Theorem 5.13). Control over the conditional variances is analogous to controlling the quadratic variation of a martingale. If the conditional variance is not sufficiently well-behaved, then the limiting process may fail to have a consistent or well-defined growth rate and thereby fail to be a Gaussian process.

Definition 1.8. *We call a collection of random variables $X_{n,m}$ and σ-algebras $\mathcal{F}_{n,m}$ (for $n \geq 1$, $0 \leq m \leq k_n$) a martingale difference array if every $X_{n,m}$ is $\mathcal{F}_{n,m}$-measurable and if $\mathbb{E}[X_{n,m}|\mathcal{F}_{n,m-1}] = 0$ for $1 \leq m \leq k_n$.*

This is called a martingale difference array because $X_{n,m}$ is the difference of the martingale $S_{n,m} = \sum_{i=0}^{m} X_{n,i}$ for $0 \leq m \leq k_n$. At times, we deviate in immaterial ways from this definition, for example, letting the range of m be $0 \leq m \leq k_n + n$. We use the following formulation of a martingale CLT, which can be found in Hall and Heyde [13], Corollary 3.1:

Theorem 1.9. *Let ($X_{n,m}$ and $\mathcal{F}_{n,m}, n \geq 1, 0 \leq m \leq k_n$) be a mean-zero, square-integrable martingale difference array. Suppose that*

- *for all $\varepsilon > 0, \sum_{m=1}^{k_n} \mathbb{E}[X_{n,m}^2 \mathbf{1}_{\{|X_{n,m}|>\varepsilon\}}|\mathcal{F}_{n,m-1}]$ converges to 0 in probability,*
- *$V_{n,k_n}^2 = \sum_{m=1}^{k_n} \mathbb{E}[X_{n,m}^2|\mathcal{F}_{n,m-1}]$ converges to 1 in probability.*

Then $S_{n,k_n} = \sum_{m=0}^{k_n} X_{n,m}$ converges in distribution to a standard Gaussian.

The first condition is a Lindeberg-style condition, which requires that the large increments (i.e., larger than ε) collectively contribute little to the conditional variance. The second condition asks that the conditional variance be concentrated around 1.

1.4.2 Variance decomposition

We use the filtration given by $\sigma_m = \sigma(Y_v : v \in \mathbb{Z}^D, v \cdot e_1 < m)$, that is, σ_m is the σ-algebra generated by the colors of all vertices whose first coordinate is strictly less than m. This can be thought of as the history prior to layer m. For $v \in \mathbb{Z}^D$ and $n \geq 1$, set $Y_v^m = \mathbb{E}[Y_v|\sigma_m]$ and $S_n^m = \mathbb{E}[S_n|\sigma_m]$.

This lets us define a martingale which converges to $G^{(n)}(1) = S_n/\sqrt{\operatorname{Var} S_n}$. For each n we will associate a k_n, to be determined later. For $-k_n \leq m \leq n$, set

$$G_n^m = S_n^m/\sqrt{\operatorname{Var} S_n}.$$

S_n is not a martingale, as it has correlated increments. Instead, our sequence of martingales will be a sequence of Doob martingales, the n-th of which converges to S_n. S_n^m is a martingale which converges to S_n as $m \to \infty$ (actually, it is equal to S_n as soon as $m > n$) and is the martingale that the martingale difference array defines. This gives a martingale difference array with $\sqrt{\operatorname{Var} S_n} X_{n,0} = S_n^{-k_n}$ and

$$\sqrt{\operatorname{Var} S_n} X_{n,i} = S_n^{-k_n+i} - S_n^{-k_n+i-1} \tag{1.4.1}$$

for $1 \leq i \leq n + k_n$. The filtration is

$$\mathcal{F}_{n,i} = \sigma_{-k_n+i} \tag{1.4.2}$$

for $0 \leq i \leq n + k_n$.

Up to rescaling by $\sqrt{\operatorname{Var} S_n}$, the increments of the martingale difference array are thus $S_n^{m+1} - S_n^m$. We see that

$$S_n^{m+1} - S_n^m = \sum_{i=1}^{n} \left(Y_{ie_1}^{m+1} - Y_{ie_1}^m\right). \tag{1.4.3}$$

Recall that for a vertex v, the notation $v \to x$ means that x is an ancestor of v (not necessarily the direct parent), and for all vertices, $v \to v$ is true. Set

$$C_x^n = \sum_{i=1}^{n} \mathbb{P}(ie_1 \to x) \tag{1.4.4}$$

and set

$$A_m = \{x \in \mathbb{Z}^D : x \cdot e_1 = m\}, \tag{1.4.5}$$

the set of integer vectors with first component equal to m.

Proposition 1.10. *Almost surely,*

$$S_n^{m+1} - S_n^m = \sum_{x \in A_m} C_x^n \left(Y_x - Y_x^m\right).$$

And for any $v \in \mathbb{Z}^D$ and $m \in \mathbb{Z}$,

$$Y_v^{m+1} - Y_v^m = \sum_{x \in A_m} (Y_x - Y_x^m)\mathbb{P}(v \to x).$$

Proof. We expand on Equation 1.4.3. First, assume $m \leq v \cdot e_1$. Then we can decompose Y_v as

$$Y_v = \sum_{x \in A_m} Y_x \mathbf{1}_{\{v \to x\}},$$

since v has exactly one ancestor in A_m. For $x \in A_m$, Y_x is σ_{m+1} measurable, so $Y_x^{m+1} = Y_x$. And for $x \in A_m$, the indicator $\mathbf{1}_{\{v \to x\}}$ is independent of both σ_m and σ_{m+1}. Taking the conditional expectations with respect to σ_m and σ_{m+1} respectively, these facts tells us that

$$Y_v^m = \sum_{x \in A_m} Y_x^m \mathbb{P}(v \to x) \text{ and } Y_v^{m+1} = \sum_{x \in A_m} Y_x \mathbb{P}(v \to x). \tag{1.4.6}$$

Combining, we get

$$Y_v^{m+1} - Y_v^m = \sum_{x \in A_m} (Y_x - Y_x^m)\mathbb{P}(v \to x).$$

Since $\sum_{x \in A_m} \mathbb{P}(v \to x) = 1$, this is a weighted average of $Y_x - Y_x^m$, where the weighting is given by the distribution of the random walk started at v when it hits A_m.

Now we treat $m > v \cdot e_1$. In this case, Y_v is measurable with respect to both σ_m and σ_{m+1}. So $Y_v^{m+1} - Y_v^m = Y_v - Y_v = 0$. And in this case, for $x \in A_m$, we have $\mathbb{P}(v \to x) = 0$.

Thus, regardless of the value of m, it is true that

$$Y_v^{m+1} - Y_v^m = \sum_{x \in A_m} (Y_x - Y_x^m)\mathbb{P}(v \to x). \tag{1.4.7}$$

This proves the second claim. Specializing to $v = ie_1$, Equation 1.4.7 lets us express

$$S_n^{m+1} - S_n^m = \sum_{i=1}^{n} \sum_{x \in A_m} (Y_x - Y_x^m)\mathbb{P}(ie_1 \to x) \tag{1.4.8}$$

Since Y_x is assumed to be bounded (a finite fourth moment also suffices), it has a finite first absolute moment. For fixed i, we compute

$$\mathbb{E} \sum_{x \in A_m} \mathbb{P}(ie_1 \to x)|Y_x - Y_x^m| \leq 2\,\mathbb{E}|Y_0| \sum_{x \in A_m} \mathbb{P}(ie_1 \to x) < \infty,$$

so $\sum_{x \in A_m} \mathbb{P}(ie_1 \to x)|Y_x - Y_x^m|$ is finite almost surely. This means that almost surely, we can swap the sums in Equation 1.4.8, which proves the claim. $\square$

By our definition of the martingale difference array, V_{n,k_n}^2 is $\mathrm{Var}[S_n]$ times

$$\sum_m \mathbb{E}\left[\left(S_n^{m+1} - S_n^m\right)^2 \,|\, \sigma_m\right] = \sum_m \mathrm{Var}\left[S_n^{m+1} - S_n^m \,|\, \sigma_m\right] = \sum_m \mathrm{Var}[S_n^{m+1} \,|\, \sigma_m],$$

where the first equality holds because $\mathbb{E}[S_n^{m+1} - S_n^m|\sigma_m] \equiv 0$ and the second equality holds because S_n^m is σ_m-measurable.

Lemma 1.11. *The conditional variance of an increment can be decomposed as*

$$\mathrm{Var}[S_n^{m+1} - S_n^m|\sigma_m] = \sum_{x\in A_m} (C_x^n)^2 \,\mathrm{Var}[Y_x|\sigma_m].$$

Proof. We have established in Proposition 1.10 that

$$S_n^{m+1} - S_n^m = \sum_{x\in A_m} C_x^n \,(Y_x - Y_x^m).$$

Since Y_x^m is σ_m-measurable, the variance of $S_n^{m+1} - S_n^m$ is the same as the variance of $\sum_{x\in A_m} C_x^n Y_x$. Under this perspective, the lemma is saying that the random variables in the sum are conditionally uncorrelated. Indeed, the key observation is that $(Y_x)_{x\in A_m}$ is a collection of conditionally independent random variables when conditioned on σ_m. For distinct $x, x' \in A_m$, the values of Y_x and $Y_{x'}$ are determined by choosing a random parent for x and x', with Y_x and $Y_{x'}$ being assigned the values of the respective parents. Crucially, conditioned on σ_m, the choice of parent of x is independent from the choice of parent of x'.

We confirm this intuitive idea with a rigorous computation. Recalling that for $x \in A_m$, we have

$$Y_x = \sum_{z\in A_{m-1}} \mathbf{1}_{\{x\to z\}} Y_z$$

and

$$Y_x^m = \mathbb{E}[Y_x \mid \sigma_m] = \sum_{z\in A_{m-1}} \mathbb{P}(x \to z) Y_z,$$

the definition of conditional covariance gives that for $x, x' \in A_m$, $\mathrm{Cov}[Y_x, Y_{x'} \mid \sigma_m]$ is equal to

$$\sum_{z,z' \in A_{m-1}} \mathbb{E}\left[Y_z Y_{z'} \left(\mathbf{1}_{\{x \to z\}} - \mathbb{P}(x \to z)\right) \left(\mathbf{1}_{\{x' \to z'\}} - \mathbb{P}(x' \to z')\right) \mid \sigma_m\right].$$

Because Y_z and $Y_{z'}$ are σ_m-measurable, and since the joint distribution of the parents of x and x' is independent of σ_m, this is equal to

$$\sum_{z,z' \in A_{m-1}} Y_z Y_{z'}\, \mathbb{E}\left[\left(\mathbf{1}_{\{x \to z\}} - \mathbb{P}(x \to z)\right) \left(\mathbf{1}_{\{x' \to z'\}} - \mathbb{P}(x' \to z')\right)\right].$$

For $x \neq x'$, the terms in the expectation are independent with mean 0, so the conditional covariance is 0. For $x = x'$, we simply have $\mathrm{Cov}[Y_x, Y_x \mid \sigma_m] = \mathrm{Var}[Y_x \mid \sigma_m]$, which proves the lemma. □

This gives us a partial grasp on the conditional variance of the increments $S_n^{m+1} - S_n^m$. We refine this in the following lemma, which is a conditional version of the formula $\mathrm{Var}[X] = \frac{1}{2}\mathbb{E}(X - \tilde{X})^2$ where $\tilde{X}$ is an i.i.d. copy of X.

Lemma 1.12. *For $x \in A_m$, we can decompose*

$$\mathrm{Var}[Y_x \mid \sigma_m] = \frac{1}{2} \sum_{z,z' \in A_{m-1}} (Y_z - Y_{z'})^2 \mathbb{P}(x \to z)\mathbb{P}(x \to z').$$

Proof. We know $Y_x = \sum_{z \in A_{m-1}} Y_z \mathbf{1}_{\{x \to z\}}$ and

$$Y_x^m = \sum_{z \in A_{m-1}} Y_z \mathbb{P}(x \to z).$$

By definition,

$$\mathrm{Var}[Y_x | \sigma_m] = \mathbb{E}[(Y_x - Y_x^m)^2 | \sigma_m].$$

We present two proofs. For the first, start with the desired expression, add and subtract Y_x^m, and expand the square. Then the $(Y_z - Y_x^m)^2$ terms give the conditional variance, and the $(Y_z - Y_x^m)(Y_{z'} - Y_x^m)$ terms sum to 0:

$$\begin{aligned}\sum_{z,z' \in A_{m-1}} &(Y_z - Y_x^m + Y_x^m - Y_{z'})^2\, \mathbb{P}(x \to z)\mathbb{P}(x \to z') \\ &= \sum_{z,z'} (Y_z - Y_x^m)^2\, \mathbb{P}(x \to z)\mathbb{P}(x \to z') \\ &\quad + \sum_{z,z'} (Y_x^m - Y_{z'})^2\, \mathbb{P}(x \to z)\mathbb{P}(x \to z') \\ &\quad + 2\sum_{z,z'} (Y_z - Y_x^m)(Y_x^m - Y_{z'})\, \mathbb{P}(x \to z)\mathbb{P}(x \to z').\end{aligned}$$

Since $\sum_{z' \in A_{m-1}} \mathbb{P}(x \to z') = 1$, we get that

$$\sum_{z,z'} (Y_z - Y_x^m)^2\, \mathbb{P}(x \to z)\mathbb{P}(x \to z') = \sum_z (Y_z - Y_x^m)^2\, \mathbb{P}(x \to z),$$

and similarly for the sum involving z'. This is just the conditional variance:

$$\begin{aligned}\mathrm{Var}[Y_x \mid \sigma_m] &= \mathbb{E}[(Y_x - Y_x^m)^2 \mid \sigma_m] \\ &= \mathbb{E}\Big[\sum_z (Y_x - Y_x^m)^2 \mathbf{1}_{\{x \to z\}} \mid \sigma_m\Big] \\ &= \mathbb{E}\Big[\sum_z (Y_z - Y_x^m)^2 \mathbf{1}_{\{x \to z\}} \mid \sigma_m\Big], \text{ since } Y_z = Y_x \text{ on the event } \{x \to z\} \\ &= \sum_z (Y_z - Y_x^m)^2\, \mathbb{P}(x \to z),\end{aligned}$$

where the last equality is true because Y_z and Y_x^m are σ_m-measurable and $\mathbf{1}_{\{x \to z\}}$ is independent of σ_m. For the cross-term from the product, we see that

$$\begin{aligned}\sum_{z,z'} &(Y_z - Y_x^m)(Y_x^m - Y_{z'})\, \mathbb{P}(x \to z)\mathbb{P}(x \to z') \\ &= \Big(\sum_z (Y_z - Y_x^m)\mathbb{P}(x \to z)\Big)\Big(\sum_{z'} (Y_x^m - Y_{z'})\mathbb{P}(x \to z')\Big) = 0,\end{aligned}$$

since $Y_x^m = \sum_{z \in A_{m-1}} Y_z\, \mathbb{P}(x \to z)$. Collecting the previous two computations proves the lemma.

Here's an alternate proof, which is in some ways more direct. As with non-conditional variance,

$$\mathrm{Var}[Y_x \mid \sigma_m] = \mathbb{E}[(Y_x - Y_x^m)^2 \mid \sigma_m] = \mathbb{E}[Y_x^2 \mid \sigma_m] - (Y_x^m)^2.$$

We already have an expression for Y_x^m, and we can similarly compute

$$\begin{aligned}
\mathbb{E}[Y_x^2 \mid \sigma_m] &= \mathbb{E}\Big[\Big(\sum_{z \in A_{m-1}} Y_z \mathbf{1}_{\{x \to z\}}\Big)^2 \mid \sigma_m\Big] \\
&= \sum_{z, z' \in A_{m-1}} \mathbb{E}\left[Y_z Y_{z'} \mathbf{1}_{\{x \to z\}} \mathbf{1}_{\{x \to z'\}} \mid \sigma_m\right] \\
&= \sum_{z \in A_{m-1}} \mathbb{E}\left[Y_z^2 \mathbf{1}_{\{x \to z\}} \mid \sigma_m\right] \\
&= \sum_{z \in A_{m-1}} Y_z^2 \mathbb{P}(x \to z),
\end{aligned}$$

where the second-to-last equality holds because $\{x \to z\}$ and $\{x \to z'\}$ are disjoint unless $z = z'$ and the last equality holds because Y_z is σ_m-measurable and $\mathbf{1}_{\{x \to z\}}$ is independent of σ_m. Then the lemma follows from computing

$$\begin{aligned}
&\sum_{z, z' \in A_{m-1}} (Y_z - Y_{z'})^2 \mathbb{P}(x \to z) \mathbb{P}(x \to z') \\
&= \sum_{z, z'} Y_z^2 \mathbb{P}(x \to z) \mathbb{P}(x \to z') + Y_{z'}^2 \mathbb{P}(x \to z) \mathbb{P}(x \to z') - 2 Y_z Y_{z'} \mathbb{P}(x \to z) \mathbb{P}(x \to z') \\
&= \sum_z Y_z^2 \mathbb{P}(x \to z) + \sum_{z'} Y_{z'}^2 \mathbb{P}(x \to z') - 2\Big(\sum_z Y_z \mathbb{P}(x \to z)\Big)\Big(\sum_{z'} Y_{z'} \mathbb{P}(x \to z')\Big) \\
&= 2 \sum_z Y_z^2 \mathbb{P}(x \to z) - 2\Big(\sum_z Y_z \mathbb{P}(x \to z)\Big)^2 \\
&= 2\left(\mathbb{E}[Y_x^2 \mid \sigma_m] - (Y_x^m)^2\right).
\end{aligned}$$

□

Corollary 1.13. *We have that*

$$\mathrm{Var}[S_n]V_{n,k_n}^2 = \sum_{m=-k_n}^{n} \sum_{x\in A_m} (C_x^n)^2 \sum_{z,z'\in A_{m-1}} \frac{1}{2}(Y_z - Y_{z'})^2 \mathbb{P}(x\to z)\mathbb{P}(x\to z').$$

Remark 1.14. Since

$$\mathbb{E}V_{n,k_n}^2 = \mathrm{Var}\left[\frac{S_n - S_n^{-k_n}}{\sqrt{\mathrm{Var}\, S_n}}\right] \text{ and } \mathbb{E}\frac{1}{2}(Y_z - Y_z')^2 \le \mathrm{Var}[Y_z],$$

we also get that

$$\mathrm{Var}[S_n - S_n^{-k_n}] \le \mathrm{Var}[Y_0] \sum_{m=-k_n}^{n} \sum_{x\in A_m} (C_x^n)^2 \sum_{z,z'\in A_{m-1}} \mathbb{P}(x\to z)\mathbb{P}(x\to z'). \qquad (1.4.9)$$

To make use of Corollary 1.13, we need control of the coefficients. The following lemmas use the local limit theorem to get quantitative bounds on various coefficients that appear in the conditional variance decomposition.

Lemma 1.15. *Suppose μ satisfies Assumption A. Let $(Z_n)_{n\ge1}$ be a random walk on $\mathbb{Z}^d$ with i.i.d. μ-distributed increments. Let ℓ be as in Assumption A. Then there is a finite constant C such that for n sufficiently large and for $z\in\mathbb{Z}^d$,*

$$\mathbb{P}(Z_n = z) \le Cn^{-d/\alpha}\ell^{-d}(n).$$

Proof. The local limit theorem (Theorem 1.3) tells us that as $n\to\infty$,

$$\sup_{z\in\mathbb{Z}^d}\left| n^{d/\alpha}\,\ell^d(n)\mathbb{P}(Z_n = z) - g\left(\frac{z}{n^{1/\alpha}\ell(n)}\right)\right| \to 0.$$

We show in Proposition A.1 in the appendix that if a stable law is not supported on a hyperplane, it has a bounded density. Let C be a strict upper bound for the density. Since the limit of the supremum is 0, eventually it is less than C. So for n sufficiently large, for all $z \in \mathbb{Z}^d$,

$$n^{d/\alpha}\,\ell^d(n)\mathbb{P}(Z_n = z) - g\left(\frac{z}{n^{1/\alpha}\ell(n)}\right) \leq C,$$

which tells us that $n^{d/\alpha}\ell^d(n)\mathbb{P}(Z_n = z) \leq 2C$. Rearranging gives the result. $\square$

Lemma 1.16. *Suppose $\alpha < d$. Then there is a constant C such that for all $n \geq 1$ and for all $m \in \mathbb{Z}$,*

$$\sum_{x \in A_m} (C_x^n)^2 \leq Cn.$$

Proof. We bound

$$\sum_{x \in A_m} (C_x^n)^2 \leq \left(\sum_{x \in A_m} C_x^n\right)\left(\max_{x \in A_m} C_x^n\right). \tag{1.4.10}$$

When $m \leq i$, the random walk started at ie_1 intersects the hyperplane A_m at exactly one vertex and so $\mathbb{P}(ie_1 \to \cdot)$ defines a probability distribution on A_m, showing that $\sum_{x \in A_m} \mathbb{P}(ie_1 \to x) = 1$. For $m > i$, $\mathbb{P}(ie_1 \to \cdot)$ is identically zero on A_m. As everything is nonnegative, we can swap the order of summation, giving

$$\sum_{x \in A_m} C_x^n = \sum_{i=1}^{n} \sum_{x \in A_m} \mathbb{P}(ie_1 \to x) \leq n. \tag{1.4.11}$$

We can bound

$$\max_{x \in A_m} C_x^n = \max_{x \in A_m} \sum_{i=1}^{n} \mathbb{P}(ie_1 \to x) \leq \sum_{i=1}^{n} \max_{x \in A_m} \mathbb{P}(ie_1 \to x)$$

and we show that $\mathbb{P}(ie_1 \to x)$ is small for all x.

If $i < m$, then $\mathbb{P}(ie_1 \to x) = 0$ for $x \in A_m$. For $i \geq m$, with the interpretation that $\mathbb{P}(ie_1 \to \cdot)$ on A_m is the probability distribution of step $i - m$ of a random walk with independent μ-distributed increments, Lemma 1.15 tells us that there are positive constants C' and m_0' such that if $i - m > m_0'$, for $x \in A_m$,

$$\mathbb{P}(ie_1 \to x) \le C'(i-m)^{-d/\alpha}\ell^{-d}(i-m).$$

Fix $\varepsilon > 0$ small enough that $-d/\alpha + \varepsilon < -1$ (we assumed $\alpha < d$). Slowly varying functions grow more slowly than any power function with positive exponent. This means there is a positive integer $m_0 \ge m_0'$ depending on ε such that for $i - m \ge m_0$, we get that $C'\,\ell^{-d}(i-m) \le (i-m)^{\varepsilon}$. When $0 \le i - m < m_0$, we use the trivial bound that probabilities are at most 1. And when $i - m < 0$, the probability is 0. This lets us bound

$$\begin{aligned}
\sum_{i=1}^{n} \max_{x \in A_m} \mathbb{P}(ie_1 \to x) &\le \sum_{i=-\infty}^{\infty} \max_{x \in A_m} \mathbb{P}(ie_1 \to x) \\
&= \sum_{i=m}^{\infty} \max_{x \in A_m} \mathbb{P}(ie_1 \to x) \\
&\le \sum_{i=m}^{m+m_0-1} 1 + \sum_{i=m+m_0}^{\infty} (i-m)^{-d/\alpha+\varepsilon} \\
&= m_0 + \sum_{i=m_0}^{\infty} i^{-d/\alpha+\varepsilon}.
\end{aligned}$$

This last expression is finite, since $-d/\alpha + \varepsilon < -1$, and is independent of m and of n. So there is some absolute constant C, depending only on μ, such that for all $x \in \mathbb{Z}^d$ and $n \ge 1$,

$$C_x^n \le C. \tag{1.4.12}$$

Combining this with Equation 1.4.10 gives the result. $\square$

We also need a version that is better for large $|m|$, although it has worse dependence on n.

Lemma 1.17. *Let $C_x^n = \sum_{i=1}^{n} \mathbb{P}(ie_1 \to x)$. For any $\varepsilon > 0$, there is an $m_0 < 0$ such* *that for $m < m_0$ and $n \geq 1$,*

$$\sum_{x \in A_m} (C_x^n)^2 \leq n^{2+\varepsilon} \, |m|^{-d/\alpha+\varepsilon}.$$

Proof. The proof is identical to that of Lemma 1.16 until we reach the statement that there is a constant C such that for $i - m$ sufficiently large, for $x \in A_m$,

$$\mathbb{P}(ie_1 \to x) \leq C(i-m)^{-d/\alpha} \ell^{-d}(i-m).$$

Since we want to show the statement for $m < m_0$ with $m_0 < 0$, we are free to assume $m < 0$ and $i > 0$. With these assumptions, the bound $\mathbb{P}(ie_1 \to x) \leq C|m|^{-d/\alpha} \ell^{-d}(i-m)$ holds for $x \in A_m$. From here, fix $\varepsilon \in (0,1)$. Since slowly varying functions grow more slowly than any polynomial, for $i - m$ sufficiently large, $C\,\ell^{-d}(i-m) \leq \frac{1}{4}(i-m)^{\varepsilon} \leq \frac{1}{2}(i^{\varepsilon} + |m|^{\varepsilon})$. Since $i - m \geq |m|$, there is some $m_0 < 0$ for which the bound $\mathbb{P}(ie_1 \to x) \leq \frac{1}{2}|m|^{-d/\alpha}(|m|^{\varepsilon} + |i|^{\varepsilon})$ holds for all $i \geq 1$ and all $x \in A_m$ with $m < m_0$. So for $m < m_0$,

$$\begin{aligned}
\max_{x \in A_m} C_x^n &\leq \frac{1}{2} \sum_{i=1}^{n} |m|^{-d/\alpha}(|m|^{\varepsilon} + |i|^{\varepsilon}) \\
&\leq \frac{1}{2} n \, |m|^{-d/\alpha+\varepsilon} + \frac{1}{2} n^{1+\varepsilon} |m|^{-d/\alpha} \\
&\leq n^{1+\varepsilon} |m|^{-d/\alpha+\varepsilon},
\end{aligned}$$

giving a final bound of

$$\sum_{x \in A_m} (C_x^n)^2 \leq n^{2+\varepsilon} |m|^{-d/\alpha+\varepsilon}.$$

□

1.4.3 Asymptotic negligibility

The asymptotic negligibility for the martingale central limit theorem which we use is a Lindeberg-style condition: we want to show for all $\varepsilon > 0$,

$$\sum_{m=-k_n}^{n} \mathbb{E}[X_{n,m}^2 \mathbf{1}_{\{|X_{n,m}|>\varepsilon\}} | \sigma_m]$$

converges to 0 in probability. The basic idea is that, conditional on σ_m, the increment $X_{n,m}$ becomes a sum of independent random variables. Hoeffding's inequality then gives an exponentially small bound on $\mathbb{P}(|X_{n,m}| > \varepsilon)$. Since $X_{n,m}^2$ is trivially bounded by n^2, it suffices to show that there is a choice of k_n which grows at most polynomially in n.

Lemma 1.18. *Suppose $\alpha < d$. There is a constant C such that for n sufficiently large, for all $m \in \mathbb{Z}$, and for all $\varepsilon > 0$, the following bound holds:*

$$\mathbb{P}(|S_n^{m+1} - S_n^m|^2 \geq \varepsilon \sqrt{\operatorname{Var} S_n} \mid \sigma_{m-1}) \leq 2 \exp\left(-C\varepsilon^2 n^{2-d/\alpha} \ell^{-d}(n)\right).$$

Proof. The main insight is that $S_n^{m+1} - S_n^m$ is the sum of conditionally independent random variables. We use a conditional Hoeffding's inequality. The Hoeffding inequality says that if $(X_i)_{i\geq 1}$ are independent random variables which are bounded by $a_i \leq X_i \leq b_i$ almost surely, and if $W_n = X_1 + \cdots + X_n$, we have the bound

$$\mathbb{P}(|W_n - \mathbb{E}W_n| \geq t) \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{1.4.13}$$

Recall from Equation 1.4.8 that

$$S_n^{m+1} - S_n^m = \sum_{x \in A_m} C_x^n (Y_x - Y_x^m).$$

The variables Y_x^m are σ_m-measurable by their definitions, and the random variables

$(Y_x)_{x\in A_m}$ are conditionally independent given σ_m. Specifically, conditional on σ_m, each Y_x is a random variable which takes value Y_z with probability $\mathbb{P}(x \to z)$. As $(\mathbf{1}_{\{x\to z\}})_{z\in A_{m-1}}$ is independent of $(\mathbf{1}_{\{x'\to z\}})_{z\in A_{m-1}}$ for $x \neq x'$, and since these are two collections are independent when conditioned on σ_m, we get the conditional independence of $(Y_x)_{x\in A_m}$.

Here is where we assume that the color distribution Y_x is bounded. Suppose $a \le Y_x \le b$ almost surely and set $c = b - a$. Conditional Hoeffding tells us

$$\mathbb{P}(|S_n^{m+1} - S_n^m| \ge \varepsilon\sqrt{\operatorname{Var} S_n} \mid \sigma_m) \le 2\exp\left(-\frac{2\varepsilon^2 \operatorname{Var}[S_n]}{c^2 \sum_{x\in A_m}(C_x^n)^2}\right),$$

and the right side is no longer random. By Proposition 1.6, $\operatorname{Var}[S_n] \sim C' n^{3-d/\alpha}\ell^{-d}(n)$ for some constant C'. So for n sufficiently large, $\operatorname{Var}[S_n] \ge \frac{1}{2}C' n^{3-d/\alpha}\ell^{-d}(n)$. We know from Lemma 1.16 that there is some constant C'' for which $\sum_{x\in A_m}(C_x^n)^2 \le C'' n$ for all $n \ge 1$. So for n sufficiently large,

$$\mathbb{P}(|S_n^{m+1} - S_n^m| \ge \varepsilon\sqrt{\operatorname{Var} S_n} \mid \sigma_m) \le 2\exp\left(-\frac{C'\varepsilon^2 n^{3-d/\alpha}\ell^{-d}(n)}{2c^2 C'' n}\right).$$

□

The lemma requires $\alpha < d$, and as long as $\alpha > d/2$, the exponent of n is positive. This shows that if $\alpha \in (d/2, d)$, the probability that the martingale increments are too big decays like a stretched exponential in n. We sum $O(k_n)$ many of these terms, so it remains to show that k_n can be chosen not too large while still satisfying the other conditions of the martingale central limit theorem. Polynomial dependence on n, for example, suffices.

Lemma 1.19. *For any $\delta > 0$, the sequence $k_n = \lceil n^{1+\delta} \rceil$ satisfies*

$$\operatorname{Var}[S_n^{-k_n}]/\operatorname{Var}[S_n] \to 0$$

as $n \to \infty$.

Proof. As $(S_n^m)_{m\in\mathbb{Z}}$ is a martingale with respect to the filtration $(\sigma_m)_{m\in\mathbb{Z}}$, it has a backwards limit $S_n^{-\infty}$ with convergence almost surely and in L^2. The limit is measurable with respect to the tail σ-algebra $\sigma_{-\infty} = \cap_m \sigma_m$. Moreover, $(S_n^{-\infty}) = \mathbb{E}[S_n \mid \sigma_\infty]$. For details, see Durrett [10] Chapter 4.6, in particular Theorems 4.6.1 and 4.6.2, or Doob [8] sections XI.15 and XI.16. The convergence in L^2 only requires that S_n has a finite second moment.

This lets us break down the variance of $S_n^{-k_n}$ as

$$\mathrm{Var}[S_n^{-k_n}] = \mathrm{Var}[S_n^{-\infty}] + \sum_{m=-\infty}^{-k_n-1} \mathrm{Var}[S_n^{m+1} - S_n^m]. \tag{1.4.14}$$

The law of total variance splits $\mathrm{Var}[S_n^{m+1} - S_n^m]$ into the expected conditional variance plus the variance of the conditional expectation. Since

$$\mathbb{E}[S_n^{m+1} - S_n^m \mid \sigma_m] = 0 \text{ almost surely,}$$

only the expected conditional variance remains. Using the law of total variance and Lemma 1.11, we get

$$\mathrm{Var}[S_n^{m+1} - S_n^m] = \mathbb{E}\,\mathrm{Var}[S_n^{m+1} - S_n^m \mid \sigma_m] = \sum_{x\in A_m} (C_x^n)^2\, \mathbb{E}\,\mathrm{Var}[Y_x \mid \sigma_m]. \tag{1.4.15}$$

Since the distribution of $(Y_x)_{x\in\mathbb{Z}^D}$ is translation-invariant, $\mathbb{E}\,\mathrm{Var}[Y_x|\sigma_m]$ is just some absolute constant C, independent of x or m. Combining Equation 1.4.14 and Equation 1.4.15 with the bound given by Lemma 1.17, for $\varepsilon \in (0, d/\alpha - 1)$, we get that

$$\begin{aligned}\mathrm{Var}[S_n^{-k_n}] &\le \mathrm{Var}[S_n^{-\infty}] + C \sum_{m=-\infty}^{-k_n-1} n^{2+\varepsilon}|m|^{-d/\alpha+\varepsilon}\\ &\le \mathrm{Var}[S_n^{-\infty}] + C' n^{2+\varepsilon}(k_n)^{1-d/\alpha+\varepsilon}\end{aligned}$$

for some constant C'. We will show shortly that $\mathrm{Var}[S_n^{-\infty}] = 0$. Assuming that fact, since $1 - d/\alpha + \varepsilon < 0$, and using the bound $\mathrm{Var}[S_n] \geq n^{3-d/\alpha-\varepsilon}$ for large n, some algebra reveals that choosing $k_n \geq n^a$ for $a > \frac{d/\alpha-1+2\varepsilon}{d/\alpha-1-\varepsilon}$ gives $\mathrm{Var}[S_n^{-k_n}]/\mathrm{Var}[S_n] \to 0$ as $n \to \infty$. But ε was arbitrary and $\frac{d/\alpha-1+2\varepsilon}{d/\alpha-1-\varepsilon}$ approaches 1 as $\varepsilon \to 0$. So $k_n \geq n^a$ for any $a > 1$ is sufficient to guarantee that $\mathrm{Var}[S_n^{-k_n}]/\mathrm{Var}[S_n] \to 0$ as $n \to \infty$.

To finish the proof, we must show that the tail sigma-algebra is trivial. Here's the idea. Fix two vertices, x and y. In the model, Y_x^m and Y_y^m are obtained by taking random walks which proceed until they hit a point z in A_{m-1}, at which point Y_x assumes the value of Y_z. And Y_x^m is the expectation of Y_x over the randomness of the choice of $z \in A_{m-1}$. The local limit theorem lets us work directly with the distribution of the random walks. Y_x^m and Y_y^m use random walks with the same increment distributions, but started at x and y respectively. If we can show that for $|m|$ large, the two distributions are approximately equal, then the expectations are close.

We can make this precise to show that $Y_x^{-\infty} = Y_y^{-\infty}$ almost surely. If this holds, $S_n^{-\infty}$ is a sum of n random variables which are almost surely identical: $\mathrm{Var}[S_n^{-\infty}] = \mathrm{Var}[nY_0^{-\infty}] = n^2 \mathrm{Var}[Y_0^{-\infty}]$. But $\mathrm{Var}[S_n^{-\infty}] \leq \mathrm{Var}[S_n]$, and $\mathrm{Var}[S_n] = o(n^2)$ by Proposition 1.6, so the only possibility is that $\mathrm{Var}[Y_0^{-\infty}] = 0$, which means that $Y_0^{-\infty}$ is almost surely constant. This means that the tail sigma-algebra is trivial.

Let's show that $Y_x^{-\infty} = Y_y^{-\infty}$ almost surely. Since things are translation invariant, it suffices to consider the case where $x = 0$ and $y = e_j$, with the rest of $\mathbb{Z}^D$ following inductively.

The goal is to show that the two distributions induced on A_m by the random walks started at x and y are very similar for large $|m|$. To this end, we split the density of the stable law U into a region of high density and low density. The two distributions have small relative error on the high density region and small total mass on the low density region. We make use of the continuity of g, the density of the stable law. Fix

$k > 1$ and set

$$B_{1/k} = \{x \in \mathbb{R}^d : g(x) \geq 1/k\}$$

(this is the "high density" region). Analogously, letting $\bar{z}$ denote the projection of z onto its last d coordinates, set

$$B_{1/k}^{(m)} = \left\{ z \in A_m : \frac{\bar{z}}{|m|^{1/\alpha}\ell^d(|m|)} \in B_{1/k} \right\}.$$

By the local limit theorem (Theorem 1.3), there is some m_0 such that for $m < m_0$, for all $z \in A_m$,

$$g\left(\frac{\bar{z}}{|m|^{1/\alpha}\ell(|m|)}\right) - \frac{1}{k^2} \leq |m|^{d/\alpha}\ell^d(|m|)\mathbb{P}(0 \to z) \leq g\left(\frac{\bar{z}}{|m|^{1/\alpha}\ell(|m|)}\right) + \frac{1}{k^2}. \quad (1.4.16)$$

Note that as $k \to \infty$, $B_{1/k} \to \mathbb{R}^d$. For vertices $x, z \in \mathbb{Z}^D$ with $\tilde{m} = (x - z) \cdot e_1 > 0$, set

$$g_{x,z} = g\left(\frac{\bar{z} - \bar{x}}{\tilde{m}^{1/\alpha}\ell(\tilde{m})}\right)\tilde{m}^{-d/\alpha}\ell^{-d}(\tilde{m}).$$

This is what the local limit theorem predicts that $\mathbb{P}(x \to z)$ should be. For $z \in B_{1/k}^{(m)}$, using Equation 1.4.16, and the fact that $g(x) \geq 1/k$ on $B_{1/k}$, we have

$$\left|\frac{\mathbb{P}(0 \to z) - g_{0,z}}{g_{0,z}}\right| = \left|\frac{|m|^{d/\alpha}\ell^d(|m|)\mathbb{P}(0 \to z) - |m|^{d/\alpha}\ell^d(|m|)g_{0,z}}{|m|^{d/\alpha}\ell^d(|m|)g_{0,z}}\right| \leq \frac{1/k^2}{1/k} = \frac{1}{k}. \quad (1.4.17)$$

Similar arguments show the same $1/k$ bound when replacing 0 by e_j. The only caveat is that if $j = 1$, then $|m|$ must be replaced by $|m| + 1$, and the corresponding bounds are true for z in

$$\left\{ z \in A_m : \frac{\bar{z}}{(|m|+1)^{1/\alpha}\ell^d(|m|+1)} \in B_{1/k} \right\} \text{ or } \left\{ z \in A_m : \frac{\bar{z} + e_j}{(|m|)^{1/\alpha}\ell^d(|m|)} \in B_{1/k} \right\}$$

for $j = 1$ and $j > 1$ respectively. For $|m|$ large, these sets are nearly identical to $B_{1/k}^{(m)}$.

To be rigorous, we define $\tilde{B}_{1/k}^{(m)}$ as the intersection of the two sets and $B_{1/k}^{(m)}$.

From Equation 1.4.6, we know $|Y_x^m - Y_y^m| = \left|\sum_{z\in A_m} Y_z(\mathbb{P}(x\to z) - \mathbb{P}(y\to z))\right|$. So we bound

$$\mathbb{E}|Y_x^m - Y_y^m| \le \sum_{z\in\tilde{B}_{1/k}^{(m)}} \mathbb{E}|Y_z|\,|\mathbb{P}(x\to z)-\mathbb{P}(y\to z)| + \sum_{z\in A_m\setminus\tilde{B}_{1/k}^{(m)}} \mathbb{E}|Y_z|\,|\mathbb{P}(x\to z)-\mathbb{P}(y\to z)|.$$

Call the first sum (*) and the second (**).

The triangle inequality says that

$$\begin{aligned}\frac{|\mathbb{P}(x\to z)-\mathbb{P}(y\to z)|}{\mathbb{P}(x\to z)} \le{}& \frac{|\mathbb{P}(x\to z) - g_{x,z}|}{g_{x,z}}\,\frac{g_{x,z}}{\mathbb{P}(x\to z)}\\ &+ \frac{|g_{x,z}-g_{y,z}|}{g_{x,z}}\,\frac{g_{x,z}}{\mathbb{P}(x\to z)}\\ &+ \frac{|g_{y,z}-\mathbb{P}(y\to z)|}{g_{y,z}}\,\frac{g_{y,z}}{g_{x,z}}\,\frac{g_{x,z}}{\mathbb{P}(x\to z)}.\end{aligned} \tag{1.4.18}$$

When $z\in\tilde{B}_{1/k}^{(m)}$, $x=0$, and $y=e_j$, the ratio $g_{x,z}/\mathbb{P}(x\to z)$ is bounded below and above by $1-1/k$ and $1+1/k$. As $|m|\to\infty$, the continuity of g tells us that $g_{y,z}/g_{x,z}$ converges to 1 and $|g_{x,z}-g_{y,z}|/g_{x,z}$ goes to 0. Using Equation 1.4.17 and Equation 1.4.18, we have that for $|m|$ sufficiently large,

$$|\mathbb{P}(x\to z)-\mathbb{P}(y\to z)| = \frac{|\mathbb{P}(x\to z)-\mathbb{P}(y\to z)|}{\mathbb{P}(x\to z)}\mathbb{P}(x\to z) \le \frac{3}{k}\mathbb{P}(x\to z). \tag{1.4.19}$$

The constant 3 is because we need a number strictly larger than 2 to accommodate the approximations in Equation 1.4.18.

So using Equation 1.4.19, the first sum (*) can be bounded

$$\sum_{z\in\tilde{B}_{1/k}^{(m)}} \mathbb{E}|Y_z|\,|\mathbb{P}(x\to z)-\mathbb{P}(y\to z)| \le \frac{3}{k}\sum_{z\in\tilde{B}_{1/k}^{(m)}} \mathbb{E}|Y_z|\mathbb{P}(x\to z) \le \frac{3\mathbb{E}|Y_0|}{k}.$$

For the second sum (**), we can use the trivial bound $|\mathbb{P}(x\to z)-\mathbb{P}(y\to z)| \le$

$\mathbb{P}(x \to z) + \mathbb{P}(y \to z)$. As $k \to \infty$, the sets $B_{1/k}$ converge to all of $\mathbb{R}^d$, so the mass associated with $A_m \setminus \tilde{B}_{1/k}^{(m)}$ goes to 0. Thus, $\mathbb{E}|Y_x^m - Y_y^m| \to 0$ as $m \to -\infty$. Since the backwards martingale limit theorem gives convergence of Y_x^m to $Y_x^{-\infty}$ in L^1, we get that

$$\mathbb{E}|Y_x^{-\infty} - Y_y^{-\infty}| \leq \mathbb{E}|Y_x^{-\infty} - Y_x^m| + \mathbb{E}|Y_x^m - Y_y^m| + \mathbb{E}|Y_y^m - Y_y^{-\infty}|.$$

Each of these three terms on the right hand side goes to 0 as $m \to -\infty$, telling us that $\mathbb{E}|Y_x^{-\infty} - Y_y^{-\infty}| = 0$, i.e., $Y_x^{-\infty}$ and $Y_y^{-\infty}$ are almost surely equal.

There is another approach to show $Y_x^{-\infty} = Y_y^{-\infty}$ almost surely which is inspired by the proof of Proposition 1 in Hammond-Sheffield [14]. Briefly, the idea is to construct a coupling of the random walks started at x and y which coalesce with high probability. Note that this is not the standard random walk we work with, which has independent increments prior to coalescence. The sketch is included in Remark A.2. $\square$

1.4.4 Concentration of conditional variance

The other condition of the martingale central limit theorem we use is that

$$V_{n,k_n}^2 = \sum_{m=1}^{k_n} \mathbb{E}[X_{n,m}^2 | \mathcal{F}_{n,m-1}]$$

converges to 1 in probability as $n \to \infty$. We use a second moment method by expressing V_{n,k_n}^2 as a weighted sum of some nonnegative random variables, showing that no coefficient is too large relative to the sum of coefficients, and showing that the random variables in the sum have decaying correlations.

Lemma 1.20. *For $z, z' \in A_{m-1}$, set $c(n, z, z') = \sum_{x \in A_m} (C_x^n)^2\, \mathbb{P}(x \to z)\mathbb{P}(x \to z')$. Then*

$$\frac{\max_{-k_n \leq m \leq n} \sup_{z,z' \in A_{m-1}} c(n, z, z')}{\sum_{m=-k_n}^{n} \sum_{z,z' \in A_{m-1}} c(n, z, z')} \to 0 \text{ as } n \to \infty. \tag{1.4.20}$$

Proof. We use a multi-functional Hölder inequality, which says that if $p_1, p_2, p_3, p_4 \in$

$[1, \infty]$ with $p_1^{-1} + p_2^{-1} + p_3^{-1} + p_4^{-1} = 1$, then $\|f_1 f_2 f_3 f_4\|_1 \leq \|f_1\|_{p_1} \|f_2\|_{p_2} \|f_3\|_{p_3} \|f_4\|_{p_4}$.

We apply this to $c(n, z, z')$ with $f_1(x) = f_2(x) = C_x^n$, $f_3(x) = \mathbb{P}(x \to z)$, and $f_4(x) = \mathbb{P}(x \to z')$. Since $\mathbb{P}(x \to z) \leq 1$ and $\sum_{x \in A_m} \mathbb{P}(x \to z) = 1$,

$$\|f_3\|_{p_3}^{p_3} = \sum_{x \in A_m} \mathbb{P}(x \to z)^{p_3} \leq 1,$$

and similarly for f_4.

We recall Equation 1.4.12 and Equation 1.4.11 from the proof of Lemma 1.16. They tell us that there is some constant C such that

$$\max_{x \in A_m} C_x^n \leq C \text{ and } \sum_{x \in A_m} C_x^n \leq n$$

From here, we know that $f_1(x)^{p_1 - 1} \leq C^{p_1 - 1}$, so $f_1(x)^{p_1} \leq C^{p_1 - 1} f_1(x)$, and $\sum_x f_1(x)^{p_1} \leq C^{p_1 - 1} \sum_x f_1(x)$, giving the bound

$$\|f_1\|_{p_1}^{p_1} \leq C^{p_1 - 1} \sum_{x \in A_m} C_x^n = C^{p_1 - 1} n,$$

and the analogous bound holds for f_2. Putting this together, we get the bound

$$c(n, z, z') = \|f_1 f_2 f_3 f_4\|_1 \leq (C^{p_1 - 1} n)^{1/p_1} (C^{p_2 - 1} n)^{1/p_2} = C^{2 - 1/p_1 - 1/p_2} n^{1/p_1 + 1/p_2}. \tag{1.4.21}$$

The denominator in Equation 1.4.20 is

$$\sum_{m=-k_n}^{n} \sum_{z, z' \in A_{m-1}} \sum_{x \in A_m} (C_x^n)^2 \, \mathbb{P}(x \to z) \mathbb{P}(x \to z') \tag{1.4.22}$$

Combining Equation 1.4.9, Lemma 1.19, and Corollary 1.7, for n sufficiently large, we can lower bound Equation 1.4.22 by $n^{1+\delta}$ for some $\delta > 0$. Since $1/p_1 + 1/p_2 < 1$, combining the lower bound on the denominator with the upper bound on $c(n, z, z')$

from Equation 1.4.21 proves the lemma. $\square$

Lemma 1.21. *For $\varepsilon > 0$, there is some N such that if $z, z' \in A_{m-1}$ and $\tilde{z}, \tilde{z}' \in A_{\tilde{m}-1}$ with $|m - \tilde{m}| > N$ or $\min(|z-\tilde{z}|, |z-\tilde{z}'|, |z'-\tilde{z}|, |z'-\tilde{z}'|) > N$, then*

$$\left|\mathrm{Cov}\left[(Y_z - Y_{z'})^2, (Y_{\tilde{z}} - Y_{\tilde{z}'})^2\right]\right| < \varepsilon.$$

Proof. The main idea is that, with high probability, the random vectors $(Y_z, Y_{z'})$ and $(Y_{\tilde{z}}, Y_{\tilde{z}'})$ have values determined by independent random variables. We approximate $(Y_z, Y_{z'})$ and $(Y_{\tilde{z}}, Y_{\tilde{z}'})$ by random vectors which are truly independent, then show that this approximation is good.

By the symmetry of covariance, we lose no generality in assuming that $\tilde{m} \leq m$. Recall that the random graph was generated by assigning a random vector R_v to every vector in $\mathbb{Z}^D$. The parent of v is $v + R_v$. With $a(v) = v + R_v$, for any vertex, the sequence $(v, a(v), a^2(v), \dots)$, where a^i indicates the i-fold composition of a, is a random walk started at v. Set $A_v = \{a^i(v) : i \geq 0\}$, the set of ancestors of v. The random walks started at two vertices have independent and identical increments unless they intersect, at which point they coalesce and have identical increments.

We augment the probability space with a collection of random vectors $(\bar{R}_v)_{v \in Z^D}$ which is jointly independent of the original probability space. Define $\bar{Y}_z$ and $\bar{Y}_{z'}$ as follows. Set $z_0 = z, z'_0 = z'$. Let $i_0 = \inf\{i \geq 0 : a^i(z) \in A_{\tilde{z}} \cup A_{\tilde{z}'} \text{ or } a^i(z') \in A_{\tilde{z}} \cup A_{\tilde{z}'}\}$. In the event that the components containing z and z' are each disjoint from the components containing $\tilde{z}$ and $\tilde{z}'$, this is $+\infty$. For $i < i_0$, set $z_{i+1} = z_i + R_{z_i}$ and $z'_{i+1} = z'_i + R_{z'_i}$. For $i \geq i_0$, set $z_{i+1} = z_i + \bar{R}_{z_i}$ and $z'_{i+1} = z'_i + \bar{R}_{z'_i}$.

Now define $\bar{Y}_z$ and $\bar{Y}'_z$: if $i_0 = \infty$, then $\bar{Y}_z = Y_z$ and $\bar{Y}_{z'} = Y_{z'}$. If $i_0 < \infty$ and the sequences (z_i) and (z'_i) intersect (and thus coalesce), then $\bar{Y}_z = \bar{Y}_{z'}$ is sampled from the color distribution distribution, independently of all other randomness. And if $i_0 < \infty$ and the sequences are disjoint, sample $\bar{Y}_z$ and $\bar{Y}_{z'}$ independently from the color distribution.

This construction ensures that $(\bar{Y}_z, \bar{Y}_{z'})$ has the same joint distribution as $(Y_z, Y_{z'})$. And on the event $A = \{i_0 = \infty\}$, they are equal: $(\bar{Y}_z, \bar{Y}_{z'}) = (Y_z, Y_{z'})$. We have constructed the vector $(\bar{Y}_z, \bar{Y}_{z'})$ to be independent of $(Y_{\tilde{z}}, Y_{\tilde{z}'})$.

Now, for any function f, we can bound the covariance

$$\begin{aligned}
\mathrm{Cov}[f(Y_z, Y_{z'}), f(Y_{\tilde{z}}, Y_{\tilde{z}'})] &= |\mathbb{E}f(Y_z, Y_{z'})f(Y_{\tilde{z}}, Y_{\tilde{z}'}) - \mathbb{E}f(Y_z, Y_{z'})\mathbb{E}f(Y_{\tilde{z}}, Y_{\tilde{z}'})| \\
&\leq |\mathbb{E}f(Y_z, Y_{z'})f(Y_{\tilde{z}}, Y_{\tilde{z}'}) - \mathbb{E}f(Y_z, Y_{z'})f(Y_{\tilde{z}}, Y_{\tilde{z}'})\mathbf{1}_A| \\
&\quad + \left|\mathbb{E}f(Y_z, Y_{z'})f(Y_{\tilde{z}}, Y_{\tilde{z}'})\mathbf{1}_A - \mathbb{E}f(\bar{Y}_z, \bar{Y}_{z'})f(Y_{\tilde{z}}, Y_{\tilde{z}'})\mathbf{1}_A\right| \\
&\quad + \left|\mathbb{E}f(\bar{Y}_z, \bar{Y}_{z'})f(Y_{\tilde{z}}, Y_{\tilde{z}'})\mathbf{1}_A - \mathbb{E}f(\bar{Y}_z, \bar{Y}_{z'})f(Y_{\tilde{z}}, Y_{\tilde{z}'})\right| \\
&\quad + \left|\mathbb{E}f(\bar{Y}_z, \bar{Y}_{z'})f(Y_{\tilde{z}}, Y_{\tilde{z}'}) - \mathbb{E}f(Y_z, Y_{z'})\mathbb{E}f(Y_{\tilde{z}}, Y_{\tilde{z}'})\right|
\end{aligned}$$

The second term is 0 because $(\bar{Y}_z, \bar{Y}_{z'}) = (Y_z, Y_{z'})$ on A. The fourth term is 0 because $(\bar{Y}_z, \bar{Y}_{z'})$ is independent of $(Y_{\tilde{z}}, Y_{\tilde{z}'})$ and because $(\bar{Y}_z, \bar{Y}_{z'})$ has the same joint distribution as $(Y_z, Y_{z'})$. The first and third are handled similarly, so we only describe the first, which we bound as

$$\begin{aligned}
|\mathbb{E}f(Y_z, Y_{z'})f(Y_{\tilde{z}}, Y_{\tilde{z}'}) - \mathbb{E}f(Y_z, Y_{z'})f(Y_{\tilde{z}}, Y_{\tilde{z}'})\mathbf{1}_A| &= |\mathbb{E}f(Y_z, Y_{z'})f(Y_{\tilde{z}}, Y_{\tilde{z}'})\mathbf{1}_{A^c}| \\
&\leq \mathbb{E}\,|f(Y_z, Y_{z'})f(Y_{\tilde{z}}, Y_{\tilde{z}'})\mathbf{1}_{A^c}| \\
&\leq \mathbb{E}\left[|f(Y_z, Y_{z'})f(Y_{\tilde{z}}, Y_{\tilde{z}'})|\middle|A^c\right]\mathbb{P}(A^c) \\
&\leq \mathbb{P}(A^c)\sqrt{\mathbb{E}[f(Y_z, Y_{z'})^2|A^c]\,\mathbb{E}[f(Y_{\tilde{z}}, Y_{\tilde{z}'})^2|A^c]}.
\end{aligned}$$

With $f(a,b) = (a-b)^2$, we use the general bound $(a+b)^4 \leq 16a^4 + 16b^4$ to bound

$$\mathbb{E}[(Y_z - Y_{z'})^4|A^c] \leq 16\,\mathbb{E}[Y_z^4|A^c] + 16\,\mathbb{E}[Y_{z'}^4|A^c].$$

But the marginal distributions of Y_z and $Y_{z'}$ are just the color distribution, so it suffices that the color distribution has a finite fourth moment. So

$$\sqrt{\mathbb{E}[(Y_z, Y_{z'})^4 | A^c]\, \mathbb{E}[(Y_{\tilde{z}}, Y_{\tilde{z}'})^4 | A^c]}$$

can be bounded by some absolute constant depending only on the color distribution.

At this point, we have that $\mathrm{Cov}[(Y_z, Y_{z'})^2, (Y_{\tilde{z}}, Y_{\tilde{z}'})^2]$ is bounded by some absolute constant times $\mathbb{P}(A^c)$. We must show that if $|m - \tilde{m}|$ is large, or if $\min(|z - \tilde{z}|, |z - \tilde{z}'|, |z' - \tilde{z}|, |z' - \tilde{z}'|)$ is large, then $\mathbb{P}(A^c)$ is small.

First, consider what happens as $|m - \tilde{m}| \to \infty$. It suffices to show that as $|m - \tilde{m}| \to \infty$,

$$\sup_{v \in A_m, \tilde{v} \in A_{\tilde{m}}} |\mathbb{P}(v \sim \tilde{v})| \to 0.$$

Since we have assumed that $\tilde{m} \leq m$, using arguments analogous to those in the proof of Proposition 1.5, $\mathbb{P}(v \sim \tilde{v}) = C \sum_{i=0}^{\infty} \mathbb{P}(Z_{|m-\tilde{m}|+i} + \tilde{Z}_i = a)$, where C is some positive constant and a is the projection of $v - \tilde{v}$ onto the last d coordinates. By the local limit theorem, since the density is bounded,

$$\mathbb{P}(Z_{|m-\tilde{m}|+i} + \tilde{Z}_i = a) \leq C' |m - \tilde{m} + 2i|^{-d/\alpha} \ell^{-d}(|m - \tilde{m} + 2i|).$$

By the arguments in the proof of Proposition 1.5, this is summable, so as $|m - \tilde{m}| \to \infty$, this goes to 0. Thus, $\mathbb{P}(A^c)$ can be made arbitrarily small by making $|m - \tilde{m}|$ large.

Next, we show that it is also sufficient that $\min(|z - \tilde{z}|, |z - \tilde{z}'|, |z' - \tilde{z}|, |z' - \tilde{z}'|)$ is large. It suffices to show that as $|z - \tilde{z}| \to \infty$, $\mathbb{P}(z \sim \tilde{z}) \to 0$. Since the underlying random walks are transient, $\mathbb{P}(Z_i - \tilde{Z}_i = x \text{ for some } i \geq 0) \to 0$ as $|x| \to \infty$. We really want to show that $\mathbb{P}(Z_{|m-\tilde{m}|+i} - \tilde{Z}_i = \mathrm{proj}(z - \tilde{z}) \text{ for some } i \geq 0) \to 0$ as $|\mathrm{proj}(z - \tilde{z})| \to \infty$, where the projection is onto the last d coordinates. Since we have taken care of the case where $|m - \tilde{m}|$ goes to ∞, we may assume that $|m - \tilde{m}|$ is bounded as $|z-\tilde{z}| \to \infty$. But $\max_{k\leq|m-\tilde{m}|} \mathbb{P}(|Z_k| > t) \to 0$ as $t \to \infty$, and $Z_{|m-\tilde{m}|+i}$ has the same distribution as $Z_i + Z'_{|m-\tilde{m}|}$ for an independent copy $(Z'_k)_{k\geq 0}$. So for any $t > 0$, we can upper bound $\mathbb{P}(Z_{|m-\tilde{m}|+i} - \tilde{Z}_i = \text{proj}(z-\tilde{z}) \text{ for some } i \geq 0)$ by

$$\mathbb{P}(|Z'_{|m-\tilde{m}|}| \geq t) + \mathbb{P}(|Z'_{|m-\tilde{m}|}| < t, Z_i - \tilde{Z}_i = \text{proj}(z-\tilde{z}) - Z'_{|m-\tilde{m}|} \text{ for some } i \geq 0).$$

When $\text{proj}(z-\tilde{z}) \to \infty$, given that $|Z'_{|m-\tilde{m}|}| < t$, we get that $\text{proj}(z-\tilde{z}) + Z'_{|m-\tilde{m}|} \to \infty$ as well. So for fixed t, $\mathbb{P}(|Z'_{|m-\tilde{m}|}| \geq t) + \mathbb{P}(|Z'_{|m-\tilde{m}|}| < t, Z_i - \tilde{Z}_i = \text{proj}(z-\tilde{z}) - Z'_{|m-\tilde{m}|} \text{ for some } i \geq 0)$ goes to 0 as $\text{proj}(z-\tilde{z}) \to \infty$, showing that $\mathbb{P}(Z_{|m-\tilde{m}|+i} - \tilde{Z}_i = \text{proj}(z-\tilde{z}) \text{ for some } i \geq 0)$ is asymptotically bounded by $\max_{k\leq|m-\tilde{m}|} \mathbb{P}(|Z_k| > t)$. But t was arbitrary, so we can take t to ∞ to get that $\mathbb{P}(Z_{|m-\tilde{m}|+i} - \tilde{Z}_i = \text{proj}(z-\tilde{z}) \text{ for some } i \geq 0)$ goes to 0. Thus, as $|z-\tilde{z}| \to \infty$, we have $\mathbb{P}(z \sim \tilde{z}) \to 0$. Finally, we conclude that $\mathbb{P}(A^c)$ can be made arbitrarily small by making $\min(|z-\tilde{z}|, |z-\tilde{z}'|, |z'-\tilde{z}|, |z'-\tilde{z}'|)$ large. □

We combine the previous two facts in a standard second moment argument.

Lemma 1.22. *Let I_n be a sequence of countable index sets. Suppose that for $n \geq 1$ and $i \in I_n$, we have $a_{n,i} \geq 0$, $X_i \geq 0$, $0 < c \leq \mathbb{E}X_i$, and $\text{Var}[X_i] \leq \sigma^2$. Set $S_n = \sum_{i\in I_n} a_{n,i}X_i$ and $M_n = \max_{i\in I_n} a_{n,i}$. Assume*

$$\lim_{n\to\infty} \frac{M_n}{\sum_{i\in I_n} a_{n,i}} = 0, \tag{1.4.23}$$

and that for all $\varepsilon > 0$, there is some N such that for every i, there is a B_i with $|B_i| \leq N$ and for all $j \notin B_i$, $\text{Cov}[X_i, X_j] \leq \varepsilon$.

Then S_n is asymptotically concentrated around its mean in the sense that for all $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|S_n - \mathbb{E}S_n| \geq \delta\, \mathbb{E}S_n) = 0.$$

Proof. We break down

$$\operatorname{Var}[S_n] = \sum_{i \in I_n} \sum_{j \in I_n \cap B_i} a_{n,i} a_{n,j} \operatorname{Cov}[X_i, X_j] + \sum_{i \in I_n} \sum_{j \in I_n \setminus B_i} a_{n,i} a_{n,j} \operatorname{Cov}[X_i, X_j].$$

We bound the first sum by

$$M_n N \sigma^2 \sum_{i \in I_n} a_{n,i},$$

since $|B_i| \leq N$, since $\operatorname{Cov}[X_i, X_j] \leq \sigma^2$, and since $a_{n,j} \leq M_n$. We bound the second sum by

$$\varepsilon \sum_{i \in I_n} \sum_{j \in I_n \setminus B_i} a_{n,i} a_{n,j} \leq \varepsilon \Big(\sum_{i \in I_n} a_{n,i} \Big)^2,$$

since $\operatorname{Cov}[X_i, X_j] \leq \varepsilon$ when $j \notin B_i$. We also know $\mathbb{E} S_n \geq c \sum_{i \in I_n} a_{n,i}$. So $\operatorname{Var}[S_n] \leq M_n N \sigma^2 \mathbb{E} S_n / c + \varepsilon (\mathbb{E} S_n / c)^2$.

Chebyshev's inequality tells us that

$$\mathbb{P}(|S_n - \mathbb{E} S_n| \geq \delta \, \mathbb{E} S_n) \leq \frac{\operatorname{Var}[S_n]}{\delta^2 (\mathbb{E} S_n)^2} \leq \frac{M_n N \sigma^2 \mathbb{E} S_n / c + \varepsilon (\mathbb{E} S_n / c)^2}{\delta^2 (\mathbb{E} S_n)^2} = \frac{M_n N \sigma^2}{c \delta^2 \mathbb{E} S_n} + \frac{\varepsilon}{c^2 \delta^2}.$$

Since $\frac{M_n}{\sum_{i \in I_n} a_{n,i}} \to 0$ and $\mathbb{E} S_n \geq c \sum_{i \in I_n} a_{n,i}$, the term $\frac{M_n N \sigma^2}{c \delta^2 \mathbb{E} S_n}$ goes to 0 as $n \to \infty$, showing that $\lim_{n \to \infty} \mathbb{P}(|S_n - \mathbb{E} S_n| \geq \delta \, \mathbb{E} S_n) \leq \frac{\varepsilon}{c^2 \delta^2}$. But ε was arbitrary, so the limit is 0. □

Corollary 1.23. *If $\delta > 0$, with $k_n = \lceil n^{1+\delta} \rceil$, we have concentration of conditional variance:*

$$V_{n,k_n}^2 = \sum_{m=1}^{k_n} \mathbb{E}[X_{n,m}^2 | \mathcal{F}_{n,m-1}] \to 1$$

in probability as $n \to \infty$.

Proof. Corollary 1.13 gives a decomposition of V_{n,k_n}^2 as the sum of correlated random variables $(Y_z - Y_{z'})^2$ times some coefficients

$$\operatorname{Var}[S_n]^{-1} c(n, z, z') = \operatorname{Var}[S_n]^{-1} \sum_{x \in A_m} (C_x^n)^2 \, \mathbb{P}(x \to z) \mathbb{P}(x \to z').$$

Lemma 1.20 tells us that the assumption in Equation 1.4.23 of Lemma 1.22 is satisfied, and Lemma 1.21 says that the covariance decay assumption of Lemma 1.22 is satisfied. We apply Lemma 1.22 to conclude that V_{n,k_n}^2 concentrates around its mean.

We know $\operatorname{Var}[S_n] \, \mathbb{E} V_{n,k_n}^2 = \operatorname{Var}[S_n - S_n^{-k_n}]$. Lemma 1.19 tells us that as $n \to \infty$, $\operatorname{Var}[S_n^{-k_n}] / \operatorname{Var}[S_n] \to 0$, so

$$\frac{\operatorname{Var}[S_n - S_n^{-k_n}]}{\operatorname{Var}[S_n]} \to 1.$$

This lets us conclude that the conditional variance goes to 1 in probability as $n \to \infty$. $\square$

This completes the proof of asymptotic gaussianity of $S_n / \operatorname{Var} S_n$. We briefly indicate the modifications needed for full joint Gaussianity. When following the proof, the only thing that changes is the definition of C_x^n. If we have a sequence $0 < c_1 < c_2 < \cdots < c_r$ and coefficients $a_1, \ldots, a_r$ and want to prove the asymptotic Gaussianity of

$$S_n = a_1 \sum_{i=1}^{c_1 n} Y_{ie_1} + a_2 \sum_{i=c_1 n}^{c_2 n} Y_{ie_1} + \cdots + a_r \sum_{i=c_{r-1} n}^{c_r n} Y_{ie_1},$$

following the analysis leading to Equation 1.4.4 shows that the appropriate definition is

$$C_x^n = a_1 \sum_{i=1}^{c_1 n} \mathbb{P}(ie_1 \to x) + a_2 \sum_{i=c_1 n}^{c_2 n} \mathbb{P}(ie_1 \to x) + \cdots + a_r \sum_{i=c_{r-1} n}^{c_r n} \mathbb{P}(ie_1 \to x).$$

The key proofs to modify are those of Lemma 1.16, Lemma 1.17, and Lemma 1.20. Everything else is virtually unchanged.

1.5 Conclusion

We have shown that the stochastic process has the correct asymptotic variance and is asymptotically Gaussian. Since the process has stationary increments, this is sufficient to conclude that the processes converge to a fractional Brownian motion. The decay of coalescence probabilities of heavy-tailed random walks is key for this model to give the correlation decay of a fractional Brownian motion.

We briefly mention two interesting directions for future study. First, is there a way to define a joint process which doesn't give only fractional Brownian motion when summing along a ray, but instead gives a fractional Gaussian field? Second, is there a way to give a model negatively correlated increments such that the limiting process is a fractional Brownian motion with Hurst parameter less than $1/2$?

CHAPTER 2

COALESCING BRANCHING RANDOM WALKS

The contents of this chapter include joint work with Michael Damron and Bounghun Bock.

2.1 Introduction

We start by introducing the model and some notation, following the work of Bock, Damron, and Hanson [3]. Throughout this chapter, the underlying graph is the infinite d-dimensional integer lattice, whose vertex set is $\mathbb{Z}^d$, with edges between vertices which are distance 1 apart. However, the definitions apply to any simple graph (i.e., a graph which has no loops or multi-edges), so we state the definitions generally.

Let $G = (V, E) = (V(G), E(G))$ be a simple graph. Let $(\omega(e))_{e \in E}$ be a collection of real valued random variables, thought of as random edge weights for the edges of G. They are not assumed to be independent. Letting $\langle x, y \rangle$ denote a directed edge from x to y and (x, y) an undirected edge between x and y, we define the directed nearest neighbor graph $\mathcal{N}_D$ as the graph whose vertex set is $V(G)$ and whose edge set is

$$\{\langle x, y \rangle : (x, y) \in E(G) \text{ and } \omega((x, y)) \leq \omega((x, z)) \text{ for all } z \text{ with } (x, z) \in E(G)\}.$$

In words, at each vertex, there is a directed edge pointing to its nearest neighbor (or all nearest neighbors if there is a tie for the minimum incident edge weight). If $x \in V$ has finite degree in G, then there is an incident edge with minimum weight, and so x has out-degree at least 1 in $\mathcal{N}_D$ (provided x has degree at least 1 in G). If all the weights $(\omega(e))_{e\in E}$ are distinct, then every vertex has out-degree at most 1 in $\mathcal{N}_D$.

We generalize this by defining the directed k-nearest neighbors graph k-$\mathcal{N}_D$ in which there is a directed edge from each vertex to its nearest k neighbors, and recover $\mathcal{N}_D$ as 1-$\mathcal{N}_D$. Specifically, k-$\mathcal{N}$ is defined as the directed graph with vertex set $V(G)$, and $\langle x, y\rangle$ is in the edge set if $(x, y) \in E(G)$ and $\omega((x, y)) \leq \omega((x, z))$ for all but at most $k-1$ vertices z with $(x, z) \in E(G)$.

So there is an edge from x to y if the weight of the edge between x and y is a lowest weight edge after discarding up to $k-1$ smaller edges. When every vertex has finite degree in G, every vertex of k-$\mathcal{N}_D$ has out-degree at least k (provided the vertices have degree at least k in G). When all edge weights are distinct, we see that every vertex of k-$\mathcal{N}_D$ has out-degree at most k. Infinite degree vertices may fail to have a minimum weight edge.

Lastly, we define the (undirected) nearest neighbors graph $\mathcal{N}$ and the (undirected) k-nearest neighbors graph k-$\mathcal{N}$ as the undirected versions of $\mathcal{N}_D$ and k-$\mathcal{N}_D$. A goal is to characterize which graphs can be equal to k-$\mathcal{N}_D$ or k-$\mathcal{N}$ for some set of edge weights. For various theorems, we may make various assumptions on the edge weight distribution. In order of increasing strength, we may make no assumption on the edge weight distribution, we may assume that the edge weights are almost surely distinct and that the distribution is translation invariant, we may assume that the edge weights are almost surely distinct and that the distribution is translation invariant and ergodic, or we may assume that the edge weights are i.i.d. and continuous.

2.2 Coalescing branching random walks

Continuing with the theme of this thesis, the use of random walks and coalescence as a tool, we start with a proof that if $k \geq d+2$ and the edge weight distribution is i.i.d. and continuous, then almost surely every vertex of k-$\mathcal{N}$ is in the same component. We achieve this by defining a collection of coalescing branching random walks, one started at each vertex, and show that the random walks at each pair of vertices coalesce with probability 1.

In the case $k \geq d+2$ with i.i.d. weights, we can take advantage of the independence of the edges in the positive direction to define a coalescing branching random walk whose analysis will let us show that the infinite component is unique.

As a warm-up, we prove that if $d \leq 3$, then it suffices that $k \geq d+1$. The proof strategy fails for $d > 3$ due to the simple random walk being transient in higher dimensions. However, the failure suggests a way to proceed: if we ask that $k \geq d+2$, then we can analyze things using a natural coalescing branching random walk, which we show is recurrent in all dimensions.

Theorem 2.1. *Assume the edge weights are independently and identically distributed Uniform*$[0,1]$ *random variables. For $d \leq 3$, if $k \geq d+1$, there is almost surely a unique infinite connected component in k-$\mathcal{N}_D$. Moreover, every vertex is almost surely in this component.*

Proof. Almost surely, all edge weights are distinct, so we restrict ourselves to outcomes in the probability 1 event that all edge weights are distinct. For $x \in \mathbb{Z}^d$, set $N_x = \{\langle x, x+\mathbf{e}_i\rangle : 1 \leq i \leq d\}$, where $\mathbf{e}_i$ is the standard basis vector in the i-th direction. We refer to these as the edges at x in the positive direction. Since each vertex in $\mathbb{Z}^d$ has degree $2d$, there must be at least $k-d$ of the edges from N_x present in k-$\mathcal{N}_D$. Moreover, by looking at only the weights of the edges in N_x, we can identify $k-d$ edges which are necessarily present in k-$\mathcal{N}_D$: the $k-d$ edges of lowest weight. For $k \geq d+1$, define k-$\mathcal{N}^+$ to be the subgraph of k-$\mathcal{N}_D$ which, for each vertex x, contains exactly the $k-d$ lowest weight edges from N_x. By construction, the set of edges starting at x and the set of edges starting at y in k-$\mathcal{N}^+$ are independent random sets for all distinct vertices x and y. If we show that every vertex from k-$\mathcal{N}^+$ is in the same connected component, then we can conclude that there is a unique infinite component in k-$\mathcal{N}_D$.

With this setup, we prove uniqueness when $d \leq 3$. If we couple the graphs 1-$\mathcal{N}_D, 2\text{-}\mathcal{N}_D, \ldots, 2d\text{-}\mathcal{N}_D$ in the natural way, each is a subgraph of the next. So if every vertex in $(d+1)$-$\mathcal{N}_D$ is almost surely in the same connected component, this is also true for all $k \geq d+1$. We therefore assume without loss of generality that $k = d+1$. To show that all vertices are in the same component, we must show that every pair of vertices is in the same component. So fix two arbitrary vertices x and y.

In k-$\mathcal{N}^+$, every vertex has out-degree one, so there is a unique directed path starting at each vertex. Define $x_0 = x$ and x_i to be the vertex obtained by walking along the directed path started at x for i steps. Define $(y_i)_{i \geq 0}$ analogously. It suffices to show that the directed paths started at x and y intersect with probability 1.

For $z \in \mathbb{Z}^d$, let $c(z)$ be the sum of the components: if $z = (z_1, z_2, \ldots, z_d) \in \mathbb{Z}^d$, then $c(z) = z_1 + \cdots + z_d \in \mathbb{Z}$. We may assume that $c(x) = c(y)$; if not, the vertex for which the sum is lower may be replaced by a vertex in its path which makes the sums equal. Define $w_i = x_i - y_i$. We note that $c(w_i) = 0$ for all i. If we have $w_i = 0$ for any i, then the directed paths started at x and y intersect, showing that x and y are in the same component.

Consider the distribution of the process $(w_i)_{i \geq 0}$. If $w_i = 0$, then $x_i = y_i$, which tells us that $x_{i+1} = y_{i+1}$ and thus $w_{i+1} = 0$. If $w_i \neq 0$, then $x_i \neq y_i$. When this holds, the increments $x_{i+1} - x_i$ and $y_{i+1} - y_i$ are identically distributed and, as they are defined in terms of disjoint sets of edge weights, are independent. Each such increment $x_{i+1} - x_i$ is uniformly distributed in $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$, the standard basis vectors. So when $w_i \neq 0$, $w_{i+1} - w_i$ is 0 with probability $1/d$ and otherwise is uniformly distributed on the $d(d-1)$ vectors which have one $+1$ entry and one -1 entry. And when $w_i = 0$, $w_{i+1} = 0$ deterministically. It suffices to show that this process hits 0 almost surely when started at an arbitrary initial condition w_0 with $c(w_0) = 0$.

We see that w_i defines a (lazy) random walk on the graph $L_0(\mathbb{Z}^d)$ whose vertex set is $\{v \in \mathbb{Z}^d : c(v) = 0\}$ and whose vertices are adjacent if their entries differ by 1 in two coordinates and are equal elsewhere. Lazy means there is a positive probability that $w_{i+1} = w_i$. And the origin is an absorbing state for this random walk. In $d = 1$, there is nothing to show, as the graph contains only the vertex 0. In $d = 2$, this graph is isomorphic to $\mathbb{Z}$. In $d = 3$, this is isomorphic to the two-dimensional triangular lattice (sometimes also called the hexagonal lattice) in which each vertex has degree 6.

It is well-known that, under basic assumptions (mean zero, finite variance, independent increments), random walks in 1 and 2 dimensions are recurrent, so the random walk visits the origin with probability 1. Thus, almost surely, every pair of vertices is in the same connected component of k-$\mathcal{N}^+$, showing that there is a unique infinite component in k-$\mathcal{N}_D$.

To make this claim about recurrence on the triangular lattice more formal, and also to preview a proof technique for the higher dimensional case, we can identify the vertices of $L_0(\mathbb{Z}^3)$ with the vertices of $\mathbb{Z}^2$ by considering only the first 2 coordinates of each vertex in $L_0(\mathbb{Z}^3)$. Since the sum of the coordinates is 0, this uniquely determines the third coordinate. Theorem 4.1.1 of [20], amongst countless other references, states recurrence for aperiodic random walks on $\mathbb{Z}^d$ whose independent increments have mean zero and finite variance. The increment of the first two coordinates is 0 with probability 1/3 and is otherwise uniformly distributed on $\{(1,0),(-1,0),(0,1),(0,-1),(1,-1),(-1,1)\}$. This distribution has mean zero and finite variance. Because it is lazy and its support spans $\mathbb{Z}^2$, it is aperiodic. Thus, it is recurrent. $\square$

Observe how this proof fails for $d > 3$. The set of vertices whose sum is 0 forms a $(d-1)$-dimensional subspace. Random walks on more than 3 dimensions are transient, so there is a positive probability that the random walks in the positive direction do not hit the origin. This suggests a modification: if we take $k = d + 2$, there are 2 edges we could potentially follow in the positive direction. This corresponds to a coalescing branching random walk with branching factor of 2. As an interlude, we make some formal definitions about coalescing branching random walks and prove some facts about them.

Definition 2.2. *Let $k \geq 2$ and G be a graph with minimum degree at least k. For $v \in V(G)$ and $i \geq 0$, let $N_i^{(k)}(v)$ be a random set of size k which is chosen uniformly amongst the size k subsets of the neighbors of v. The random variables $(N_i^{(k)}(v))_{i \geq 0, v \in V(G)}$ are assumed to be independent. Using this, we define a set-valued random process $(X_i)_{i \geq 0}$ where each $X_i \subseteq V(G)$.*

Let X_0 be a nonempty subset of $V(G)$. Given X_i, define

$$X_{i+1} = \bigcup_{v \in X_i} N_i^{(k)}(v). \tag{2.2.1}$$

*A **coalescing branching random walk** with branching factor k on G is any random process which has the distribution of $(X_i)_{i \geq 0}$ for some nonempty choice of X_0.*

To explain the "coalescing branching" part of the name, we take the interpretation that for $x_i \in X_i$, the k neighbors $N_i^{(k)}(x_i)$ are the children of x_i. Thus, moving from X_i to X_{i+1}, every vertex branches into k children. If we defined a process where there could be multiple particles occupying a given vertex, we would have a branching random walk. Instead, X_{i+1} is defined as a union of the children of the vertices in X_i. Thus, if two distinct vertices $x_i, x_i' \in X_i$ both have the same child, only one copy of that child exists in X_{i+1}. This explains the "coalescing" part of the name.

The following theorem resolves an open question posed in Blitz [2].

Theorem 2.3. *When $k = 2$, every coalescing branching random walk on $\mathbb{Z}^d$ almost surely occupies the origin infinitely often.*

The theorem follows as a corollary of Lemma 2.5 and Theorem 2.6.

Since a point in a coalescing branching random walk has k children, a natural thing to do is to "follow" just one child. At every time step, there are k children to choose from. Choosing which child to follow allows a controller to exert influence.

Definition 2.4. *Fix a graph G. For $v \in V(G)$, let $N(v)$ be the neighbors of v. A* ***policy*** *on a graph G is a function π which maps from $V(G) \times V(G)^\infty$ to the space of probability distributions on $V(G)$ such that*

$$\mathrm{supp}(\pi(v\,;(v_i))) \subseteq N(v).$$

For $\varepsilon > 0$, an ***ε-controlled random walk*** *on a graph G, controlled by policy π, is a stochastic process $(v_i)_{i\in\mathbb{Z}}$ on $V(G)$ such that with probability $1-\varepsilon$, v_{i+1} is chosen uniformly amongst the neighbors of v_i and with probability ε, v_{i+1} is chosen according to the policy:*

$$v_{i+1} \sim \pi(v_i\,;(v_k)_{k\le i}).$$

We see that a policy takes the current state of the random walk and the history of the random walk and outputs a probability distribution on the neighbors of the current state. With probability ε, this next state is sampled according to the probability distribution; otherwise, the next state is a uniform random neighbor. Strictly speaking, the definition of a policy requires an infinite history. We abuse the definition slightly to allow π to accept a finite history.

Lemma 2.5. *Suppose G has minimum degree at least k and finite maximum degree. Then the coalescing branching random walk with branching factor k on G started at $X_0 = \{v\}$ stochastically dominates every ε-controlled random walk started at v with ε sufficiently small.*

Proof. We couple the coalescing branching random walk $(X_i)_{i\ge 0}$ and the ε-controlled random walk $(x_i)_{i\ge 0}$ so that $x_i \in X_i$ for every $i \ge 0$.

Suppose $x_i \in X_i$. Set $\pi_i = \pi(x_i\,; (x_k)_{k\leq i})$. Sample

$$v_i^* \sim \pi_i.$$

If v_i^* is in $N_i^{(k)}(x_i)$, then set $x_{i+1} = v_i^*$. Otherwise, set x_{i+1} equal to a random element of $N_i^{(k)}(x_i)$. The probability that v_i^* is in $N_i^{(k)}(x_i)$ is $\frac{k}{\deg(x_i)}$. So the probability that $x_{i+1} = v$ is

$$\pi_i(v)\frac{k}{\deg(x_i)} + (1-\pi_i(v))\left(1 - \frac{k}{\deg(x_i)}\right)\left(\frac{1}{\deg(x_i)-1}\right).$$

Since $\deg(x_i)$ is bounded (the graph is assumed to have finite maximum degree), this is stronger than the ε control for ε sufficiently small. To account for the fact that the control is too strong, we can modify the policy distribution by mixing it with the uniform distribution to achieve exactly an ε level of control. And by construction, $x_{i+1} \in X_{i+1}$. □

Theorem 2.6. *For any $\varepsilon > 0$, there is a policy π under which the ε-controlled random walk on $\mathbb{Z}^d$ hits the origin infinitely often.*

Proof. We present the proof in full detail for $d = 2$. The proof for $d > 2$ is not conceptually different, but is notationally cumbersome. We use $(z_i)_{i\geq 0}$ for the controlled random walk. The walk may be started at an arbitrary vertex z_0.

Using the policy, we bias the first coordinate of z_1 towards 0. Once it hits 0, we switch to biasing the second coordinate towards 0. In $d = 2$, we alternate between minimizing the two coordinates. (For $d > 2$, the appropriate thing to do is to cycle through minimizing all d coordinates.) Using the history, we can tell which coordinate the policy should try to minimize. If all coordinates are 0 (and z_i is at the origin), the policy just chooses a uniform neighbor.

Consider the process that the coordinate being minimized experiences. With probability ε, the policy moves it towards 0. And with probability $1-\varepsilon$, the particle z_i moves to a uniform random neighbor. In a uniform random movement, the coordinate changes by $+1$ with probability $1/2d$ and -1 with probability $1/2d$. With probability $1-1/d$, the coordinate is unchanged.

If the current value of the coordinate being minimized is n, the expected number of steps it takes to hit 0 is $C(\varepsilon)|n|$ for some finite constant $C(\varepsilon) < \infty$. This is linear in $|n|$.

Since the transition probabilities only depend on the current location $z_i \in \mathbb{Z}^d$ and on which coordinate is being minimized, the controlled random walk defines a Markov chain, with state space $\mathbb{Z}^d \times \{1, 2, \ldots, d\}$, where the last entry keeps track of which coordinate is being minimized.

Let's analyze the chain. We define $t_n^{(1)}$ to be the index corresponding to the n-th time at which the first coordinate hits 0 while being minimized. (If the second coordinate is being minimized and the first coordinate hits 0 by chance, this does not count). Similarly, let $t_n^{(2)}$ be the index for the n-th time the second coordinate of $(z_i)_{i\geq 0}$ hits 0 while being minimized. Observe that

$$0 \leq t_1^{(1)} \leq t_1^{(2)} \leq t_2^{(1)} \leq t_2^{(2)} \leq \ldots.$$

We define $a_n^{(1)}$ to be the absolute value of the second coordinate of z_i when the first coordinate of z_i hits 0 for the n-th time:

$$a_n^{(1)} = |z_{t_n^{(1)}} \cdot \mathbf{e}_2|.$$

Similarly, define

$$a_n^{(2)} = |z_{t_n^{(2)}} \cdot \mathbf{e}_1|.$$

Finally, let $(a_n)_{n\geq 1}$ be the sequence which alternates between terms from $(a_n^{(1)})_{n\geq 1}$

and $(a_n^{(2)})_{n\geq 1}$:

$$a_1 = a_1^{(1)}, a_2 = a_1^{(2)}, a_3 = a_2^{(1)}, a_4 = a_2^{(2)}, \text{ etc.}$$

Note that when $a_i = 0$, we must also have $z_{i'} = 0$ for the index i' which corresponds to the appropriate $t_j^{(k)}$.

Observe that $(a_n)_{n\geq 1}$ is a random process which satisfies the Markov property. When $a_k \neq 0$, the distribution of a_{k+1} is that of $|S_N|$ where $S_k = X_1 + \cdots + X_k$, X_i are i.i.d. mean zero random variables which take value in $\{-1, 0, +1\}$, and N is a random variable that corresponds to the time between consecutive $t_k^{(1)}$ and $t_k^{(2)}$ or $t_k^{(2)}$ and $t_{k+1}^{(1)}$. We noted that N has mean $C(\varepsilon)a_k$. And by Cauchy-Schwarz,

$$\mathbb{E}[|S_n|] \leq \sqrt{\mathbb{E}[S_n^2]} = C'\sqrt{n}.$$

Then, using Jensen's inequality,

$$\mathbb{E}[|S_N|] = \mathbb{E}\,\mathbb{E}[|S_N| \mid N] \leq C'\mathbb{E}[\sqrt{N}] \leq C'\sqrt{\mathbb{E}N}.$$

So

$$\mathbb{E}[a_{k+1}|a_k] \leq C'\sqrt{C(\varepsilon)a_k}.$$

We see that there is a finite constant C for which $\mathbb{E}[a_{k+1}|a_k] \leq C\sqrt{a_k}$. If $a_k \geq 4C^2$, then $\mathbb{E}[a_{k+1}|a_k]/a_k \leq C(a_k)^{-1/2} \leq 1/2$, so $(a_k)_{k\geq 1}$ decays exponentially in expectation until a_k falls below $4C^2$.

Let $i_0 = \lceil \log_2(a_k/4C^2) \rceil$. Then $\mathbb{E}[a_{k+i_0}|a_k] \leq 4C^2$. By Markov's inequality, $\mathbb{P}(a_{k+i_0} \geq 8C^2|a_k) \leq 1/2$. Because $(a_k)_{k\geq 0}$ is a Markov process, almost surely a_k is below $8C^2$ infinitely often. And when $a_k < 8C^2$, the probability that $a_{k+1} = 0$ is bounded away from 0. This lets us conclude that a_k hits 0 infinitely often, and thus so does the walk $(z_i)_{i\geq 1}$.

This proves things for $d = 2$. The proof idea is almost identical for $d > 2$, but notation makes it much more cumbersome. The full details are written out in the proof of Lemma 2.8. $\square$

In the proof of the following theorem, as well as the lemma which follows it, we see how the ideas of controlled random walks and coalescing branching random walks can arise naturally in the analysis of other processes.

Theorem 2.7. *Assume the edge weights are independently and identically distributed Uniform*$[0,1]$ *random variables. If* $k \geq d+2$*, then there is almost surely a unique infinite connected component in* k-$\mathcal{N}_D$*. Moreover, every vertex is almost surely in this component.*

Proof. We condition on the probability 1 event that all edge weights are distinct so that the k minimum edges are well-defined. Recall the definitions of N_x, k-$\mathcal{N}^+$, $c(x)$, and $L_0(\mathbb{Z}^d)$ from the proof of Theorem 2.1. It suffices to prove the case $k = d+2$. Fix two vertices x and y which we may assume satisfy $c(x) = c(y)$.

Instead of the random walks $(x_i)_{i\geq 0}$ and $(y_i)_{i\geq 0}$, we use set-valued coalescing branching random walks. For $i \geq 0$, let X_i be the set of vertices accessible from x via a directed path in k-$\mathcal{N}^+$ of length i. Y_i is defined analogously.

Consider the stochastic process $(X_i)_{i\geq 0}$. It starts with $X_0 = \{x\}$. For each vertex v in $\mathbb{Z}^d$, there are two edges in k-$\mathcal{N}^+ \cap N_v$. Call the two vertices to which these two edges point the *children* of v. Note that the two children of v are chosen uniformly amongst size 2 subsets of $\{x + \mathbf{e}_i : 1 \leq i \leq d\}$, and this choice is independent for distinct vertices v_1, v_2. Then X_{i+1} is exactly the set containing the children of the vertices in X_i. We see that the process $(X_i)_{i\geq 0}$ may be viewed as a coalescing branching random walk, with branching factor of 2. It is branching because each point in X_i chooses 2 children to occupy the next state of the random walk. And it is coalescing because if two vertices in X_i both have the same child, there is only a single copy of that child in X_{i+1}.

We define $W_i = X_i - Y_i$, where the difference of sets means $A - B = \{a - b : a \in A, b \in B\}$. Strictly speaking, the distribution of $(W_i)_{i\geq 0}$ is not as simple as that of a coalescing branching random walk. This is because we can have $w = x_1 - y_1 = x_2 - y_2$ with $w \in W_i$ and distinct vertices $x_1 \neq x_2 \in X_i$, $y_1 \neq y_2 \in Y_i$. Then W_{i+1} contains points given both by the difference of the children of x_1 and y_1, and the points coming from x_2 and y_2. Moreover, because W_i is obtained by taking all pairwise differences of X_i and Y_i, the children of distinct vertices from W_i are not independent from each other.

However, for the proof that the coalescing branching random walk is recurrent, we only need to track one particle at a time, and observe its two potential children, choosing only one of them. If we track one particle from $(X_i)_{i\geq 0}$ and one particle from $(Y_i)_{i\geq 0}$, this gives at least two options for the children of w_i, and this is sufficient for the proof to go through.

Thus, by the proof of Lemma 2.8, this coalescing branching random walk almost surely occupies the origin eventually, showing that there is a directed path starting at x and one starting at y which intersect. We conclude that, every pair of vertices is in the same component, showing that k-$\mathcal{N}^+$, and thus k-$\mathcal{N}_D$, contains a unique infinite component almost surely. □

Lemma 2.8. *Let $(Z_i)_{i\geq 0}$ be a coalescing branching random walk with branching factor 2 on the d-dimensional integer lattice $\mathbb{Z}^d$, started at an arbitrary initial condition $Z_0 = \{v\}$. Then almost surely, $0 \in Z_i$ for infinitely many $i \geq 0$. The same is true for walks on the graph $L_0(\mathbb{Z}^d)$ whose vertices are $V = \{v \in \mathbb{Z}^d : c(v) = 0\}$ with edges between vertices which differ by 1 in exactly two coordinates.*

Proof. Let's start with the proof in the integer lattice case. The proof for $L_0(\mathbb{Z}^d)$ only requires a slight modification.

The idea is to follow the trajectory of just one particle. At each time step, it has two children. We disregard one and only follow the child with the better location. It suffices to show that this single trajectory hits the origin.

Concretely, we define a $\mathbb{Z}^d$-valued random walk $(z_i)_{i\geq 0}$ such that $z_i \in Z_i$ for every $i \geq 0$ and show that $(z_i)_{i\geq 0}$ hits the origin infinitely often. We bias one coordinate at a time of z_i towards 0. Once that coordinate hits 0, we switch to the next nonzero coordinate, cycling through the d coordinates (and if $z_i = 0$, we choose z_{i+1} without any bias). When moving from z_i to z_{i+1}, we have two increments to choose from which are uniformly sampled without replacement from $\{\pm\mathbf{e}_j : 1 \leq j \leq d\}$. While minimizing coordinate j, we choose the increment according to the following rule. Suppose the j-th coordinate is positive for simplicity, with the obvious modification if it is negative. If the increment $-\mathbf{e}_j$ is available, choose it. Otherwise, choose uniformly amongst the available increments which are not $\mathbf{e}_j$. Thus, with probability $1/d$, we have $z_{i+1} = z_i - \mathbf{e}_j$, and with probability $1 - 1/d$, the increment is uniform in $\{\pm\mathbf{e}_k : k \neq j\}$. This means that the number of steps of the process it takes for the coordinate being minimized to decrease by 1 follows a geometric distribution with mean d. And if the coordinate being minimized has value n, the number of steps it takes to hit 0 follows a negative binomial distribution with mean $|n|d$.

This defines a Markov chain with state space $\mathbb{Z}^d \times \{1, 2, \ldots, d\}$, where the last entry keeps track of which coordinate is being minimized. The accessible states are those for which the coordinate being minimized is nonzero. And 0 is special, as it doesn't have any coordinate to minimize.

Related to this, we define another process. Let a_i be the absolute value of the (i mod d)-th coordinate just before it gets minimized for the $\lceil i/d \rceil$-th time. In other words, we start the process by minimizing the first coordinate, and it has some value. This initial is a_1. When this finishes, we start minimizing the second coordinate. And a_2 is the absolute value of the second coordinate when it starts to get minimized. And a_{d+1} is the absolute value of the first coordinate when it starts to get minimized a second time, after all coordinates have been minimized. If a coordinate is 0 when it is its turn to get minimized, the corresponding a_i is 0.

Consider the evolution of this process. After the i-th coordinate finishes getting minimized, it has value 0. Then $d-1$ other coordinates are minimized, during which time the i-th coordinate experiences drift. After a coordinate has been minimized, the number of steps that it takes for the other coordinates to get minimized is a sum of $a_k + a_{k+1} + \cdots + a_{k+d-1}$ geometric distributions each with mean d. Thus, a_{k+d} has distribution of the absolute value of a random walk which has taken a random number of steps, where the expected number of steps is $d(a_k + a_{k+1} + \cdots + a_{k+d-1})$. The increments of the random walk are 0 with probability $1-1/d$ (it is uniform which coordinate changes) and ± 1 with probability $1/2d$ each. If X_n is a sum of n random variables with this distribution, then $\mathrm{Var}[X_n] = 1/d$ we can bound $\mathbb{E}|X_n| \leq \sqrt{\mathbb{E}X_n^2} = \sqrt{n/d}$. If there are N random steps, we can also bound

$$\mathbb{E}|X_N| = \mathbb{E}\,\mathbb{E}\left[|X_N| \,\middle|\, N\right] \leq \mathbb{E}\sqrt{N/d} \leq \sqrt{\mathbb{E}N/d},$$

where the last inequality is Jensen's.

Set $S_k = a_{k-d+1} + a_{k-d+2} + \cdots + a_k$. We note that if S_k is ever 0, this corresponds to z_i hitting the origin for some i. Since the expected number of steps is dS_k, we get the bound

$$\mathbb{E}\left[S_{k+1} \mid S_k\right] \leq \sqrt{S_k}.$$

We want to get a similar bound on $\mathbb{E}[S_{k+d}|S_k]$. Suppose that S_k is large. For induction, suppose that we have shown that $\mathbb{E}[a_{k+i}|S_k] \leq c_i\sqrt{S_k}$. We have that $S_{k+i} \leq S_k + a_{k+1} + \cdots + a_{k+i}$. So $\mathbb{E}[a_{k+i}|S_k]$

We show that $\mathbb{E}[S_{k+d}|S_k] \leq C\sqrt{S_k}$ for some finite constant C. First, we have shown $\mathbb{E}[a_{k+1}|S_k] \leq \sqrt{S_k}$. And $S_{k+1} \leq S_k + a_{k+1}$, so $\mathbb{E}[S_{k+1}|S_k] \leq \mathbb{E}[S_k + a_{k+1}|S_k] \leq S_k + \sqrt{S_k} \leq 2S_k$. Inductively, assume that $\mathbb{E}[a_{k+i}|S_k] \leq c_i\sqrt{S_k}$ and that $\mathbb{E}[S_{k+i}|S_k] \leq$

$b_i S_k$. Then

$$\mathbb{E}[a_{k+i+1}|S_k] = \mathbb{E}[\mathbb{E}[a_{k+i+1}|S_{k+i}]|S_k] \leq \mathbb{E}[\sqrt{S_{k+i}}|S_k] \leq \sqrt{\mathbb{E}[S_{k+i}|S_k]} \leq \sqrt{b_i S_k}.$$

And

$$\mathbb{E}[S_{k+i+1}|S_k] \leq \mathbb{E}[\mathbb{E}[S_{k+i+1}|S_{k+1}]|S_k] \leq \mathbb{E}[b_i S_{k+1}|S_k] \leq b_i b_1 S_k.$$

So the inequalities hold with $c_{i+1} = \sqrt{b_i}$ and $b_{i+1} = b_i b_1$. We finally bound

$$\mathbb{E}[S_{k+d}|S_k] = \mathbb{E}[a_{k+1} + \cdots + a_{k+d}|S_k] \leq (c_1 + \cdots + c_d)\sqrt{S_k}.$$

Thus, there is a finite constant C for which $\mathbb{E}[S_{k+d}|S_k] \leq C\sqrt{S_k}$. Set $\tilde{S}_k = S_{dk}$. Note that $(\tilde{S}_k)_{k\geq 1}$ is a Markov process. If $\tilde{S}_k \geq 4C^2$, then $\mathbb{E}[\tilde{S}_{k+1}|\tilde{S}_k]/\tilde{S}_k \leq C(\tilde{S}_k)^{-1/2} \leq 1/2$, so it decays exponentially in expectation until it reaches $4C^2$.

Let $i_0 = \left\lceil \log_2(\tilde{S}_k/4C^2) \right\rceil$. Then $\mathbb{E}[\tilde{S}_{k+i_0}|\tilde{S}_k] \leq 4C^2$. By Markov's inequality, $\mathbb{P}(\tilde{S}_{k+i_0} > 8C^2|\tilde{S}_k) \leq 1/2$. Because this is a Markov process, this tells us that almost surely $\tilde{S}_k$ is below $8C^2$ infinitely often. And when $\tilde{S}_k < 8C^2$, the probability that $\tilde{S}_{k+1} = 0$ is bounded away from 0. In particular, it is the probability that the next d steps in $(a_i)_{i\geq 1}$ all have value 0. Since each is generated from the sum of at most $\tilde{S}_k$ random variables, the probability that each is equal to 0 is a quantity bounded away from 0. This lets us conclude that $\tilde{S}_k$ hits 0 infinitely often, and thus so does the walk $(z_i)_{i\geq 1}$.

For the case where the random walk is on the lattice of points whose sum is 0, we use largely the same proof. We omit the computations and instead indicate the key change. Instead of minimizing all d coordinates, we disregard completely the final coordinate. This gives a process on the first $d-1$ coordinates. We make the observation that when the first $d-1$ coordinates are all 0, this forces the last one to be 0 as well. So it is sufficient to consider the process on the first $d-1$ coordinates.

The increments of the random walk on $L_0(\mathbb{Z}^d)$ are the vectors with one $+1$ entry and one -1 entry. If we greedily minimize one coordinate, this will result in the other coordinates getting pushed in the opposite direction, which ruins the analysis. The key insight is that if, while minimizing coordinate i, we only utilize increments which are -1 in coordinate i and $+1$ in coordinate d, then we can collect all of the bias into the d-th coordinate, which doesn't matter as we are ignoring it. More specifically, we always have 2 increments to choose from. While minimizing coordinate i, if either of these is the increment which is -1 in coordinate i and $+1$ in coordinate d, we choose that increment. Of course, the signs are reversed if coordinate i is negative and we are maximizing it. Otherwise, we choose uniformly among the two provided increments. This means that, while minimizing coordinate i, it experiences constant drift toward the origin, and the other coordinates, except for coordinate d, experience centered noise with no drift. Because the i-th coordinate can increase, the amount of time it takes to hit 0 no longer follows a negative binomial distribution, but the distribution of how long this takes still has expectation which is linear in the absolute value of the coordinate being minimized. And the other coordinates experience unbiased drift. This is all that is required for the above argument to go through. $\square$

2.3 Some bounds in the i.i.d. case

In this section, as in the previous, we assume that G is the d-dimensional infinite integer lattice $\mathbb{Z}^d$, with edges between vertices whose L^1 distance is exactly 1. We assume $k \leq 2d$. Since vertices in this graph have finite degree, every vertex of k-$\mathcal{N}_D$ has out-degree at least k. Here, we specifically assume that the edge weights are i.i.d. and a.s. distinct. Without loss of generality, we may assume they follow a Uniform$[0, 1]$ distribution.

Theorem 2.9. *Let $x \in \mathbb{Z}^d$. Set $N_x = \{\langle x, x+\mathbf{e}_i\rangle : 1 \leq i \leq d\}$, where $\mathbf{e}_i$ is the standard basis vector in the i-th direction. Assume the edge weights are independently and* *identically distributed Uniform*$[0,1]$ *random variables. For each nonnegative integer* k*,*

$$\mathbb{P}\left(\textit{no edge from } N_x \textit{ is present in } k\text{-}\mathcal{N}_D\right) \leq \left(\frac{1}{2}\right)^k \tag{2.3.1}$$

Proof. The case $k = 0$ is trivial, as probabilities are at most 1, and the case $k > d$ follows from the pigeonhole principle: x has $2d$ incident edges, and N_x contains d of those edges. If $k > d$, then one edge from k-$\mathcal{N}_D$ must be in N_x. So suppose that $1 \leq k \leq d$.

If $\vec{E}(x)$ denotes the set of directed edges starting at x, then let $C = \vec{E}(x) \setminus N_x$. The sets N_x and C each contain exactly d edges. If no edge from N_x is present in k-$\mathcal{N}_D$, then the k smallest edges are all in C. This has probability $\binom{d}{k}/\binom{2d}{k}$, since there are $\binom{d}{k}$ ways to choose the k edges in C and $\binom{2d}{k}$ total ways to choose the k edges. Then we see

$$\binom{d}{k}\Big/\binom{2d}{k} = \frac{d!k!(2d-k)!}{k!(d-k)!(2d)!} = \frac{d}{2d}\cdot\frac{d-1}{2d-1}\cdots\frac{d-k+1}{2d-k+1} \leq \left(\frac{1}{2}\right)^k,$$

which gives the result. $\square$

The previous result is a sort of correlation inequality. By symmetry, as the number of edges in N_x is the same as the number of edges incident to x which are not in N_x, the probability that the i-th smallest edge is in N_x is $1/2$. The event in question in Equation 2.3.1 is the event that the k smallest edges are all not in N_x. If the events that the i-th smallest edge is present in N_x were independent, the intersection would have probability 2^{-k}. Equation 2.3.1 says that things are less likely than if the events were independent, i.e., they are negatively correlated.

If we do the same computation but with a set of edges of size n instead of size $|N_x| = k$, the probability in question becomes

$$\binom{2d-n}{k}\Big/\binom{2d}{k} = \frac{(2d-n)!(2d-k)!}{(2d-n-k)!(2d)!} =\leq \left(\frac{2d-n}{2d}\right)^k.$$

Taking $n = d$ recovers the previous theorem.

Corollary 2.10. *With the same assumptions as Theorem 2.9, if $k \to \infty$ (and $d \geq k$ goes to ∞ as well),*

$$\mathbb{P}\,(\textit{at least one edge from } N_x \textit{ is present in } k\text{-}\mathcal{N}_D) \to 1.$$

For an edge e and a graph G, we use the notation $e \in G$ to mean that e is in the edge set of G.

Theorem 2.11. *Assume the edge weights are independently and identically distributed Uniform$[0,1]$ random variables. For distinct undirected edges $e_1, \dots, e_n$ of $\mathbb{Z}^n$,*

$$\mathbb{P}\left(\bigcap_{i=1}^{n} \{e_i \notin (d+1)\text{-}\mathcal{N}\}\right) \leq e^{-n/d}. \tag{2.3.2}$$

Proof. For a vertex $x \in \mathbb{Z}^d$, we define N_x as the set of edges $\{\langle x, x + \mathbf{e}_i \rangle : 1 \leq i \leq d\}$, where $\mathbf{e}_i$ is the standard basis vector in the i-th direction. Let $\tilde{N}_x$ be the version of N_x containing undirected edges instead of directed edges. For distinct vertices x and y, the sets $\tilde{N}_x$ and $\tilde{N}_y$ are disjoint, and moreover the sets $\tilde{N}_v$ for $v \in \mathbb{Z}^d$ partition the edges of the graph $\mathbb{Z}^d$.

Without loss of generality, we may assume that there are l distinct vertices $x_1, \dots, x_l$ and some integers $0 = n_0 < n_1 < \cdots < n_l = n$ such that $\{e_{n_{i-1}+1}, \dots, e_{n_i}\} \subseteq \tilde{N}_{x_i}$. The event

$$A_i := \bigcap_{j=n_{i-1}+1}^{n_i} \left\{\omega(e_j) > \min\{\omega(e) : e \in \tilde{N}_{x_i}\}\right\}$$

is implied by the event in Equation 2.3.2, since if $\omega(e_j)$ is the minimum weight in $\tilde{N}_{x_i}$, then e_i is one of the the $d+1$ lowest weight edges incident with x_i. So it suffices to show $\mathbb{P}(\cap_{i=1}^{l} A_i) \leq e^{-n/d}$. Moreover, the events A_i for $i = 1, \dots, l$ are independent, as they are defined in terms of disjoint sets of independent random variables. The event A_i is equivalent to the event that one of the $d-(n_i-n_{i-1})$ edges in $\tilde{N}_{x_i} \setminus \{e_{n_{i-1}+1}, \ldots, e_{n_i}\}$ is the minimum weight edge in $\tilde{N}_{x_i}$, which has probability $1-(n_i-n_{i-1})/d \leq e^{-(n_i-n_{i-1})/d}$. By independence, using the fact that $\sum_{i=1}^{n}(n_i - n_{i-1}) = n$, we get that $\mathbb{P}(\cap_{i=1}^{l} A_i) \leq e^{-n/d}$, which proves the result. $\square$

We may generalize this to the case $k \geq d+1$ in the following way.

Theorem 2.12. *Assume the edge weights are independently and identically distributed Uniform*$[0,1]$ *random variables. For $d+1 \leq k \leq 2d$ and distinct undirected edges $e_1, \ldots, e_n$,*

$$\mathbb{P}\left(\bigcap_{i=1}^{n}\{e_i \notin k\text{-}\mathcal{N}\}\right) \leq \left(2-\frac{k}{d}\right)^{\frac{n}{2d-k}} \tag{2.3.3}$$

Proof. We use the definition of $\tilde{N}_x$ from the proof of Theorem 2.11, as well as the assumption, with no loss of generality, that there are l distinct vertices $x_1, \ldots, x_l$ and some integers $0 = n_0 < n_1 < \cdots < n_l = n$ such that $\{e_{n_{i-1}+1}, \ldots, e_{n_i}\} \subseteq \tilde{N}_{x_i}$. For $1 \leq i \leq l$, we define the event

$$A_i := \bigcap_{j=n_{i-1}+1}^{n_i} \left\{\omega(e_j) > (k-d)^{th}\text{-min of } \{\omega(e) : e \in \tilde{N}_{x_i}\}\right\},$$

where the $(k-d)^{th}$ minimum of the set is almost surely well defined because the set contains k almost surely distinct elements. Each event A_i is implied by the event in Equation 2.3.3, so it suffices to show $\mathbb{P}(\cap_{i=1}^{l} A_i) \leq (2-k/d)^{n/(2d-k)}$.

The event A_i is equivalent to the event that the weights of some $k-d$ edges in the set $\tilde{N}_{x_i} \setminus \{e_{n_{i-1}+1}, \ldots, e_{n_i}\}$, which contains $d-(n_i-n_{i-1})$ edges, are the $k-d$ minimum weights among edges in $\tilde{N}_{x_i}$. This event has probability $\binom{d-(n_i-n_{i-1})}{k-d}/\binom{d}{k-d}$, since there are $\binom{d}{k-d}$ ways to choose which $k-d$ elements are the minimum $k-d$ and, out of those, $\binom{d-(n_i-n_{i-1})}{k-d}$ gives the number of ways in which none of the $n_i - n_{i-1}$ elements of $\{e_{n_{i-1}+1}, \ldots, e_{n_i}\}$ are in the minimum $k-d$. This probability simplifies to

$$\begin{aligned}\frac{(d-(k-d))!\,(d-(n_i-n_{i-1}))!}{(d-(n_i-n_{i-1})-(k-d))!\,d!} &= \frac{2d-k}{d}\;\frac{2d-k-1}{d-1}\cdots\frac{2d-k-(n_i-n_{i-1})+1}{d-(n_i-n_{i-1})+1}\\ &\leq \left(2-\frac{k}{d}\right)^{n_i-n_{i-1}},\end{aligned}$$

since each term in the product is at most $(2d-k)/d$. The events A_i are defined in terms of disjoint sets of random variables, they are independent, which lets us we conclude that $\mathbb{P}(\cap_{i=1}^{l} A_i) \leq (2-k/d)^n$. $\square$

CHAPTER 3
PHYSICS INFORMED NEURAL NETWORKS

The contents of this chapter include joint work with Biraj Dahal, Wenjing Liao, and Molei Tao.

3.1 Overview

The study of numerical methods for solving partial differential equations (PDEs) is in many ways mature. Nonetheless, common numerical methods, which discretize on a grid, have an exponential dependence on the dimension of the domain. While many common PDEs have four-dimensional spacetime as their natural domain, others are in higher dimensions. For example, many physical PDEs such as the Boltzmann, Vlasov, and Fokker-Planck equations, have position, velocity, and time as their phase space, leading to seven-dimensional systems. The Black-Scholes equation on n assets is a $n+1$-dimensional PDE. The Hamilton-Jacobi-Bellman equation is a high-dimensional PDE which arises in control problems.

Some of these are at times intractable. For example, finding the ground state energy of a quantum system is in general NP-hard (or QMA-hard). Nonetheless, there has been success applying techniques from modern machine learning to solving PDEs. One approach is with physics informed neural networks (PINNs) [26, 7].

The idea behind PINN relies on the fact that neural networks are universal function approximators, and the fact that all derivatives of a neural network can be effectively computed. The degree to which the derivatives of the neural network do not satisfy a PDE can be used as a loss function, which can be minimized to obtain an approximate solution to the PDE. A catch is that, while all derivatives of the neural network can be computed, one must specify a point in the domain at which

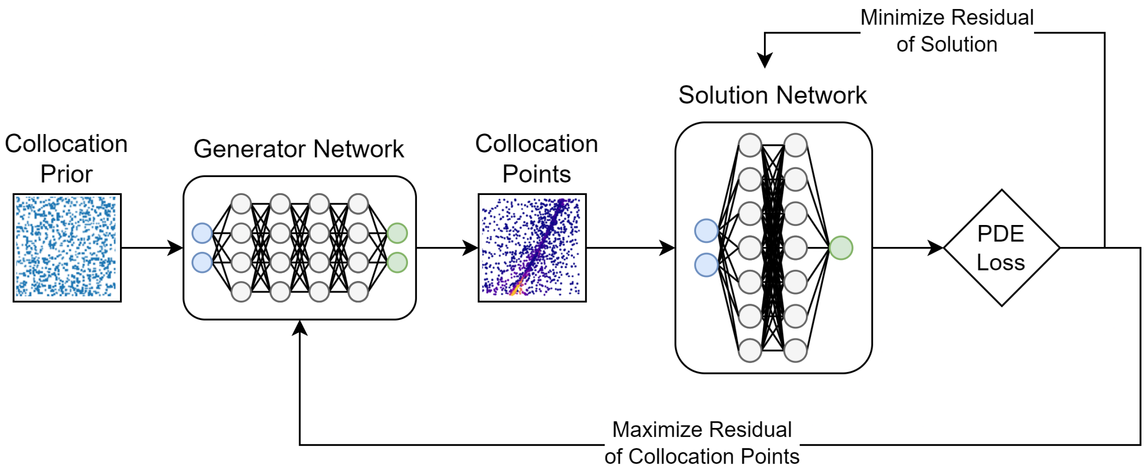

Figure 3.1: Diagram representing the training scheme of the proposed method. The generator network tries to maximize the PDE loss by generating high-loss collocation points, while the solution network tries to minimize it by improving the PDE loss at the collocation points.

to evaluate the derivative. The choice of which points to evaluate the derivatives at has significant influence on the performance of the trained network [6, 36].

In this chapter, we introduce a model which generates points for the neural network's derivatives to be evaluated at, which we call **collocation points**, and show that this can significantly accelerate the training of the neural network and can improve the accuracy of the converged neural network. We achieve this with a low computational overhead cost associated with running the generative model. Figure 3.1 shows the training scheme.

3.2 Background

3.2.1 Neural networks

Feedforward neural networks are parametric functions from $\mathbb{R}^n$ to $\mathbb{R}^m$. A feedforward neural network architecture is determined by n, m, the sizes of hidden layers, and choice of activation function. Fix $L, p, m, n \in \mathbb{N}$ and $\sigma : \mathbb{R} \to \mathbb{R}$. Then a feedforward neural network with L hidden layers and width p from $\mathbb{R}^m$ to $\mathbb{R}^n$ is given by

$$f_\theta(x) = (A_{L+1} \circ \sigma \circ A_L \circ \sigma \circ \cdots \circ \sigma \circ A_1),$$

where $A_1 : \mathbb{R}^n \to \mathbb{R}^p$, $A_{L+1} : \mathbb{R}^p \to \mathbb{R}^m$, and $A_i : \mathbb{R}^p \to \mathbb{R}^p$ for $1 < i < L+1$ are affine linear transformations, and θ is the vector containing all the numbers which specify these affine linear transformations. The activation function σ is applied element-wise. The hidden layers in principle can vary in size and the activation functions between layers can vary, but we restrict ourselves to the class of feedforward networks described above.

Since the neural network is the composition of affine linear transformations and an activation function, it inherits the smoothness properties of the activation function. A common choice is hyperbolic tangent, which has strong smoothness properties; another is ReLU, which does not. As a parametric function from $\mathbb{R}^n$ to $\mathbb{R}^m$, it makes sense to talk about derivatives of arbitrary order of f_θ. Moreover, there are libraries such as PyTorch which allow the numerical values of arbitrary derivatives to be computed at any point in $\mathbb{R}^n$. More generally, tools such as PyTorch are written in such a way that any appropriately coded function can be differentiated using a process known as automatic differentiation. As a corollary, any function written appropriately in such a framework can be directly optimized using a gradient-based optimizer such as gradient descent or Adam. How successful the optimization is depends on the choice of optimizer, hyperparameters, and the function being minimized.

3.2.2 PDEs and loss functions

Many PDEs with boundary conditions can be specified by $\Omega, \mathcal{D}$, and $\mathcal{B}$, where $\Omega \subseteq \mathbb{R}^d$ is the domain, which we assume to be a closed set, $\partial\Omega$ is its boundary, and $\mathcal{D} \in L^2(\Omega; \mathbb{R}^k)$ and $\mathcal{B} \in L^2(\partial\Omega; \mathbb{R}^k)$ are (partial) differential operators. A strong solution to the PDE with boundary conditions is a function $u^* : \Omega \to \mathbb{R}^k$ which satisfies

$$(\mathcal{D}u^*)(x) = f \text{ for } x \in \Omega,$$
$$(\mathcal{B}u^*)(x) = g \text{ for } x \in \partial\Omega,$$

for some choices of f and g. If $k > 1$, this can be thought of as a system of PDEs. Without loss of generality, we assume that f and g are identically 0.

If $u_\theta : \Omega \to \mathbb{R}^k$ is a parametric function, then we we would like for $\mathcal{D}u_\theta \approx 0$ on Ω and $\mathcal{B}u_\theta \approx 0$ on $\partial\Omega$. If the boundary value problem is well-posed, then the solution depends continuously (under some appropriate norms) on f and g. Hence, assuming f and g are 0, if $\mathcal{D}u_\theta$ and $\mathcal{B}u_\theta$ are small under appropriate norms, then $u^* - u_\theta$ is also small under an appropriate norm. This leads to the formulation of so-called physics-informed neural networks: let u_θ be a neural network and use gradient-based optimization algorithms to find a value of θ which makes $\|\mathcal{D}u_\theta\|$ and $\|\mathcal{B}u_\theta\|$ small. We call the function $\mathcal{D}u_\theta$ the PDE residual.

In practice, it is often possible to choose the architecture of the neural network in such a way that forces $\mathcal{B}u_\theta \equiv 0$; we refer to this as hard boundary enforcement. To illustrate the principle, if $u_\theta : \mathbb{R} \to \mathbb{R}$ and we want to enforce $u_\theta(0) = u_\theta(1) = 0$, we can choose $u_\theta(x) = (x)(1 - x)f_\theta(x)$, where $f_\theta : \mathbb{R} \to \mathbb{R}$ is a feedforward neural network. This works for Dirichlet boundary conditions [19], and there are techniques to handle other boundary conditions [30].

The other option is soft boundary enforcement, in which $\mathcal{B}u_\theta$ is treated in a similar manner as $\mathcal{D}u_\theta$. For simplicity, we use hard boundary enforcement in this chapter.

With $\mathcal{D}$ and Ω specified, and with hard boundary enforcement ensuring that u_θ satisfies the boundary condition, the goal is to minimize an appropriate norm of Du_θ. While the theoretically appropriate norm to ensure well-posedness varies problem-to-problem, and although Sobolev norms often come with the strongest theoretical guarantees, it is common to use a L^2 norm when optimizing PINNs. Computing the higher derivatives of Du_θ required for Sobolev norms comes with added complexity and computational cost, and directly minimizing the value of Du_θ works well in practice. To this end, we define the squared L^2 loss function

$$L(u) = \|\mathcal{D}u\|_{L^2}^2 = \int_\Omega |(\mathcal{D}u)(x)|^2 \, dx.$$

This is an integral which must be numerically approximated. In standard PINN approaches, for $X = (x_1, \ldots, x_n) \in \Omega^n$, the loss is approximated as

$$\|\mathcal{D}u\|_X^2 := \sum_{i=1}^n |(\mathcal{D}u)(x_i)|^2.$$

The choice of points X can vary over the course of optimization. For example, X may be a fixed grid of points on Ω, it may be a fixed set of random points, or it may be randomly resampled for every optimization step. These points are known as collocation points.

3.2.3 Literature review

Minimizing a PDE residual loss of a neural network evaluated at collocation points has been done early as the 1990s [7, 19], but this approach has seen more research in recent years [26, 27, 37, 16].

There has been a spate of methods selecting collocation points in a way that is influenced by the PDE residual. Several methods are compared in Wu et. al. [36], three of which are residual-based adaptive sampling methods. The other methods are non-adaptive strategies such as uniform random points, a fixed grid of points, or a Hammersley sequence of points. The residual-based adaptive sampling methods typically sample from a larger number of points using some non-adaptive strategy, evaluate the PDE residual at these points, and somehow keep or emphasize the points which have high PDE residual during training. In low dimensions, it is feasible to sample the domain densely and get good coverage of the PDE residual. In higher dimensions, it becomes infeasible to sample densely.

One adaptive method, referred to as residual-based adaptive refinement with greed (RAR-G) [36] or just, in the paper originally proposing it, residual-based adaptive refinement [24], samples a dense set of points, evaluates the PDE residual at these points, and then adds the m points with greatest residual to the set of points used for PINN training. These points are used for some number of standard PINN training steps, until another round of dense sampling adds m more points to the set of training points, and the process continues. "Refinement" refers to the fact that each time, m more points are added, and none of the previously used training points are discarded.

A second method [28] uses the idea of importance sampling to select collocation points following a distribution whose density is proportional to loss or PDE residual. This is generalized by a method called residual-based adaptive distribution (RAD) [36] in which the density sampled is proportional to a power of the loss, plus a constant, with the exponent and the constant serving as hyperparameters:

$$p(x) \propto \frac{|(\mathcal{D}u)(x)|^\alpha}{\mathbb{E}\left[\,|(\mathcal{D}u)(Z)|^\alpha\right]} + c. \tag{3.2.1}$$

The random variable Z is distributed according to the probability measure on Ω. The PDE residual $\mathcal{D}u$ can be generalized to another loss, for example using the Sobolev loss.

Another method, residual-based adaptive refinement with distribution (RAR-D) [36], blends the previous two approaches. While RAR-G retains the m sampled points with the highest PDE residual, in RAR-D, m points are sampled according to the distribution in Equation 3.2.1. As in RAR-G, these m points are added to the training set, so that each iteration, the set of collation points grows by m. In other words, RAR-D acts as RAD but with refinement instead of pure resampling.

The preceding methods all rely on sampling the domain and evaluating the PDE residual at many points to identify a much smaller set of points to train the solution network at. However, this comes with a computational opportunity cost, as every point where the loss function is evaluated is a point where the gradient of the loss function could be computed. For (scalar-valued) loss functions, the computational cost of computing the gradient of the loss function is at most a small constant, typically 3 to 4, times the cost of evaluating the loss function ([12], Chapter 3.3). Hence, while evaluating the loss at many points can identify good points to train at, this incurs a computational opportunity cost, since one could directly train at more points instead. This suggests that lighter weight generative models for producing good collocation points would be valuable.

One method using a generative model to generate PINN collocation points is deep adaptive sampling (DAS-PINNs) [31]. Inspired by importance sampling, the generative model aims to sample from a distribution whose density is proportional to the PDE residual or other loss function. The authors use a bespoke architecture which they call a KR-Net that approximates the density, so that the distribution may be (approximately) sampled from. The aim of this sampling is to reduce the variance of the estimator of the true loss integral $\int_\Omega |(\mathcal{D}u)(x)|^2 \, dx$.

The above methods are defined entirely in terms of the loss evaluated at points in the domain, and none make use of the gradient of the loss to improve the generated points. We describe two methods which do.

In one [35], which we refer to as L^∞-PINN, the use of L^2 loss for PINN is called into question. For each training step, some collocation points are randomly sampled from a base distribution. The points are then individually optimized to maximize the PDE residual using 20 steps (a tuneable hyperparameter) of a gradient-ascent style algorithm. However, it is not a traditional gradient-ascent. Instead of using the gradient $\nabla_x(\mathcal{D}u)(x)$ of the PDE residual with respect to each collocation point, the method uses the sign of this gradient: $\text{sign}(\nabla_x(\mathcal{D}u)(x))$, causing every entry to be ± 1. After the collocation points have been optimized, the solution network is trained for one step on these points. This is repeated with newly sampled and optimized collocation points each iteration. The main drawback is that $20+1$ backward passes through the solution network are required for every update of the solution network.

Another such method, titled point adaptive collocation method for artificial neural networks (PACMANN) [34], is similar. In it, the points are not reset every time and instead receive gradient updates every training step. They explore various gradient-based optimization algorithms for moving the collocation points. Adam is a strong performer.

3.3 Method

In the proposed method, the collocation points to evaluate the PDE residual loss at are generated by a generative model, which is a neural network that is trained to generate points which maximize the PDE residual loss. We fix a latent space Q with probability measure ρ from which we can easily sample. For nice geometries, it is convenient to take $Q = \Omega$ with the uniform measure. A generator network is a function $g_\phi : Q \to \Omega$, which we take to be a neural network parameterized by ϕ. This induces a probability measure on Ω: the pushforward of ρ under g_ϕ. If $Z \sim \rho$, then $g_\phi(Z)$ follows the distribution of the pushforward, so we can easily sample collocation points from it by sampling z_i from ρ and setting $x_i = g_\phi(z_i)$. The training objective for the generator network is to maximize the PDE residual loss at the generated collocation points. That is, for a fixed function u, we want solve the maximization problem

$$\max_{\phi} \int_Q \left|(\mathcal{D}u)(g_\phi(z))\right|^2 \, d\rho(z).$$

Again, we cannot evaluate this integral directly, so we estimate it by sampling points $z_1, \ldots, z_n \sim \rho$ and approximating with $||Du||_X$ for $X = (g_\phi(z_1), \ldots, g_\phi(z_n))$.

The method is to use this generator network to generate higher loss points for the solution network to train at. The networks are trained in an adversarial manner. See Figure 3.1.

3.3.1 Solution and generator networks

For the solution network, we use a modified feedforward network with tanh activation function and two hidden layers. Such a network is capable of approximating any function under Sobolev norm arbitrarily well if the network is sufficiently wide [5]. So let f_θ be a tanh feedforward neural network. For hard boundary enforcement, with dirichlet boundary conditions of the form $u(x) = b(x)$ for $x \in \partial\Omega$, we need a continuous function $N_{\partial\Omega}$ defined on Ω which is 0 on the boundary and nonzero elsewhere. Then

$$u_\theta(x) = N_{\partial\Omega}(x) f_\theta(x) + b(x)$$

satisfies the boundary conditions. For rectangular domains $\Omega = \prod_{i=1}^d [a_i, b_i]$, a convenient choice of enforcement function is

$$N_{\partial\Omega}(x) = \prod_{i=1}^d \frac{2(x - a_i)(x - b_i)}{b_i - a_i}$$

For the generator network, we similarly use a modified feedforward network with ReLU activation function. ReLU is acceptable here as there is no need for the generated distribution to have strong smoothness properties. However, another constraint arises with the generator network. In order for the generated points to cover the entire domain, we need the support of the pushforward distribution to be all of Ω. In terms of the function g_ϕ, this corresponds to the function being surjective. This is achieved by enforcing that g_ϕ is equal to the identity function on $\partial\Omega$. Additional constraints can be imposed as appropriate for the problem. For example, in spacetime problems, one may have a desire to weight all times equally.

Lemma 3.1 (Generator Approximation)**.** *Let Ω be bounded and convex and ρ be the uniform measure on Ω. Let $\mu \in \mathbb{P}(\Omega)$ be a probability measure with Hölder class density $f_\mu \in \mathcal{C}^\alpha(\Omega)$ with respect to ρ for some $\alpha \in (0,1)$. Then there is a surjectivity enforced neural network g_ϕ such that*

$$\mathbb{W}_1(\mu, (g_\phi)_\sharp \rho) < \epsilon.$$

Proof. See Appendix B for the full proof, including the extension to the case where $\Omega = \Omega_{\text{space}} \times [0, T]$ is a spacetime domain, and a time-preserving generator neural network is used. $\square$

Lemma 3.2. *Suppose $\Omega \subseteq \mathbb{R}^d$ is homeomorphic to the standard d-ball. Let $f : \overline{\Omega} \to \mathbb{R}^d$ be continuous, and let $N \in \mathcal{C}_0\left(\overline{\Omega}\right)$ be any continuous function that vanishes on the boundary $\partial\Omega$. Then the function $\tilde{f}(x) = N(x)f(x) + x$ defined on $\overline{\Omega}$ contains Ω in its range.*

Proof. Without loss of generality suppose Ω is the open ball in d-dimensions, so $\partial\Omega = S^{d-1}$ and $\overline{\Omega} = D^d$. For all $x \in \partial\Omega$, we have $\tilde{f}(x) = x$, so $\tilde{f}|_{\partial\Omega}$ is the identity. Note that this means $\partial\Omega \subseteq \tilde{f}(\overline{\Omega})$. In addition, we have the inclusion map $\iota : \partial\Omega \to \overline{\Omega}$.

Now suppose for the sake of contradiction that $y \in \overline{\Omega}$ such that $y \notin \tilde{f}(\overline{\Omega})$. Since $\partial\Omega \subseteq \tilde{f}(\overline{\Omega})$, this means $y \in \Omega$. This allows us to define a map $\phi : \overline{\Omega} \to \partial\Omega$ as follows: given any x, consider the unique ray starting at y passing through $\tilde{f}(x)$, and $\phi(x)$ is the intersection of that ray with $\partial\Omega$. Note that $\phi|_{\partial\Omega}$ is the identity. Thus the composite map $\phi \circ \iota : \partial\Omega \to \partial\Omega$ is the identity map. Thus the induced map on the homology group

$$(\phi \circ \iota)_* : H_{d-1}(\partial\Omega) \to H_{d-1}(\partial\Omega)$$

is the identity homomorphism. However, we must have that $(\phi \circ \iota)_* = \phi_* \iota_*$, but this is impossible because the function ι_* is the zero function:

$$\iota_* : H_{d-1}(\partial\Omega) \to H_{d-1}(\overline{\Omega}), \quad \text{where } H_{d-1}(\overline{\Omega}) = 0,$$

but the homology group $H_{d-1}(\partial\Omega)$ is not 0. □

3.3.2 Training

Once the loss functions and generator and solution network architectures have been fixed, the optimization algorithm has to be chosen. The basic framework is to use gradient-based optimization algorithms to minimize the PDE loss with respect to θ and maximize the PDE loss with respect to ϕ. We use Adam as the optimizer for both and alternate between some number s_u of θ updates and some number of s_g of ϕ updates. The basic algorithm is described in Algorithm 1. It requires a number of training cycles N, the distribution ρ on latent space, the number of steps s_u and s_g for training u and g within each training cycle, the number of sample points n, and optimizers for u and for g. The optimizers typically have additional hyperparameters which must be specified.

Algorithm 1: GG-PDE

```
Data: N, ρ, s_u, s_g, n, Opt_u, Opt_g
Result: Trained u_θ
for i = 1, ..., N do
    for j = 1, ..., s_u do
        Sample z_1, ..., z_n ~ ρ
        x_k ← g_φ(z_k)
        L_u ← (1/n) Σ_{k=1}^{n} |D u_θ(x_k)|^2
        θ ← Opt_u(θ, ∇_θ L_u)
    for j = 1, ..., s_g do
        Sample z_1, ..., z_n ~ ρ
        x_k ← g_φ(z_k)
        L_g ← -(1/n) Σ_{k=1}^{n} |D u_θ(x_k)|^2
        φ ← Opt_g(φ, ∇_φ L_g)
```

Additional penalties may be added for either model as desired, for example, any points g_ϕ generates which are not in Ω may be discarded and g_ϕ may receive a penalty for generating these points. Ideally, such a penalty function should be differentiable with the (negative) gradient pointing towards Ω, so that g_ϕ can learn to generate points in the domain.

3.4 Numerical experiments

Given a proposed approximate solution u_θ to the PDE, there are two primary ways to measure the error. If an exact or high-fidelity solution u^* is available (for example, from a traditional numerical solver), we can measure the relative error

$$\frac{\|u_\theta - u^*\|}{\|u^*\|}.$$

This integral can be approximated with a fixed grid on bounded domains, or approximated by sampling random points in the domain. The relative error is not observable unless the ground truth is already known.

The error observable during training is the PDE residual

$$\|\mathcal{D}u_\theta\|,$$

which can be approximated with points from a grid, random points, or with the points generated by g_ϕ. Numerical experiments were run on an NVIDIA RTX 6000, and using the library PyTorch.

3.4.1 Low-dimensional

As a simple example, we use the transport equation in $d+1$ dimensions on the spacial domain $[0, L]^d$ with time between 0 and 1. Let $v \in \mathbb{R}^d$ be the speed. Boundary conditions should be specified on the parts of the boundary which the characteristic

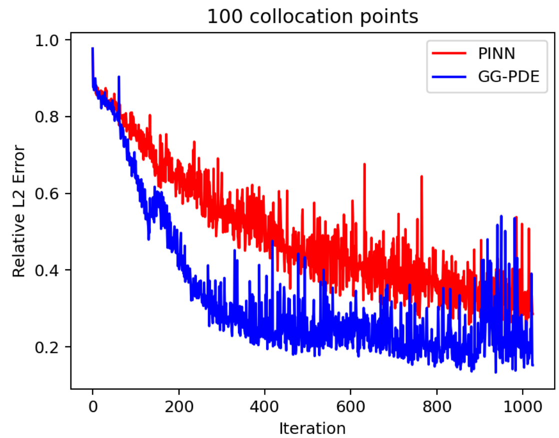

Figure 3.2: Error versus training iteration of GG-PDE and PINN for the (2+1)d transport equation with 100 collocation points

curves do not hit. For example, if all components of v are positive, then the boundary conditions should be specified on the parts of the boundary with (at least) one coordinate equal to 0. Call this set B. The transport equation with initial condition f_0 and zero Dirichlet boundary conditions is given by

$$\begin{aligned} u_t + v \cdot \nabla u = 0, \\ u(x,0) = f_0(x), x \in (0,L)^d, \\ u(x,t) = 0, (x,t) \in B \end{aligned}$$

As initial conditions, we use a bump of width $h = 0.3$ and the speed vector v which has all entries equal to 1. We set $L = 1.3$. The initial condition is a shift of

$$\tilde{f}_0(x) = \prod_{i=1}^{d} \frac{16}{h^4}(x_i(1-x_i))^2 \mathbf{1}_{[0,h]}(x_i),$$

which has continuous first derivatives. Then set $f_0(x) = \tilde{f}_0(x - c)$, where c is a vector with all entries 0.1, which serves as padding away from the boundary conditions. The exact solution is $u(x,t) = f_0(x - vt)$.

The adversarial training gives both an improvement in final accuracy, as shown

Table 3.1: Comparison of final relative L^2 test error of GG-PDE and PINN on the transport equation with varying number of collocation points (mean with standard deviation in parenthesis over three trials).

	50 points	100 points
GG-PDE	0.186 (0.030)	0.182 (0.011)
PINN	0.244 (0.017)	0.221 (0.046)

in Table 3.1, and in convergence speed, as shown in Figure 3.2.

3.4.2 High-dimensional

In low dimensions it is feasible, although not necessarily efficient, to use a dense set of collocation points. In contrast, in high dimensions, a dense set of points is exponentially large and cannot be used. It is difficult to know a priori which regions of the domain are the most relevant, and it is infeasible in high dimensions to sample the domain densely to discover the important regions. The adversarial training in GG-PDE leverages the gradient of the PDE residual to find important regions of the domain. Moreover, so long as the generator network is not too large compared to the solution network, the computational overhead of using adversarial training is fairly low, resulting in a significant increase in training efficiency compared to standard PINN.

To illustrate this, we consider a version of the Linear Quadratic Gaussian control problem, which is a special case of the Hamilton-Jacobi-Bellman (HJB) equation:

$$\begin{aligned} u_t + \Delta u - \mu \|\nabla_x u\|^2 = 0, \\ u(0,1) = g(x), \quad x \in \mathbb{R}^n. \end{aligned} \tag{3.4.1}$$

This problem is known to have a unique solution of

$$u(x,t) = -\frac{1}{\mu} \ln \left(\mathop{\mathbb{E}}_{y \sim \mathbb{N}} \left[\exp(-\mu\, g(x - \sqrt{2(1-t)}y) \right] \right), \tag{3.4.2}$$

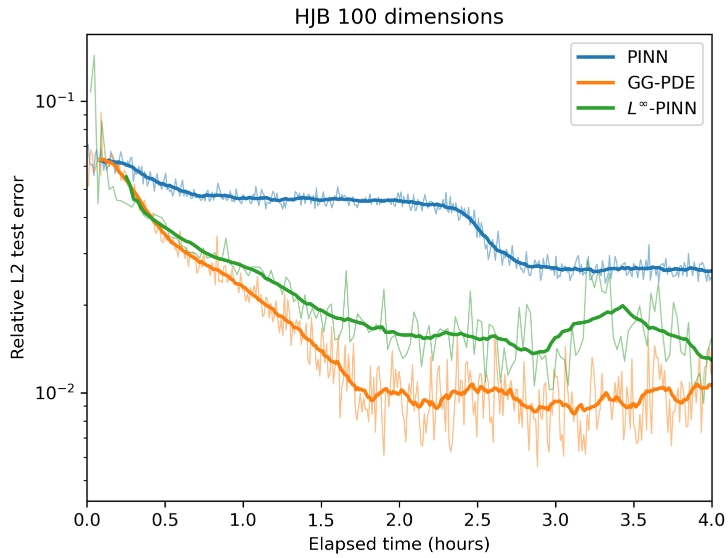

Figure 3.3: Relative L^2 test error (log scale) versus training time for the HJB equation in 100 dimensions with terminal condition $g(x) = \ln\left(\frac{1+\|x\|^2}{2}\right)$. Thick lines are a rolling average.

where the expectation is taken over y following the standard normal distribution $\mathbb{N}$ in $\mathbb{R}^n$. We consider Equation 3.4.1 with $\mu = 1$, $n = 100$, and terminal condition $g(x) = \ln\left(\frac{1+\|x\|^2}{2}\right)$. This problem has been considered in other works [15, 35].

With 100 collocation points per iteration, only a very small portion of the 100 dimensional space can be explored, and the choice of points can greatly influence training. In Figure 3.3, the relative L^2 test error is plotted against elapsed training time to demonstrate the efficiency and accuracy of GG-PDE compared to standard PINN.

Empirically, we find that GG-PDE minimizes the test error significantly more quickly than standard PINN, and the converged model stabilizes at a better error. For comparison, the method described in [35], which we refer to as L^∞-PINN, is also plotted. It is important to note that the L^∞-PINN method eventually reaches a lower test error than GG-PDE, but takes significantly longer (12 hours vs about 2 hours) to converge.

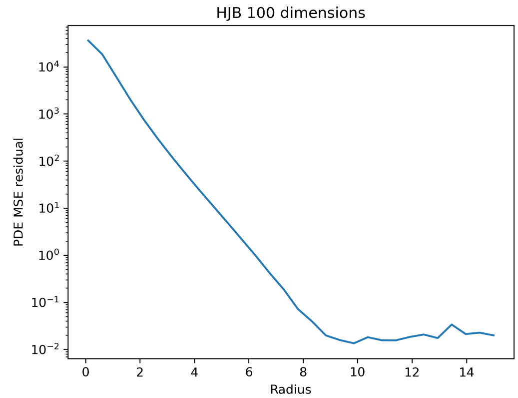

Figure 3.4: Error vs radius after 5000 training iterations of standard PINN.

The generator network adds relatively little training overhead compared to standard PINN. The training points for each iteration in L^∞-PINN are obtained by running 20 steps of gradient ascent with respect to collocation points on the PDE residual, requiring 21 forward and backward passes through the neural network per iteration. This gradient ascent starts from new random points each training iteration. In contrast, the generator of GG-PDE evolves during training, allowing it to retain information from prior training iterations.

Figure 3.4 shows that with standard PINN, the PDE residual is concentrated in lower-norm locations after training. The norm of the random points are about $\sqrt{100} = 10$ on average, and with high probability the lower-norm regions are never explored when using standard PINN. GG-PDE is able to generate points with smaller norm.

The specific settings of the experiment are as follows. As in [35], the training points for standard PINN, and the latent distribution for GG-PDE, are standard Guassian in space and uniform on $[0, 1]$ in time. Hyperparameters for L^∞-PINN are identical to those in [35], except that hard boundary enforcement is used. For all 3 methods, the solution network uses hard boundary enforcement, 100 training points per iteration, 3 hidden layers, width 4096, and tanh activation function. The learning rate is 10^{-4}

for GG-PDE and PINN. The generator neural network has 3 hidden layers, width 200, ReLU activation function, and learning rate 10^{-3}. It is time-preserving, in the sense that it acts as the identity on the time dimension, has the identity function added to the output, and receives a penalty of 0.0001 times the squared norm of the generated points. In GG-PDE, solution network is “pre-trained” for 500 steps on the same distribution as standard PINN, then trained 5 solution steps for every 1 generator step, with 50 points from the generator and 50 random. Following [35], the test points are uniformly distributed on the cube $[0, 1]^{n+1}$. Experiments are run on an NVIDIA RTX 6000. A Monte Carlo simulation over 200,000 samples is applied to Equation 3.4.2 to obtain a ground-truth solution.

3.5 Conclusion

Using a generative model to select collocation points has the potential to accelerate PINN training and improve accuracy of the converged model. We proposed and studied GG-PDE, a zero-sum game setup using adversarial training. The generator uses PDE residual gradient information to find and emphasize high residual regions.

We now discuss potential limitations. This method starts with the premise that it is advantageous to train on higher residual points. Although it is reasonable to expect that training on these points is likely to reduce overall error, this may not hold in all circumstances.

Training the generative model in GG-PDE implicitly uses the gradient of the PDE residual with respect to the collocation points to find high residual regions of the domain. As this optimization problem is non-convex, the gradient information does not necessarily point toward the high residual regions.

When the generative model finds high residual regions, training the solution network reduces the residual on these regions. After this, there may be too little gradient information left to help the generator leave this region, potentially leaving the model in a sub-optimal local minimum. This can be mitigated somewhat by enforcing surjectivity of the generator onto the domain, ensuring that all regions of the domain may be discovered. This behavior suggests that there should be a balance between the learning speeds of the generator and the solution so that there is always useful gradient information for both models.

While the motivation and purpose are different, the training algorithm of GG-PDE is similar to that of Generative Adversarial Networks (GANs). Training of GANs is known to be unstable and can exhibit mode collapse [32, 18] and can be sensitive to hyperparameter choice [25]. Mode collapse in GG-PDE can be mitigated by architecture choices such as surjectivity enforcement or by adding the identity function to output of the generator.

Point selection necessarily comes with a trade-off: by emphasizing one region of the domain, the other regions are necessarily de-emphasized. The generator may de-emphasize regions of the domain which are have low residual yet are nonetheless important for training. Training on a combination of generated points and uniformly distributed points, and/or enforcing surjectivity, can help ensure that no regions of the domain are completely ignored.

Appendices

APPENDIX A
FBM EXTRAS

A.1 Discussion of future directions

Hammond and Sheffield [14] pose an open problem of whether some negative correlation structure can give negatively correlated fractional Brownian motion, which has Hurst parameter $H \in (0, 1/2)$ and variance that scales like n^{2H}. A basic candidate model would be one in which, instead of Y_v assuming the value of its ancestor, each Y_v assumes the opposite value of its ancestor. For two vertices u and v in the same connected component, the variables Y_u and Y_v are the same if and only if the path between u and v has even length, otherwise $Y_u = -Y_v$. Under this model, the covariance between a pair Y_v and Y_u depends not only on the probability that u and v are in the same component, but also on the probability that the path between them has even length when they are in the same component. In the Hammond-Sheffield model, one might heuristically guess that for far apart vertices, the path lengths have asymptotically 1/2 chance of being even. The positive and negative correlations would cancel, leaving only second-order effects. One might guess that the second-order effects decay too quickly to give a fractional Brownian motion, but it is not clear.

For our model, the parity of the length of the path (if such a path exists) is deterministic, and is equal to the parity of $(u - v) \cdot e_1$. Furthermore, we see that the colors of vertices in a given connected component alternate in sign as we move in the e_1 direction. In this case, the variance calculation becomes

$$\mathrm{Var}[S_n] = \sigma^2 \left(n + 2 \sum_{i=1}^{n-1} (n-i) i^{1-d/\alpha} L(i) (-1)^i \right),$$

which is the same as in Proposition 1.6 except that the sum is now an alternating sum. Some heuristic analysis suggests that the asymptotic behavior of this is $\theta(n)$, the variance of standard Brownian motion.

We can also sum in other directions than e_1. The behavior depends on the value of d/α. Summing in an off-axis direction is in some sense equivalent to adding a constant to the distribution μ, that is, to add drift to the random walk. The drift accumulated after n steps is linear in n. Considering the random walks started at 0 and ne_1, we let the latter catch up to the hyperplane containing 0. Although the random walk has drift, the difference between two such random walks is a drift free random walk. Once the random walk started at ne_1 has caught up to 0, the random walks collied if their difference hits 0. Since this difference is a drift free random walk and has the same parameter α, what happens after ne_1 has caught up to 0 is the same as in the drift-free case.

There are three cases. The first is that the dispersion associated with Z_n is asymptotically greater than the linear drift (i.e., the dispersion grows super-linearly in n) and the drift is asymptotically negligible, giving the same result as if there were no drift. The second is that the dispersion grows asymptotically slower than the linear drift, making the dispersion term asymptotically negligible. The third case is that the dispersion grows linearly, making both contributions asymptotically relevant. It is unclear how to handle this case.

We know that the scaling of the random walk is such that $Z_n \approx b(n)U$ in distribution, and $b(n) \sim n^{1/\alpha}\ell(n)$. This can be read as saying that the dispersion (since U is centered) grows like $b(n)$. So $\alpha = 1$ is a critical value for whether the linear drift or the random walk's dispersion dominate. For the case that the linear drift dominates (i.e. $\alpha > 1$), we recall that the Green's function of the (drift-free) symmetric random walk asymptotically behaves like $G(x, 0) \approx |x|^{\alpha-d}$ (for $\alpha < d$, otherwise the random walk is recurrent). If the linear drift dominates, the dispersion is asymptotically negligible and the distribution of where the random walk hits the hyperplane containing

0 should concentrate around nv for some vector v, suggesting that the probability that 0 and ne_1 are in the same component should decay like $n^{\alpha-d}$. This is in contrast to the drift-free case in which the decay is like $n^{1-d/\alpha}$.

To summarize, we want to know the behavior of $\mathbb{P}(0 \sim n(e_1+x))$ for some nonzero x with $x \cdot e_1 = 0$, where $n(e_1 + x)$ should be interpreted as the nearest lattice point. If $\alpha > 1$, the linear drift dominates and we expect $\mathbb{P}(0 \sim n(e_1+x)) \sim n^{\alpha-d}$. If $\alpha < 1$, the dispersion dominates and we expect $\mathbb{P}(0 \sim n(e_1 + x)) \sim n^{1-d/\alpha}$. It is reassuring to observe that these coincide when $\alpha = 1$.

Note that, since we have assumed $d/2 < \alpha < d$ so that the e_1-direction sums to scale to fractional Brownian motion, the condition $\alpha > 1$ is always satisfied when $d \geq 2$ and impossible for $d = 1$. So when $d = 1$, the off-axis direction sums behave similarly to sums in the e_1 direction, and when $d \geq 2$, the off-axis direction sums behave fundamentally differently. While $\alpha < d$ is essential to give transience, it is possible that the assumption $\alpha > d/2$ can be relaxed to obtain something interesting in off-axis direction sums.

By Proposition 1.6, for the sum in a particular direction to have a fractional Brownian motion scaling limit, if we have $\mathbb{P}(0 \sim n(e_1 + x)) \sim n^{-s}$, we need the exponent $-s$ to be in $(-1, 0)$. If $\alpha > 1$, the connection probability exponent is $\alpha - d$, and we get the condition $\alpha \in (d - 1, d)$. For $d = 2$, this is the same as the condition $\alpha \in (d/2, d)$ which is required for the e_1 direction to have a fractional Brownian motion scaling limit. In $d \geq 3$, the condition for off-axis sums to give fractional Brownian motion cannot be satisfied.

We can summarize as follows: When $d = 1$, if $\alpha \in (1/2, 1)$, we get fractional Brownian motion in both the e_1 and off-axis directions, and the Hurst parameter is the same in either case: $\frac{1}{2}(3 - d/\alpha)$. When $d = 2$, if $\alpha \in (1, 2)$, we get fractional Brownian motion in both the e_1 and off-axis directions, but the Hurst parameter is different: $\frac{1}{2}(3 - d/\alpha)$ in the e_1-direction and $\frac{1}{2}(2 + \alpha - d)$ in off-axis directions. When

$d = 3$, if $\alpha \in (3/2, 2]$, we get fractional Brownian motion in the e_1 direction with Hurst parameter $\frac{1}{2}(3 - d/\alpha)$. No values of α give fractional Brownian motion in off-axis directions. For $d \geq 4$, neither the e_1 nor the off-axis directions give a fractional Brownian motion scaling limit for any value of α.

With this analysis of single-direction sums in place, we turn to the higher dimensional structure of the joint process. It seems clear how to define a random walk that may converge to fractional Brownian motion in terms of a single sum, but there are multiple choices for how to define the joint process. The process should be some $G(x, t)$ where $x \in \mathbb{R}^d$, or possibly some half plane. When fixing an x, the process $t \mapsto G(x, t)$ should be fractional Brownian motion. And it could be desireable for the joint process $G(x, t)$ to have stationary increments in time. It could be interesting for $G(x, t)$ to have a distribution which is translation invariant in x.

Two candidate limiting processes arise. In one, which we may call the radial model, the rays along which we take sums all emanate from the origin, but in different directions determined by x. In the other, which we may call the parallel model, the rays along which we take sums are all parallel, in the e_1 direction, but have different starting points determined by x.

First, we discuss the parallel model. We will define $G(x, t)$ as a limit in distribution of discretizations $G^{(n)}(x, t)$. In this case, we can define $S(x, t)$ for $x \in \mathbb{Z}^d$ and $t \in \mathbb{Z}$. We embed $\mathbb{Z}^d$ in $\mathbb{Z}^D$ by identifying $(x_1, x_2, ..., x_d)$ with $(0, x_1, ..., x_d)$. As before, for a vertex $v \in \mathbb{Z}^D$, we let Y_v be the color of the component containing v. Then we set

$$S(x, t) = \sum_{i=1}^{t} Y_{x+ie_1}.$$

This sum starts at x and proceeds in the e_1 direction for t steps. S is defined for integer arguments, and we can extend to $\mathbb{R}^D$ by choosing the nearest integer point. Let σ_n^2 be the variance of $S(x, n)$ (which doesn't depend on x). We will also need a spatial scaling factor $h(n)$. Then define

$$G^{(n)}(x,t) = S(h(n)x, nt)/\sigma_n.$$

This is scaled such that $G^{(n)}(x,1)$ has unit variance and is defined from n terms of the sum.

Next, we discuss the radial model. As before, will define $G(x,t)$ as a limit in distribution of discretizations $G^{(n)}(x,t)$. First, we define $S(x,t)$ for $x \in \mathbb{R}^d, t \in \mathbb{Z}_{\geq 0}$. We can identify each vector $x = (x_1, x_2, \ldots, x_d)$ with a direction $\hat{x} = (1, x_1, \ldots, x_d) \in \mathbb{R}^D$. Then $S(x,t)$ will be defined by summing in this direction. To make things precise, for $z \in \mathbb{R}^D$ let $[z]$ be the point in $\mathbb{Z}^D$ obtained by rounding the entries of z to the nearest integer. Then we can define

$$S(x,t) = \sum_{i=1}^{t} Y_{[i\hat{x}]}.$$

This is the sum started at 0 that proceeds for t steps in the $\hat{x}$ direction. Based on the discussion of off-axis sums, the correlation decay in this direction will match that of the e_1 direction if $\alpha > 1$. Since $d/2 < \alpha < d$, this happens when $d > 1$. When $\alpha > 1$, if σ_n^2 is the variance of $S(0,n)$, we can define

$$G^{(n)}(x,t) = S(x, nt)/\sigma_n.$$

It's not entirely clear what should be done in $d = 1$, since the connection probabilities decay asymptotically differently in the e_1 and off-axis directions. It has some consistency in the following sense: for any fixed x, we can increase n to n' to sample $G^{(n')}(x,t)$ in a way that preserves the history of the sum defining $G^{(n)}(x,t)$. The first tn increments of the sum are still there, now they only cover up to time tn/n'. In the parallel model, however, both x and t must both be rescaled to do something analogous.

A.2 Auxiliary proofs

Proposition A.1. *Suppose U is a nondegenerate d-dimensional stable law of parameter $\alpha \in (0, 2)$ which is not supported on a proper hyperplane. Then the density of U is a continuous and bounded.*

Proof. While the density of stable laws is not known except in specific cases (such as $d = \alpha = 1$, which is the Cauchy distribution), the characteristic function can be described explicitly. If the characteristic function is absolutely integrable, then the density is continuous. We show that if the characteristic function is not absolutely integrable, then U is supported on a hyperplane.

A d-dimensional stable law is characterized by its index α, a location vector $\delta \in \mathbb{R}^d$, and a finite measure Λ on the unit sphere in $\mathbb{R}^d$. The characteristic function of a multivariate stable law is

$$\varphi(t) = \exp\left(-\int |t \cdot s|^\alpha + i\nu(t \cdot s, \alpha)\, d\Lambda(s) + it \cdot \delta\right),$$

where

$$\nu(x, \alpha) = \begin{cases} -\mathrm{sign}(x)|x|^\alpha \tan(\pi\alpha/2), & \alpha \neq 1, \\ (2/\pi)x \ln|x|, & \alpha = 1. \end{cases}$$

Any distribution has a density which is continuous and bounded if the characteristic function is absolutely integrable: $\int_{\mathbb{R}^d} |\phi(t)|\, dt < \infty$. Since the terms $i\nu(t \cdot s, \alpha)$ and $it \cdot \delta$ are purely imaginary, they do not contribute to the magnitude of $\phi(t)$. Let $\hat{t} = t/|t|$. Since

$$\int_{\mathbb{R}^d} |\varphi(t)|\, dt = \int_{\mathbb{R}^d} \exp\left(-|t|^\alpha \int |\hat{t} \cdot s|^\alpha\, d\Lambda(s)\right) dt,$$

it suffices to show that $\int |t \cdot s|^\alpha \, d\Lambda(s) > 0$ for every unit vector t. Suppose for contradiction that there is a unit vector t_0 with $\int |t_0 \cdot s|^\alpha \, d\Lambda(s) = 0$. Then the univariate characteristic function φ_0 given by $\varphi_0(u) = \varphi(ut_0)$ is the characteristic function of the univariate random variable $t_0 \cdot U$. But

$$|\varphi_0(u)| = \exp\left(-|u|^\alpha \int |t_0 \cdot s|^\alpha \, d\Lambda(s)\right) = 1.$$

If $\tilde{U}$ is an independent copy of U, the characteristic function of $t_0 \cdot U - t_0 \cdot \tilde{U}$ is given by $\tilde{\varphi}(u) = \varphi_0(u)\varphi_0(-u) = \varphi_0(u)\overline{\varphi_0(u)} = |\varphi_0(u)|^2 = 1$. This is the characteristic function of the random variable which is almost surely 0. So $t_0 \cdot U$ is equal to $t_0 \cdot \tilde{U}$ almost surely. But they are independent, which means that $t_0 \cdot U$ is almost surely constant, i.e., U is almost surely supported on a hyperplane. This contradiction lets us conclude that $\int |t \cdot s|^\alpha \, d\Lambda(s) > 0$ for every unit vector t, and the characteristic function of U is absolutely integrable, so the density is continuous and bounded. □

A.3 Variance heuristics

There are a couple more ways to estimate the variance of S_n.One computation involves a Green's function and the other uses the characteristic functions.

For the characteristic function approach, we can upper bound the probability in question by $\mathbb{E} \sum_{i \geq 0} \mathbf{1}_{\{X_{n+i} = X_i\}}$ (in fact, they differ by the constant $G_Z(0,0)$, where G_Z is a Green's function defined later in this proof). We let φ be the characteristic function of the increment distribution, and $b(n)$ be a function that $X_n b^{-1}(n) \to U$ in distribution for a stable law U. A note in [22] (in the proof of Corollary 3.2, after equation 3.d) states that

$$\int_{(-\pi,\pi]^d} \varphi(\xi)^i \varphi(-\xi)^j \, d\xi \leq C b(i+j)^{-d}.$$

Recall that $b(x) = x^{1/\alpha}\ell(x)$ for some slowly varying function ℓ. Using the bound, we can estimate

$$
\begin{aligned}
\mathbb{E} \sum_{i \geq 0} \mathbf{1}_{\{X_{n+i} = X_i\}} &= \sum_{i \geq 0} (2\pi)^{-d} \int_{(-\pi,\pi]^d} \varphi(\xi)^{n+i} \varphi(-\xi)^i \, d\xi \\
&\leq C \sum_{i \geq 0} b(2i+n)^{-d} \\
&\leq C \sum_{i \geq n} b(i)^{-d}. \\
&\sim \int_n^\infty x^{-d/\alpha} \ell(x)^{-d} \, dx.
\end{aligned}
$$

We observe that $\ell(x)^{-d}$ is slowly varying:

$$
\lim_{x \to \infty} \frac{\ell(cx)^{-d}}{\ell(x)^{-d}} = \left(\lim_{x \to \infty} \frac{\ell(cx)}{\ell(x)} \right)^{-d} = 1.
$$

It is a fact ([1], Proposition 1.5.10) that for $\beta < -1$, if L is slowly varying, then

$$
\frac{x^{\beta+1} L(x)}{\int_x^\infty t^\beta L(t) \, dt} \to -\beta - 1 \text{ as } x \to \infty.
$$

Noting that $-d/\alpha < -1$, we conclude that the probability in question can be upper bounded by a slowly varying function times $n^{1-d/\alpha}$.

This is only an upper bound. We should be able to replace the upper bound with just a computation, using the fact that for n large, the characteristic function is close to that of a stable law, and we should be able to get a decent grasp on the value in that case. Also, this is the probability that a random walk started at 0 returns to 0 after $n + 2i$ steps for some $i \geq 0$. There may be a direct way to do this.

For the Green's function approach, we only sketch an informal computation which treats things as continuous and uses convenient example functions for the purpose of integration. Set $Z_i = X_{n+i} - X_i$. Then $(Z_i)_{i \geq 0}$ is a random walk with random initial condition $Z_0 = X_n$ and increment in the domain of attraction of a (different) stable law with parameter α, and we are concerned with the event that the random walk Z hits 0. The Green's function of Z is defined as

$$G_Z(x,y) = \mathbb{E}[\text{number of times } Z_i = y | Z_0 = x],$$

the expected number of times Z hits y if started at x. By the Markov property of Z,

$$G_Z(x,y) = \mathbb{P}(Z \text{ hits } y | Z_0 = x)\, G_Z(y,y),$$

and $P(Z \text{ hits } y|Z_0 = x) = G_Z(x,y)/G_Z(y,y)$ when these are finite. We know that $G_Z(x,0) \sim |x|^{\alpha-d}L(x)$ (see [22]), for some slowly varying function L. Then the probability that Z hits 0 is

$$\mathbb{E}\, G_Z(X_n, 0).$$

As an informal estimate of this expectation, we approximate the density of X_1 as (up to a constant) $\min(1,|x|)^{-(\alpha+d)}$, the density of X_n as that of $n^{1/\alpha}X_1$, and treat the Green's function as (again, up to a constant) $G_Z(x,0) = \min(1,|x|)^{\alpha-d}$. With these simplifications in place, we can explicitly integrate with respect to the radius of X_1:

$$\begin{aligned}
\mathbb{E}\, G_Z(X_n,0) &= \mathbb{E}\, \min(1,|X_n|)^{\alpha-d} \\
&= \mathbb{E}\, \min(1, n^{1/\alpha}|X_1|)^{\alpha-d} \\
&= \mathbb{E}\, \min(1, n^{1-\frac{d}{\alpha}}|X_1|^{\alpha-d}) \\
&= \int_0^1 \min(1, n^{1-\frac{d}{\alpha}} r^{\alpha-d}) r^{d-1} dr \\
&\quad + \int_1^\infty \min(1, n^{1-\frac{d}{\alpha}} r^{\alpha-d}) r^{-d+\alpha} r^{d-1} dr \\
&= n^{-\frac{d}{\alpha}} \left(\frac{1}{d} - \frac{1}{\alpha}\right) + n^{1-\frac{d}{\alpha}} \left(\frac{1}{\alpha} + \frac{1}{d}\right).
\end{aligned}$$

The dominant term here is the $n^{1-d/\alpha}$ term. We see that, when making convenient assumptions and approximations, $\mathbb{P}(X_{n+i} = \tilde{X}_i \text{ for some } i) \sim n^{1-d/\alpha}$. I expect that with some work, this can be made rigorous when using an appropriate class of actual distributions.

Next, we sketch the idea of an alternate proof of showing the tail sigma algebra is trivial.

Remark A.2. Here's the idea. Fix two vertices, x and y. In the model, Y_x^m and Y_y^m are obtained by taking random walks which proceed until they hit A_{m-1}, at which point Y_x assumes the value. And Y_x^m is the expected value of this over the random walks. So the idea is to construct a coupling of the random walks such that with high probability, they land on the same spot in A_{m-1}. Then the expected values of two random variables which are the same with high probability are very close. From here, we get that $Y_x^{-\infty} = Y_y^{-\infty}$ almost surely. (I think it's almost surely). Then $S_n^{-\infty} = nY_0^{-\infty}$. And this has either variance 0 or Cn^2. But we know S_n has variance asymptotically smaller than n^2. So this variance has to be 0, which implies that the tail is trivial. An alternative approach to use a direct formula for S_x^m as a weighted sum on the hyperplane A_{m-1} and show that the weights for S_x^m and S_y^m are very close. Perhaps the local limit theorem is strong enough for this. Then once the weights are close, the weighted sums are close. Taking m to $-\infty$ could give some convergence. Would have to do the computations to see what type of convergence this needs and see if things work out.

A.4 Remarks

Remark A.3. For off-axis sums, we need $\alpha > 1$. Since $d/2 < \alpha < d$, this is impossible in $d = 1$. This suggests that the radial model may be infeasible in $d = 1$. Another possibility, however, is that we ignore this concern, since the e_1 direction is only relevant for $x = 0$, and perhaps the rest of the model will still be interesting.

Remark A.4. The process $(Y_{x+ne_1})_{\{x:\, x\cdot e_1=0\}}$ as n varies has the Markov property in n. In other words, the joint distribution of a slice only depends on the values of Y_v for

v in the previous slice.

Remark A.5. If α is too big (tails are too skinny) then the random walks at every pair of points coalesce and there is only one component. If α is too small (tails too fat), then it seems likely that $P(0 \sim ne_1)$ will decay too quickly, and random walks will just become (usual) Brownian motion.

Specifically, we must have $d/2 < \alpha < d$ to get anything interesting. If $\alpha > d$, the random walk is recurrent and all points are in the same connected component.

Remark A.6. The requirement $d/2 < \alpha < d$ is somewhat restrictive since $\alpha \in (0, 2]$. This tells us that if $d \geq 4$, for all values of α, the coalescence probabilities decay too quickly to give a fractional Brownian motion limit (we should just get usual Brownian motion). In this case, even Gaussian tails give random walks with too small of coalescence probabilities.

Remark A.7. There may be some continuity concerns with $G^{(n)}(x, t)$. In one dimension, we can linearly interpolate. In higher, we can't. In either the radial or parallel model, things seem fine when x is fixed, but fixing t and allowing x to vary may result in sizeable jumps in the process value. In the discretization, increasing t by 1 can only change the value of the sum by 1, but changing x to a neighbor can change the value of the sum by up to t (assuming all increments are ± 1). But maybe this is not an issue at all when the spacial rescaling is appropriate. And possibly this is related to the fact that Gaussian free fields are not particularly nice.

APPENDIX B
PINN SPECIFICATIONS

B.1 Neural Network Approximation Theory

Lemma B.1. *Let $f \in C^1([0,1]^d)$. For any $n \in \mathbb{N}$ with $n \geq \max\{4, d\}$, there is a ReLU network f_θ with width C_1 and depth $C_2 n \log n$ such that*

$$\|f - f_\theta\|_{L^\infty([0,1]^d)} \leq C_3 n^{-\frac{2}{d}},$$

where the constants only depend on d, the L^∞ norm of f, and the Lipschitz constant of f.

Proof. This is Theorem 1.1 from [23] adapted with $s = 1, N = 1, L = n$. □

If we desire ϵ accuracy in approximation, this leads to the corollary:

Corollary B.2. *Let $\epsilon > 0$, and let $f \in C^1([0,1]^d)$. Then there is a ReLU network f_θ with width C_1 and depth $C_2 \epsilon^{-\frac{d}{2}} \log\left(\frac{1}{\epsilon}\right)$ such that*

$$\|f - f_\theta\|_{L^\infty([0,1]^d)} \leq \epsilon,$$

where the constants only depend on d, the L^∞ norm of f, and the Lipschitz constant of f.

Proof. This is Theorem 1.1 from [23] adapted with $s = 1, N = 1, L = n$. □

The above results concerned approximation of scalar-valued function. If our target function is vector-valued, we can approximate each component function in L^∞.

Lemma B.3. *Let $\epsilon > 0$, and let $\vec{f} \in C^1([0,1]^d; \mathbb{R}^{d_2})$. Then there is a ReLU network f_θ with width C_1 and depth $C_2 \epsilon^{-\frac{d}{2}} \log\left(\frac{1}{\epsilon}\right)$ such that*

$$\|\vec{f} - f_\theta\|_{L^\infty([0,1]^d)} \leq \epsilon,$$

where the constants only depend on d, d_2, the L^∞ norm of $\vec{f}$, and the Lipschitz constant of $\vec{f}$.

Proof. Let $i \in [d_2]$, and let f^i denote the ith component function of $\vec{f}$. By applying Corollary B.2, we obtain neural networks f_θ^i such that

$$\sup_{x\in[0,1]^d} |f^i(x) - f_\theta^i(x)| \leq \frac{\epsilon}{\sqrt{d_2}}.$$

We can apply this for all i to obtain neural networks f_θ^i. Each of these networks have width C_1 and depth $C_2\epsilon^{-\frac{d}{2}} \log\left(\frac{1}{\epsilon}\right)$ where these constants may also depend on d_2. By adding identity layers as necessary (which can be done by ReLU networks), we can assume all of the f_θ^i networks have the same depth.

Consider the neural network f_θ obtained by stacking each network f_θ^i on the same network. Then f_θ implements a d_2-dimensional vector-valued function with f_θ^i as the ith component, and for all $x \in [0,1]^d$ we have

$$\|\vec{f}(x) - f_\theta(x)\| = \left(\sum_{i=1}^{d_2} |f^i(x) - f_\theta^i(x)|^2\right)^{1/2} \leq \left(\sum_{i=1}^{d_2} \frac{\epsilon^2}{d_2}\right)^{1/2} = \epsilon,$$

which proves that $\sup_{x\in[0,1]^d} \|\vec{f}(x) - f(x)\| < \epsilon$. $\square$

We now prove Lemma 3.1 on the approximation of generators in $\mathbb{W}_1$ for Hölder densities. Note that the previous results implied bounds on the L^∞ norm. When working on bounded domains, this also implies a bound on the L^1 error, which we will utilize in the following.

Proof of Lemma 3.1. Suppose for simplicity that Ω is a cube, and without loss of generality that $\Omega = [0,1]^d$. By Caffarelli's regularity theory (Theorem 12.50 in [33]), there is a function $T \in C^{s+1}(\Omega)$ such that $\mu = T_\sharp\rho$. Let $\eta > 0$, and consider the compact set $K_\eta \subseteq \Omega$ defined as the largest convex set that is distance η from the boundary $\partial\Omega$, i.e. $K_\eta = [\eta, 1-\eta]^d$.

Then the function T is uniformly C^{s+1} on the set K_η. Suppose $\mathsf{N} \in C^{s+1}(\overline{\Omega})$ vanishes on $\partial\Omega$ and has maximum value 1, e.g. $\mathsf{N}(x_1, \ldots, x_d) = \prod_{i=1}^d 4x_i(1-x_i)$. Then the function $\tilde{T}(x) = \frac{T(x)-x}{N(x)} \in C^{s+1}(K_\eta)$. Let f_ϕ be a ReLU network such that

$$\left\| f_\phi - \tilde{T} \right\|_{L^1(K_\eta)} < \epsilon/2.$$

Extend f_ϕ to have magnitude at most $\sqrt{d}$ outside of K_η. Define

$$g_\phi(x) = f_\phi(x)\mathsf{N}(x) + x.$$

Then we have

$$\begin{aligned}
\mathbb{W}_1(T_\sharp\rho, (g_\phi)_\sharp\rho) &\leq \|T(x) - g_\phi(x)\|_{L^1(\rho)} \\
&= \int_{K_\eta} |T(x) - g_\phi(x)| \, d\rho(x) + \int_{\Omega\setminus K_\eta} |T(x) - g_\phi(x)| \, d\rho(x) \\
&= \int_{K_\eta} |\tilde{T}(x)\mathsf{N}(x) + x - (f_\phi(x)\mathsf{N}(x) + x)| \, d\rho(x) \\
&\qquad + \int_{\Omega\setminus K_\eta} |T(x) - f_\phi(x)N(x) + x| \, d\rho(x) \\
&= \int_{K_\eta} |\tilde{T}(x) - f_\phi(x)||\mathsf{N}(x)| \, d\rho(x) \\
&\qquad + \int_{\Omega\setminus K_\eta} |T(x) - f_\phi(x)N(x) + x| \, d\rho(x) \\
&\leq \int_{K_\eta} |\tilde{T}(x) - f_\phi(x)| \, d\rho(x) + \int_{\Omega\setminus K_\eta} \sqrt{d} + \sqrt{d} + 1 \, d\rho(x) \\
&\leq \|\tilde{T} - f_\phi\|_{L^1(K_\eta)} + \rho(\Omega \setminus K_\eta)(2\sqrt{d} + 1) \\
&\leq \epsilon/2 + \rho(\Omega \setminus K_\eta)(2\sqrt{d} + 1) = ().
\end{aligned}$$

We will then select η by estimating in the cube case

$$
\begin{aligned}
\rho(\Omega \setminus K_\eta) &= \rho([0,1]^d \setminus [\eta, 1-\eta]^d) \\
&\leq \sum_{i=1}^{d} \rho([0,1]^{i-1} \times [0,\eta] \times [0,1]^{d-i}) + \rho([0,1]^{i-1} \times [1-\eta, 1] \times [0,1]^{d-i}) \\
&\leq \sum_{i=1}^{d} 2\eta = 2d\eta,
\end{aligned}
$$

so we choose $\eta < \frac{\epsilon}{8d^{3/2}+4d}$ to finish () $\leq \epsilon$. Now, g_ϕ is surjective because of Lemma Theorem 3.2. In addition, the size of the neural network f_ϕ depends on ϵ and the Hölder norm of $\tilde{T}$ inside of K_η. This depends on unknown finite Hölder norm of T and the regularity of N, which is all finite. □

We now handle the case of a measure μ on a spacetime domain $\Omega \times [0,1]$, with density $q \in \mathcal{C}^\alpha(\Omega \times [0,1])$. If marginal distribution of μ in time is uniform (i.e. $\mu(\Omega \times [s,t]) = t - s$), then we can generate a distribution close to μ using a time-preserving generator. In this case, we approximate μ using a "piecewise-constant" in time measure μ_n, and then μ_n can be generated using a time-preserving generator.

Lemma B.4. *Suppose μ has density $q \in \mathcal{C}^\alpha(\Omega \times [0,1])$ with uniform marginal distribution in time. Then there is a surjectivity enforced neural network $g_\phi : \mathbb{R}^{d+1} \to \mathbb{R}^d$ whose last component is the identity such that*

$$
\mathbb{W}_1(\mu, (g_\phi)_\sharp \rho) < \epsilon.
$$

Proof sketch. For any $n \in \mathbb{N}$, define the grid $0 = t_0 < t_1 < \cdots < t_n = 1$ with $t_i = \frac{i}{n}$. Consider the function q_n which is a left-hand rule approximation in time of the density q:

$$
q_n(x,t) = \sum_{i=0}^{n-1} q(x, t_i) \mathbf{1}_{[t_i, t_{i+1})}(t).
$$

Consider the measure $\mu_n \in \mathbb{P}(\Omega \times [0,1])$ which has density q_n. The distance between q and q_n in the L^1 norm is

$$\begin{aligned}
\|q - q_n\|_{L^1(\Omega\times[0,1])} &= \int_\Omega \int_0^1 |q(x,t) - q_n(x,t)|\, dt\, dx && \text{(B.1.1)}\\
&= \sum_{i=0}^{n-1} \int_\Omega \int_{t_i}^{t_{i+1}} |q(x,t) - q(x,t_i)|\, dt\, dx && \text{(B.1.2)}\\
&\leq \sum_{i=0}^{n-1} \int_\Omega \int_{t_i}^{t_{i+1}} C\left(\frac{1}{n}\right)^{\alpha} dt\, dx && \text{(B.1.3)}\\
&= C\sum_{i=0}^{n-1} \int_\Omega \frac{1}{n^{1+\alpha}}\, dx && \text{(B.1.4)}\\
&= C\frac{n}{n^{1+\alpha}} = Cn^{-\alpha}, && \text{(B.1.5)}
\end{aligned}$$

where the constant C is the Hölder coefficient of q. By Theorem B.6, we have

$$\|q - q_n\|_{L^1} < Cn^{-\alpha} \implies W_1(\mu, \mu_n) \leq C_2 n^{-\alpha},$$

where $C_2 = \frac{C_1 \text{diam}\Omega\times[0,1]}{2}$. This means we can approximate μ by μ^n for sufficiently large n.

Now for all $t \in [0,1]$, the function $q^n(\cdot, t) \in C^\alpha(\Omega)$, so we can approximate it by neural networks similar to in the proof of Lemma Theorem 3.1. We will do this for each of the time grid points t, so for all $i \in \{0, 1, \ldots, n-1\}$, let f_θ^i be a ReLU network such that

$$\|f_\theta^i - q^n(\cdot, t_i)\|_{L^\infty(\Omega)} < \epsilon.$$

Let $\tilde{\times}^{\delta_1}$ be a neural network approximation to the multiplication function, and $\mathbf{1}_i^{\delta_2}$ be a neural network approximation to the indicator function of the interval $[t_i, t_{i+1})$. Then consider the neural network

$$g_\theta(x,t) = \sum_{i=0}^{n-1} \tilde{\times}^{\delta_1}(f_\theta^i(x), \mathbf{1}_i^{\delta_2}(t)).$$

After appropriate choices of δ_1, δ_2, it can be shown that $\mathbb{W}_1(\mu_n, (g_\theta)_\sharp \rho) < C\epsilon$, which leads to the result.

$\square$

B.2 Minor Technical Tools

Proposition B.5. *Let $f, g : \Omega \to \mathbb{R}^m$ where $\Omega \subseteq \mathbb{R}^d$. Suppose $\rho \in \mathcal{P}(\Omega)$. Then*

$$\mathbb{W}_1(f_\sharp \rho, g_\sharp \rho) \leq \|f - g\|_{L^1(\rho)}.$$

Proposition B.6. *Let Ω be compact and equipped with some probability measure λ. Suppose $\mu, \nu \in \mathcal{P}(\Omega)$ have densities f and g with respect to λ. Then*

$$\mathbb{W}_1(\mu, \nu) \leq \frac{diam(\Omega)}{2} \|f - g\|_{L^1(\lambda)}.$$

Proposition B.7 (Caffarelli's Regularity Theory)**.** *Let $\Omega \subseteq \mathbb{R}^d$ be a convex bounded open set, and suppose $\mu, \nu \in \mathcal{P}(\Omega)$ have densities q_μ, q_ν with respect to the Lebesgue measure. Suppose there is a $m > 0$ such that $m^{-1} \leq q_\mu, q_\nu \leq m$. Then if $q_\mu, q_\nu \in \mathcal{C}^s(\Omega)$ for any $s > 0$, the unique optimal transport map for the squared distance cost between μ, ν is a function $T \in C^{s+1}(\Omega)$.*

Proof. See [4] for the proof (and [33] for a more modern statement). $\square$

B.3 Optimization Remark

The loss for the generator network consists of the squared mean PDE residual plus a penalty term for generated points outside of the PDE domain Ω (in the case of bounded Ω). Specifically, if $d(x, \Omega)$ denotes the Euclidean distance from a point

$x \in \mathbb{R}^d$ to the domain, the generator loss is given by

$$\text{Loss} = \frac{1}{n}\sum_{k=1}^{n}\left(\|\mathcal{D}u_\theta(g_\phi(z_k))\|^2 - 1000000 \cdot d(g_\phi(z_k), \Omega)^2\right).$$

REFERENCES

[1] N. H. Bingham, C. M. Goldie, and J. L. Teugels, *Regular Variation* (Encyclopedia of Mathematics and its Applications). Cambridge University Press, 1987.

[2] J. Blitz, "The phase shifting coalescing-branching random walk," Senior Thesis, Princeton University, 2021.

[3] B. Bock, M. Damron, and J. Hanson, "Random nearest neighbor graphs: The translation invariant case," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 59, no. 2, pp. 849–866, 2023.

[4] L. A. Caffarelli, "The regularity of mappings with a convex potential," *Journal of the American Mathematical Society*, vol. 5, no. 1, pp. 99–104, 1992.

[5] T. De Ryck, S. Lanthaler, and S. Mishra, "On the approximation of functions by tanh neural networks," *Neural Networks*, vol. 143, pp. 732–750, 2021.

[6] T. De Ryck and S. Mishra, "Numerical analysis of physics-informed neural networks and related models in physics-informed machine learning," 2024. arXiv: 2402.10926.

[7] M. Dissanayake and N. Phan-Thien, "Neural-network-based approximations for solving partial differential equations," *Communications in Numerical Methods in Engineering*, vol. 10, no. 3, pp. 195–201, 1994.

[8] J. L. Doob, *Measure Theory* (Graduate Texts in Mathematics). Springer, 1994.

[9] R. Drogin, "Long range voter models and dynamical fractional brownian motion," 2023. arXiv: 2311.03662 `[math.PR]`.

[10] R. Durrett, *Probability: Theory and Examples*, Second. Duxbury Press, 1996.

[11] B. V. Gnedenko and A. N. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables* (Addison-Wesley Series in Statistics), Revised edition, K. L. Chung, Ed., trans. by K. L. Chung. Addison-Wesley Publishing Company, 1968.

[12] A. Griewank and A. Walther, *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, Second. Philadelphia: SIAM, 2008.

[13] P. Hall and C. C. Heyde, *Martingale Limit Theory and Its Application* (Probability and Mathematical Statistics: A Series of Monographs and Textbooks), Second, Z. W. Birnbaum and E. Lukacs, Eds. Academic Press, 1980.

[14] A. Hammond and S. Sheffield, “Power law Pólya’s urn and fractional Brownian motion,” *Probability Theory and Related Fields*, vol. 157, 2013.

[15] J. Han, A. Jentzen, and W. E, “Solving high-dimensional partial differential equations using deep learning,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 34, pp. 8505–8510, 2018.

[16] J. Hou, Y. Li, and S. Ying, “Enhancing PINNs for solving PDEs via adaptive collocation point movement and adaptive loss weighting,” *Nonlinear Dynamics*, vol. 111, no. 16, pp. 15 233–15 261, 2023.

[17] J. L. Igelbrink and A. Wakolbinger, “Asymptotic gaussianity via coalescence probabilities in the Hammond-Sheffield urn,” *Latin American Journal of Probability and Mathematical Statistics*, vol. 20, pp. 53–74, 2023.

[18] Y. Kossale, M. Airaj, and A. Darouichi, “Mode collapse in generative adversarial networks: an overview,” in *2022 8th International Conference on Optimization and Applications (ICOA)*, IEEE, 2022, pp. 1–6.

[19] I. E. Lagaris, A. Likas, and D. I. Fotiadis, “Artificial neural networks for solving ordinary and partial differential equations,” *IEEE Transactions on Neural Networks*, vol. 9, no. 5, pp. 987–1000, 1998.

[20] G. F. Lawler and V. Limic, *Random Walk: A Modern Introduction* (Cambridge Studies in Advanced Mathematics). Cambridge University Press, 2010.

[21] J.-F. Le Gall, *Brownian Motion, Martingales, and Stochastic Calculus* (Graduate Texts in Mathematics), 1st ed. 2016. Switzerland: Springer Cham, 2016.

[22] J.-F. Le Gall and J. Rosen, “The range of stable random walks,” *The Annals of Probability*, vol. 19, no. 2, pp. 650–705, 1991.

[23] J. Lu, Z. Shen, H. Yang, and S. Zhang, “Deep network approximation for smooth functions,” *SIAM Journal on Mathematical Analysis*, vol. 53, no. 5, pp. 5465–5506, 2021.

[24] L. Lu, X. Meng, Z. Mao, and G. E. Karniadakis, “DeepXDE: A deep learning library for solving differential equations,” *SIAM Review*, vol. 63, no. 1, pp. 208–228, 2021.

[25] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, “Are GANs created equal? A large-scale study,” *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[26] G. K. M. Raissi P. Perdikaris, “Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations,” *Journal of Computational physics*, vol. 378, pp. 686–707, 2019.

[27] L. McClenny and U. Braga-Neto, “Self-adaptive physics-informed neural networks using a soft attention mechanism,” 2020. arXiv: 2009.04544.

[28] M. A. Nabian, R. J. Gladstone, and H. Meidani, “Efficient training of physics-informed neural networks via importance sampling,” *Computer-Aided Civil and Infrastructure Engineering*, vol. 36, no. 8, pp. 962–977, 2021.

[29] E. L. Rvačeva, “On domains of attraction of multi-dimensional distributions,” in *Selected Translations in Mathematical Statistics and Probability*, trans. by S. G. Ghurye, vol. 2, American Mathematical Society, 1962, pp. 183–205.

[30] N. Sukumar and A. Srivastava, “Exact imposition of boundary conditions with distance functions in physics-informed deep neural networks,” *Computer Methods in Applied Mechanics and Engineering*, vol. 389, 2022.

[31] K. Tang, X. Wan, and C. Yang, “DAS-PINNs: A deep adaptive sampling method for solving high-dimensional partial differential equations,” *Journal of Computational Physics*, vol. 476, 2023.

[32] H. Thanh-Tung and T. Tran, “Catastrophic forgetting and mode collapse in GANs,” in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–10.

[33] C. Villani *et al.*, *Optimal Transport: Old and New*. Springer, 2009, vol. 338.

[34] C. Visser, A. Heinlein, and B. Giovanardi, “PACMANN: Point adaptive collocation method for artificial neural networks,” 2024. arXiv: 2411.19632.

[35] C. Wang, S. Li, D. He, and L. Wang, “Is L^2 physics informed loss always suitable for training physics informed neural network?” *Advances in Neural Information Processing Systems*, vol. 35, pp. 8278–8290, 2022.

[36] C. Wu, M. Zhu, Q. Tan, Y. Kartha, and L. Lu, “A comprehensive study of non-adaptive and residual-based adaptive sampling for physics-informed neural networks,” *Computer Methods in Applied Mechanics and Engineering*, vol. 403, 2023.

[37] Y. Zang, G. Bao, X. Ye, and H. Zhou, “Weak adversarial networks for high-dimensional partial differential equations,” *Journal of Computational Physics*, vol. 411, 2020.