Electron. Commun. Probab. **18** (2013), no. 87, 1–10.
DOI: 10.1214/ECP.v18-2784
ISSN: 1083-589X

ELECTRONIC COMMUNICATIONS in PROBABILITY

Increasing paths in regular trees

Matthew I. Roberts[*] Lee Zhuo Zhao[†]

Abstract

We consider a regular n-ary tree of height h, for which every vertex except the root is labelled with an independent and identically distributed continuous random variable. Taking motivation from a question in evolutionary biology, we consider the number of paths from the root to a leaf along vertices with increasing labels. We show that if $\alpha = n/h$ is fixed and $\alpha > 1/e$, the probability that there exists such a path converges to 1 as $h \to \infty$. This complements a previously known result that the probability converges to 0 if $\alpha \leq 1/e$.

Keywords: evolutionary biology; trees; branching processes; increasing paths.
AMS MSC 2010: Primary 60J80, Secondary 60C05; 92D15.
Submitted to ECP on May 8, 2013, final version accepted on November 7, 2013.
Supersedes arXiv:1305.0814.

1 Introduction

Consider a regular n-ary tree of height h, where $n = \lfloor \alpha h \rfloor$. To each vertex except the root attach an independent and identically distributed continuous random variable. We ask whether there is a path from the root to a leaf whose labels only increase. Nowak and Krug [9] called this *accessibility percolation* and showed that $\mathbb{P}(\text{there exists an increasing path}) \to 0$ as $n \to \infty$ if $\alpha \leq 1/e$, whereas if $\alpha > 1$ then there exists some $p > 0$ depending on α such that $\mathbb{P}(\text{there exists an increasing path}) > p$. We give a complete characterisation in terms of α, showing that there is a phase transition at $\alpha = 1/e$.

Theorem 1.1. *Suppose that* $n = \lfloor \alpha h \rfloor$. *As* $h \to \infty$,

$$\mathbb{P}(\textit{there exists an increasing path}) \to \begin{cases} 0 & \textit{if } \alpha \leq 1/e, \\ 1 & \textit{if } \alpha > 1/e. \end{cases}$$

Given the result of [9] mentioned above, it suffices to prove the second statement. In fact we will show that for any $\alpha > 1/e$, there exist $\delta > 0$ and $\eta > 0$ such that

$$\mathbb{P}(\text{there exist at least } \exp(\delta h) \text{ increasing paths}) \geq 1 - \exp(-\eta h).$$

This suggests that we might be able to discover more around the critical point $1/e$, and indeed by essentially the same methods we are able to obtain the following finer result.

[*] Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK.
E-mail: mattiroberts@gmail.com
[†] Statistical Laboratory, University of Cambridge, Cambridge, CB3 0WB, UK.
E-mail: lzz20@statslab.cam.ac.uk

Theorem 1.2. *Suppose that* $n = \left(\frac{1+\beta_h}{e}\right) h$*, where* $\beta_h \to 0$ *as* $h \to \infty$*. Then as* $h \to \infty$*,*

$$\mathbb{P}(\textit{there exists an increasing path}) \to \begin{cases} 0 & \textit{if } \log h - 2h\beta_h \to \infty, \\ 1 & \textit{if } h\beta_h / \log h \to \infty. \end{cases}$$

1.1 Biological motivation

Consider the following simplified model of evolution in a population. Each genetic type, or *genotype*, in the population has an associated fitness. A particular genotype may give rise to multiple new genotypes through mutations, which either replace the original wild genotype or disappear from the population. For a haploid asexual population, the dynamics of evolution are governed by the population size N, the selection coefficient s and the mutation rate μ [3]. We make the following two assumptions on these three parameters:

1. $Ns \gg 1$. By a classical formula of Kimura [6], only mutations which give rise to a fitter genotype can replace the wild genotype and survive.

2. μ is sufficiently small such that mutations arise and either replace the wild genotype or become extinct one at a time. Therefore, there can be at most two genotypes in the population at any given time, of which one is a direct mutant of the other.

Together, these two assumptions form what is known in the evolutionary biology literature as the *strong selection weak mutation* (SSWM) regime [4, 10]. Under such a setting, the only possible evolutionary paths of genotypes are ones with increasing fitness. In the evolutionary biology literature, these increasing paths are known as *selectively accessible* [3, 11, 12].

To analyse the number of such paths, we also require the relationship between genotype and fitness. For this, we use the *House of Cards* model [7, 8], in which every genotype has an independent and identically continuously distributed fitness. Since we only care about whether the fitnesses along a path are in increasing order, as long as the random variables are continuous, the precise distribution is not important.

The space of genotypes together with their fitnesses form a labelled graph. If we further assume that the population initally consists of one single genotype, and that separate mutations never give rise to the same genotype, then the space of genotypes becomes a rooted tree. A selectively accessible or increasing path is then a simple path from the root to a leaf along vertices with increasing labels. For the House of Cards model in the SSWM regime, we may assume that the root has the genotype of minimal fitness. This leads us precisely to the accessibility percolation model outlined above.

1.2 Other models

Our methods could be extended to consider, for example, Galton-Watson trees instead of n-ary trees.

Besides trees it is also natural to consider the House of Cards model on the n-dimensional hypercube $\{0,1\}^n$, for which there has been recent progress [2, 5]. A selectively accessible path in this setting is a path of minimal length on increasing labels from $(0,\dots,0)$ to $(1,\dots,1)$. Both papers consider the effect of varying the fitness at the zero vertex on the number of accessible paths. Hegarty and Martinsson [5] obtain the threshold for the phase transition of the existence of increasing paths as $n \to \infty$. Berestycki, Brunet and Shi [2] show that around this threshold, the number of such paths converges in distribution to the product of two independent exponential variables. As a first step, they obtain results for a particular rooted tree related to the hypercube.

Hegarty and Martinsson [5] also consider another model for the relationship between genotype and fitness, known as the *Rough Mount Fuji* model in the evolutionary biology literature [1], where a linear drift, depending on the distance to the root, is introduced to the random fitnesses. This model on n-ary trees was also considered in [9].

1.3 Notation

Throughout, we assume without loss of generality that the distribution of the labels is $U[0,1]$, and use the following crude double bound for Stirling's approximation valid for all $n \geq 1$:

$$2 < \frac{n!}{\sqrt{n}(n/e)^n} < 3.$$

We also assume that $n = \alpha h$, rather than use unwieldy $\lfloor \cdot \rfloor$ notation all the way through the article. Since there is a clear monotonicity in α in the model, no extra difficulty arises in considering cases when αh is not an integer.

Let P be the set of simple (that is, non-backtracking) paths from the root to a leaf in the tree; then $\#P = n^h$. For a path $u \in P$, write $X(u) = (X(u_1), \dots, X(u_h))$ for the (i.i.d., $U[0,1]$) labels on its vertices. For any two paths $u, v \in P$, let $a(u,v) = \max\{k : u_k = v_k\}$. Clearly $X(u_j) = X(v_j)$ for all $j \leq a(u,v)$.

Define

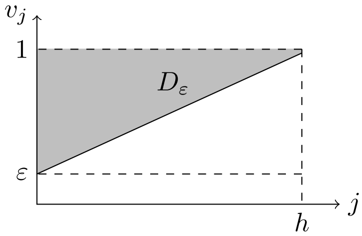
$$I = \left\{(x_1, \dots, x_h) \in [0,1]^h : x_1 < x_2 < \dots < x_h\right\},$$

and for $\varepsilon \in [0,1)$,

$$C_\varepsilon = \left\{(x_1, \dots, x_h) \in [0,1]^h : x_j \geq \varepsilon \ \forall j\right\}$$

and

$$D_\varepsilon = \left\{(x_1, \dots, x_h) \in [0,1]^h : x_j \geq \varepsilon + (1-\varepsilon)\left(\frac{j-1}{h}\right) \ \forall j\right\}.$$



Define

$$N_\varepsilon = \sum_{u \in P} \mathbb{1}_{\{X(u) \in I \cap D_\varepsilon\}},$$

and

$$N = \sum_{u \in P} \mathbb{1}_{\{X(u) \in I\}}.$$

1.4 Outline of proof

We will concentrate for the most part on proving Theorem 1, and then show how to adapt our proof to obtain Theorem 2.

We first observe that for a path u, $\{X(u) \in I\}$ is the event that h i.i.d. labels are in increasing order, which has probability $\frac{1}{h!}$. As $\#P = n^h$,

$$\mathbb{E}[N] = \frac{n^h}{h!}.$$

Using Stirling's approximation we have that

$$\mathbb{E}[N] \asymp \frac{n^h e^h}{\sqrt{h} h^h} = \frac{(\alpha e)^h}{\sqrt{h}}. \tag{1.1}$$

In particular, we see that for $\alpha \leq 1/e$, $\mathbb{E}[N] \to 0$ as $h \to \infty$, and recover the $\alpha \leq 1/e$ part of Theorem 1.1 via Markov's inequality.

Nowak and Krug [9] gave this argument, and then went on to give an upper bound on $\mathbb{E}[N^2]$, which they used to get a lower bound on the probability that $N \geq 1$. We take a similar but slightly more subtle route, in that we will work for the most part with N_ε, whose moments are slightly harder to estimate but give us more information. Of course we have

$$\mathbb{E}[N_\varepsilon] \leq \mathbb{E}[N] = \frac{n^h}{h!} \asymp \frac{(\alpha e)^h}{\sqrt{h}}.$$

In Section 2 we will show that

$$\mathbb{E}[N_\varepsilon] \geq \frac{(\alpha(1-\varepsilon)e)^h}{3h^{3/2}}$$

and in Section 3 we will see that when $\alpha(1-\varepsilon)e > 1$ and h is large,

$$\mathbb{E}[N_\varepsilon^2] \leq \mathbb{E}[N_\varepsilon] + \mathbb{E}[N_\varepsilon]^2 + c(\alpha(1-\varepsilon)e)^{2h}.$$

This will be enough to tell us that the probability that there is at least one path in N_ε is at least a constant times h^{-3} when h is large.

We then do a fairly standard trick to complete the proof of Theorem 1.1 in Section 4. We will show that there are many more than h^3 "good" subpaths in the first few levels of the tree: these are subpaths whose labels are increasing and small on the first few levels. Each of these subpaths then has a constant times h^{-3} probability of being the start of an increasing path to a leaf.

Finally in Section 5 we show how our techniques can be fine-tuned to give Theorem 1.2.

2 First moment bound

We aim to prove our lower bound on the first moment of N_ε:

Proposition 2.1.

$$\mathbb{E}[N_\varepsilon] \geq \frac{(\alpha(1-\varepsilon)e)^h}{3h^{3/2}}.$$

We shall need the following lemma.

Lemma 2.2. *Let $U_1, \ldots, U_j$ be i.i.d. $U[0,1]$ random variables. Then*

$$\mathbb{P}\left(U_1 \leq \ldots \leq U_j,\ U_1 \geq \frac{1}{j+1}, \ldots, U_j \geq \frac{j}{j+1}\right) = \frac{1}{(j+1)!}.$$

Proof. Let

$$p = \mathbb{P}\left(U_1 \leq \ldots \leq U_j,\ U_1 \geq \frac{1}{j+1}, \ldots, U_j \geq \frac{j}{j+1}\right)$$

and for each $i = 2, \ldots, j$, define

$$I_i = \int_{\frac{j}{j+1}}^{1} \int_{\frac{j-1}{j+1}}^{v_j} \cdots \int_{\frac{i}{j+1}}^{v_{i+1}} \left(\frac{v_i^{i-1}}{(i-1)!} - \frac{v_i^{i-2}}{(j+1)(i-2)!}\right) dv_i \ldots dv_j.$$

Note that

$$p = \int_{\frac{j}{j+1}}^{1} \int_{\frac{j-1}{j+1}}^{v_j} \cdots \int_{\frac{1}{j+1}}^{v_2} 1 \; dv_1 \ldots dv_j = \int_{\frac{j}{j+1}}^{1} \int_{\frac{j-1}{j+1}}^{v_j} \cdots \int_{\frac{2}{j+1}}^{v_3} \left(v_2 - \frac{1}{j+1} \right) \; dv_2 \ldots dv_j = I_2.$$

But for each $i = 2, \ldots, j-1$,

$$\begin{aligned} I_i &= \int_{\frac{j}{j+1}}^{1} \int_{\frac{j-1}{j+1}}^{v_j} \cdots \int_{\frac{i}{j+1}}^{v_{i+1}} \left(\frac{v_i^{i-1}}{(i-1)!} - \frac{v_i^{i-2}}{(j+1)(i-2)!} \right) \; dv_i \ldots dv_j \\ &= \int_{\frac{j}{j+1}}^{1} \int_{\frac{j-1}{j+1}}^{v_j} \cdots \int_{\frac{i+1}{j+1}}^{v_{i+2}} \left[\frac{v_i^i}{i!} - \frac{v_i^{i-1}}{(j+1)(i-1)!} \right]_{\frac{i}{j+1}}^{v_{i+1}} \; dv_{i+1} \ldots dv_j \\ &= \int_{\frac{j}{j+1}}^{1} \int_{\frac{j-1}{j+1}}^{v_j} \cdots \int_{\frac{i+1}{j+1}}^{v_{i+2}} \left(\frac{v_{i+1}^i}{i!} - \frac{v_{i+1}^{i-1}}{(j+1)(i-1)!} \right) \; dv_{i+1} \ldots dv_j = I_{i+1}. \end{aligned}$$

Therefore

$$\begin{aligned} p = I_2 = I_j = \int_{\frac{j}{j+1}}^{1} \left(\frac{v_j^{j-1}}{(j-1)!} - \frac{v_j^{j-2}}{(j+1)(j-2)!} \right) \; dv_j &= \left[\frac{v_j^j}{j!} - \frac{v_j^{j-1}}{(j+1)(j-1)!} \right]_{\frac{j}{j+1}}^{1} \\ &= \frac{1}{j!} - \frac{1}{(j+1)(j-1)!} = \frac{j+1-j}{(j+1)!} = \frac{1}{(j+1)!} \end{aligned}$$

as claimed. □

Proof of Proposition 2.1. By the fact that a $U[0,1]$ random variable conditioned to be at least ε is a $U[\varepsilon, 1]$ random variable,

$$\mathbb{E}[N_\varepsilon] = n^h \mathbb{P}(U \in I \cap D_\varepsilon) = n^h \mathbb{P}(U \in I \cap D_\varepsilon | U \in C_\varepsilon) \mathbb{P}(U \in C_\varepsilon) = (\alpha h(1-\varepsilon))^h \mathbb{P}(U \in I \cap D_0).$$

But by Lemma 2.2,

$$\mathbb{P}(U \in I \cap D_0) \geq \mathbb{P}(U_1 \leq 1/h) \mathbb{P}\left(U_2 < \ldots < U_h, \;\; U_i \geq \frac{i-1}{h} \;\; \forall i = 2, \ldots, h \right) = \frac{1}{h \cdot h!}.$$

Applying Stirling's approximation once more, we obtain

$$\mathbb{E}[N_\varepsilon] \geq \frac{(\alpha(1-\varepsilon)e)^h}{3h^{3/2}}.$$ □

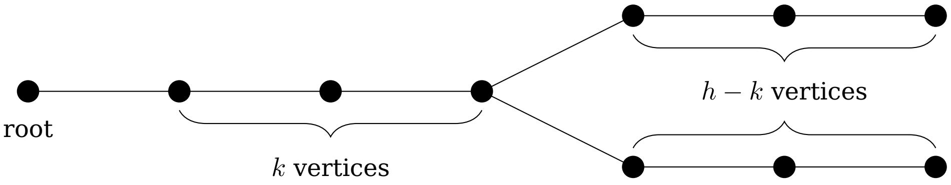
3 Second moment bound

We now aim to prove an upper bound on the second moment of N_ε:

Proposition 3.1. *If $\alpha(1-\varepsilon)e > 1$, then there exists some constant $c > 0$ such that*

$$\mathbb{E}[N_\varepsilon^2] \leq \mathbb{E}[N_\varepsilon] + \mathbb{E}[N_\varepsilon]^2 + c(\alpha(1-\varepsilon)e)^{2h}.$$

Proof. We break the second moment into a sum over *k-forks*:



To this end, for $k = 0, \dots, h$, let

$$N_\varepsilon^2(k) = \sum_{\substack{u,v \in P: \\ a(u,v)=k}} \mathbb{1}_{\{X(u),X(v) \in I \cap D_\varepsilon\}}.$$

Then

$$N_\varepsilon^2 = \sum_{k=0}^{h} N_\varepsilon^2(k).$$

Clearly $N_\varepsilon^2(h) = N_\varepsilon$, and $\mathbb{E}[N_\varepsilon^2(0)] = \mathbb{E}[N_\varepsilon]^2$.

Let $U = (U_1, \dots, U_h)$ and $V = (V_1, \dots, V_h)$ each be a sequence of i.i.d. $U[0,1]$ random variables such that $U_j = V_j$ for all $j \le k$ and U_j and V_j are independent for $j > k$. Using the fact that a uniform $[0,1]$ random variable conditioned to have value at least ε is a uniform $[\varepsilon, 1]$ random variable, we have for $k = 2, \dots, h-1$,

$$\begin{aligned}
\mathbb{E}[N_\varepsilon^2(k)] &= n^k \cdot n(n-1) \cdot n^{2h-2k-2} \cdot \mathbb{P}(U, V \in I \cap D_\varepsilon) \\
&= \left(\frac{n-1}{n}\right) n^{2h-k} \mathbb{P}(U, V \in I \cap D_\varepsilon | U, V \in C_\varepsilon) \mathbb{P}(U, V \in C_\varepsilon) \\
&= \left(\frac{n-1}{n}\right) (\alpha h)^{2h-k} (1-\varepsilon)^{2h-k} \mathbb{P}(U, V \in I \cap D_0).
\end{aligned}$$

Now,

$$\begin{aligned}
\mathbb{P}(U, V \in I \cap D_0) &= \int_{\frac{k-1}{h}}^{1} \mathbb{P}(U, V \in I \cap D_0 | U_k = x)\, dx \\
&\le \int_{\frac{k-1}{h}}^{1} \mathbb{P}(U_1 < U_2 < \dots < U_{k-1} < x) \mathbb{P}(x < U_{k+1} < U_{k+2} < \dots < U_h)^2\, dx \\
&= \int_{\frac{k-1}{h}}^{1} \frac{x^{k-1}}{(k-1)!} \cdot \frac{(1-x)^{2h-2k}}{(h-k)!^2}\, dx.
\end{aligned}$$

The curve $x^{k-1}(1-x)^{2h-2k}$ is decreasing on $x > (k-1)/(2h-k+1)$, so since $(k-1)/(2h-k+1) < (k-1)/h$,

$$\int_{\frac{k-1}{h}}^{1} \frac{x^{k-1}}{(k-1)!} \cdot \frac{(1-x)^{2h-2k}}{(h-k)!^2}\, dx \le \frac{((k-1)/h)^{k-1}}{(k-1)!} \cdot \frac{((h-k+1)/h)^{2h-2k}}{(h-k)!^2}.$$

Putting these estimates together and then applying Stirling's approximation, we obtain that for $k = 2, \dots, h-1$,

$$\begin{aligned}
\mathbb{E}[N_\varepsilon^2(k)] &\le (\alpha h(1-\varepsilon))^{2h-k} \cdot \frac{((k-1)/h)^{k-1}}{(k-1)!} \cdot \frac{((h-k+1)/h)^{2h-2k}}{(h-k)!^2} \\
&\le (\alpha(1-\varepsilon))^{2h-k} h \cdot \frac{e^{k-1}}{2(k-1)^{1/2}} \cdot \frac{e^{2h-2k+2}}{4(h-k+1)} \\
&= \frac{e}{8} \cdot \frac{(\alpha(1-\varepsilon)e)^{2h-k} h}{(k-1)^{1/2}(h-k+1)}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}[N_\varepsilon^2(1)] &\le n^{2h-1} \frac{(1-\varepsilon)^{2h-1}}{(h-1)!^2} \\
&\le \frac{e}{4} (\alpha(1-\varepsilon)e)^{2h-1}.
\end{aligned}$$

Thus if $\alpha(1-\varepsilon)e>1$, for some constant c,

$$\begin{aligned}\mathbb{E}[N_\varepsilon^2] &\le \mathbb{E}[N_\varepsilon]+\mathbb{E}[N_\varepsilon]^2+\frac{e}{4}(\alpha(1-\varepsilon)e)^{2h-1}+\sum_{k=2}^{h-1}\frac{e}{8}\cdot\frac{(\alpha(1-\varepsilon)e)^{2h-k}h}{(k-1)^{1/2}(h-k+1)}\\ &\le \mathbb{E}[N_\varepsilon]+\mathbb{E}[N_\varepsilon]^2+c(\alpha(1-\varepsilon)e)^{2h}. \qquad\square\end{aligned}$$

4 Proof of Theorem 1.1

As noted previously, it suffices to prove a lower bound when $\alpha>1/e$. Choose $\varepsilon\in(0,1)$ such that $\alpha(1-\varepsilon)e>1$. By the Paley-Zygmund inequality,

$$\mathbb{P}\left(N_\varepsilon\ge\frac{\mathbb{E}[N_\varepsilon]}{2}\right)\ge\frac{\mathbb{E}[N_\varepsilon]^2}{4\mathbb{E}[N_\varepsilon^2]}.$$

By Proposition 2.1, if we choose $\delta\in(0,\log(\alpha(1-\varepsilon)e))$ then $\mathbb{E}[N_\varepsilon]/2\ge e^{\delta h}$ for all large h, so

$$\mathbb{P}\left(N_\varepsilon>\exp(\delta h)\right)\ge\frac{\mathbb{E}[N_\varepsilon]^2}{4\mathbb{E}[N_\varepsilon^2]}.$$

But by Propositions 2.1 and 3.1, for large h and some constant c',

$$\mathbb{E}[N_\varepsilon^2]\le c'h^3\mathbb{E}[N_\varepsilon]^2.$$

Thus we get

$$\mathbb{P}(N_\varepsilon>\exp(\delta h))\ge\frac{1}{4c'h^3}. \tag{4.1}$$

Of course, we now want to improve this bound to get something exponentially close to 1 on the right-hand side. To do this, we will consider the first four levels of the tree separately from the rest. The idea is that with high probability, there are $\sim n^4$ paths from the root of length 4 whose labels are increasing and $<\varepsilon$. Each vertex at level 4 then has a subtree of $(h-4)^n$ paths of length $h-4$ and with probability $\gtrsim h^{-3}$ lots of these subpaths have labels which are increasing and $>\varepsilon$, by (4.1). So the probability that no path is increasing should look like, up to constants, $(1-h^{-3})^{n^4}$, which decays exponentially as desired. We note that our choice of four levels is only to counteract the factor of h^{-3} in (4.1) and working with any finite number of levels greater than 3 would also suffice.

We start by considering the subpaths v from the root to level 4. Although one can count subpaths whose labels satisfy $X(v_1)<\ldots<X(v_4)<\varepsilon$, we will instead count subpaths whose labels lie inside a priori intervals, allowing us to consider levels one at a time. More precisely, for $j\le4$, let M_j be the set of subpaths v from the root to level j such that $X(v_i)\in[(i-1)\varepsilon/4,i\varepsilon/4)$ for each $i=1,\ldots,j$. Observe that $\#M_1$ is the sum of n independent Bernoulli random variables of parameter $\varepsilon/4$; similarly, for $2\le j\le4$, given $\#M_{j-1}\ge k$, $\#M_j$ is at least a sum of kn independent Bernoulli random variables of parameter $\varepsilon/4$. For this reason, the following well-known form of the Chernoff bound will be useful.

Lemma 4.1. *Let $Z_1,\ldots,Z_r$ be independent Bernoulli random variables and let $Z=\sum_{i=1}^r Z_i$. Then*

$$\mathbb{P}\left(Z\le\frac{\mathbb{E}[Z]}{2}\right)\le\exp\left(-\frac{\mathbb{E}[Z]}{8}\right).$$

We can now prove our desired bound on $\#M_4$.

Lemma 4.2.

$$\mathbb{P}(\#M_4\le(n\varepsilon/8)^4)\le4\exp(-n\varepsilon^4/16384).$$

Proof. At level 1, there are n vertices, and $\mathbb{E}[\#M_1] = n\varepsilon/4$. Thus by Lemma 4.1,

$$\mathbb{P}\left(\#M_1 \le n\varepsilon/8\right) \le \exp\left(-\frac{n\varepsilon}{4\cdot 8}\right).$$

At level 2, given that $\#M_1 > n\varepsilon/8$, there are at least $n\varepsilon^2/8$ vertices whose parent had label in $[0,\varepsilon/4)$, and so

$$\mathbb{E}[\#M_2 \mid \#M_1 > n\varepsilon/8] \ge \frac{n^2\varepsilon^2}{8\cdot 4}.$$

Again by Lemma 4.1,

$$\mathbb{P}\left(\#M_2 \le (n\varepsilon/8)^2 \;\middle|\; \#M_1 > n\varepsilon/8\right) \le \exp\left(-\frac{(n\varepsilon/8)^2}{4}\right).$$

Similarly,

$$\mathbb{P}\left(\#M_3 \le (n\varepsilon/8)^3 \;\middle|\; \#M_2 > (n\varepsilon/8)^2\right) \le \exp\left(-\frac{(n\varepsilon/8)^3}{4}\right).$$

and

$$\mathbb{P}\left(\#M_4 \le (n\varepsilon/8)^4 \;\middle|\; \#M_3 > (n\varepsilon/8)^3\right) \le \exp\left(-\frac{(n\varepsilon/8)^4}{4}\right).$$

Summing these estimates gives the result. □

To complete the proof of Theorem 1.1, note that

$$\mathbb{P}(N \le \exp(\delta h)) \le \mathbb{P}(\#M_4 \le (n\varepsilon/8)^4) + \mathbb{P}(N \le \exp(\delta h),\ \#M_4 > (n\varepsilon/8)^4).$$

Suppose that $u \in M_4$, and consider the subtree of height $h-4$ rooted at the vertex u_4. In order that $N \le e^{\delta h}$, it must hold that there are no more than $e^{\delta h}$ paths in this subtree that have labels ordered and greater than ε. But we know from (4.1), since $n/(h-4) \ge n/h = \alpha$, that the probability of this event is at most $1 - c'h^{-3}$. Thus, applying also Lemma 4.2 and the inequality $1 + x \le e^x$,

$$\mathbb{P}(N \le \exp(\delta h)) \le 4\exp(-n\varepsilon^4/16384) + (1 - c'h^{-3})^{(n\varepsilon/8)^4} \le \exp(-\eta h)$$

for some $\eta > 0$, which proves Theorem 1.1.

5 Extension to $\alpha = 1/e + o(1)$: proof of Theorem 1.2

We now turn our attention to the case when $n = \alpha_h h$ where $\alpha_h = (1+\beta_h)/e$, $\beta_h \to 0$. For the first part of the theorem, it is not difficult to see from (1.1) and Markov's inequality that if $\log h - 2h\beta_h \to \infty$, then the probability that there exists an increasing path tends to 0 as $h \to \infty$. For the second part, choose ε_h such that $\varepsilon_h/\beta_h \to 0$ but $h\varepsilon_h/\log h \to \infty$ as $h \to \infty$. Then

$$\alpha_h(1-\varepsilon_h)e = (1+\beta_h)(1-\varepsilon_h) = 1 + \beta_h - \varepsilon_h - \beta_h\varepsilon_h.$$

So, for h sufficiently large, we have $\alpha_h(1-\varepsilon_h)e > 1$ and the proofs of Propositions 2.1 and 3.1 go through almost unchanged. As before we get

$$\mathbb{E}[N_{\varepsilon_h}] \ge \frac{(\alpha_h(1-\varepsilon_h)e)^h}{3h^{3/2}} \to \infty$$

and

$$\mathbb{E}[N_{\varepsilon_h}^2] \le \mathbb{E}[N_{\varepsilon_h}] + \mathbb{E}[N_{\varepsilon_h}]^2 + \frac{e}{4}(\alpha_h(1-\varepsilon_h)e)^{2h-1} + \sum_{k=2}^{h-1} \frac{e}{8}\cdot\frac{(\alpha_h(1-\varepsilon_h)e)^{2h-k}h}{(k-1)^{1/2}(h-k+1)}.$$

However, since $(\alpha_h(1-\varepsilon_h)e)$ is not constant, we cannot bound the last term, up to constants, by $(\alpha_h(1-\varepsilon_h)e)^{2h}$ as before. Instead, we see that

$$\sum_{k=2}^{h-1} \frac{1}{(k-1)^{1/2}(h-k+1)} < \int_0^{h-1} \frac{1}{\sqrt{x}(h-x)}\,dx = \frac{2}{\sqrt{h}} \tanh^{-1}\left(\sqrt{\frac{h-1}{h}}\right) \sim \frac{\log h}{\sqrt{h}},$$

which leads to a bound, up to constants, of $h^{1/2}\log h(\alpha_h(1-\varepsilon_h)e)^{2h}$. Therefore,

$$\mathbb{E}[N_{\varepsilon_h}^2] \le \mathbb{E}[N_{\varepsilon_h}] + \mathbb{E}[N_{\varepsilon_h}]^2 + ch^{1/2}\log h(\alpha_h(1-\varepsilon_h)e)^{2h} \le c'h^{15/4}\mathbb{E}[N_{\varepsilon_h}]^2$$

for some constants c and c'.

Now, the main difficulty arises in our application of Lemma 4.2. With the new exponent of $15/4$, applying our previous argument, we get

$$\mathbb{P}(N \le \exp(\delta h),\ \#M_4 > (n\varepsilon_h/8)^4) \le (1 - c'h^{-15/4})^{(n\varepsilon_h/8)^4} \le \exp(-c''h^{1/4}\varepsilon_h^4)$$

for some constant c''. However, unlike before, this probability does not converge to 0 as $h \to \infty$, because $h^{1/4}\varepsilon_h^4$ does not necessarily converge to infinity. So instead of working with a fixed number of levels, we work with $\log h$ levels. More precisely, we work with $\lfloor \log h \rfloor$ levels, but to save notation we simply write $\log h$.

Now, for $j = 1, \ldots, \log h$, we define $\widetilde{M}_j$ to be the set of subpaths v from the root to level j such that $X(v_i) \in [(i-1)\varepsilon_h/\log h, i\varepsilon_h/\log h)$ for each $i = 1, \ldots, j$. As before, by repeatedly applying Lemma 4.1 we get

$$\mathbb{P}\left(\#\widetilde{M}_j \le \left(\frac{n\varepsilon_h}{2\log h}\right)^j \,\middle|\, \#\widetilde{M}_{j-1} > \left(\frac{n\varepsilon_h}{2\log h}\right)^{j-1}\right) \le \exp\left(-\frac{1}{4}\left(\frac{n\varepsilon_h}{2\log h}\right)^j\right),$$

for $j = 1, \ldots, \log h$. Summing these bounds gives

$$\mathbb{P}\left(\#\widetilde{M}_{\log h} \le \left(\frac{n\varepsilon_h}{2\log h}\right)^{\log h}\right) \le \sum_{j=1}^{\log h} \exp\left(-\frac{1}{4}\left(\frac{n\varepsilon_h}{2\log h}\right)^j\right)$$

which converges to 0 as $h \to \infty$ by our assumption that $h\varepsilon_h/\log h \to \infty$. Therefore, with high probability, we have at least $(n\varepsilon_h/2\log h)^{\log h} = h^{\log(n\varepsilon_h/2\log h)}$ "good" increasing subpaths up to level $\log h$, each of which has probability at least $c'h^{-15/4}$ of extending to an increasing path to a leaf. Since

$$(1 - c'h^{-15/4})^{(n\varepsilon_h/2\log h)^{\log h}} \le \exp\left(-c'h^{-15/4+\log(n\varepsilon_h/2\log h)}\right) \to 0,$$

the probability that there is no increasing path from the root to a leaf converges to 0 as $h \to \infty$, completing the proof of Theorem 1.2.

References

[1] T. Aita, H. Uchiyama, T. Inaoka, M. Nakajima, T. Kokubo and Y. Husimi (2000). Analysis of a local fitness landscape with a model of the rough Mt. Fuji-type landscape: application to prolyl endopeptidase and thermolysin. *Biopolymers*, **54**, 64–79.

[2] J. Berestycki, É. Brunet and Z. Shi (2013+). How many evolutionary histories only increase fitness? Preprint, arXiv:1304.0246.

[3] J. Franke, A. Klözer, J. A. G. M. de Visser and J. Krug (2011). Evolutionary accessibility of mutational pathways. *PLoS Comput. Biol.*, **7**, e1002134. MR-2845072

[4] J. H. Gillespie (1983). Some properties of finite populations experiencing strong selection and weak mutation. *Amer. Natur.*, **121**, 691–708.

[5] P. Hegarty and A. Martinsson (2013+). On the existence of accessible paths in various models of fitness landscapes. arXiv:1210.4798.

[6] M. Kimura (1962). On the probability of fixation of mutant genes in a population. *Genetics*, **47**, 713–719.

[7] J. F. C. Kingman (1978). A simple model for the balance between selection and mutation. *J. Appl. Probab.*, **15**, 1–12. MR-0465272

[8] S. Kauffman and S. Levin (1987). Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.*, **128**, 11–45. MR-0907587

[9] S. Nowak and J. Krug (2013). Accessibility percolation on n-trees. *Europhys. Lett.*, **101**, 66004.

[10] H. A. Orr (2002). The population genetics of adaptation: The adaptation of DNA sequences. *Evolution*, **56**, 1317–1330.

[11] D. M. Weinreich, N. F. Delaney, M. A. DePristo and D. M. Hartl (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, **312**, 111–114.

[12] D. M. Weinreich, R. A. Watson and L. Chao (2005). Perspective: Sign epistasis and genetic constraints on evolutionary trajectories. *Evolution*, **59**, 1165–1174.

Acknowledgments. The first author is partly supported by EPSRC grant EP/K007440/1. This collaboration began during the 2013 UK probability meeting in Cambridge supported by EPSRC grant EP/I03372X/1 and continued at the Mathematisches Forschungsinstitut Oberwolfach supported by the ESF and Oberwolfach Leibniz Graduate Student programme. Both authors thank two anonymous referees for their helpful comments.