Increasing paths in trees
Matthew I. Roberts ∗ Lee Zhuo Zhao †
May 6, 2013

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 simple 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 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 $\alpha \leq 1/e$.

1 Introduction

Consider a regular $n$-ary tree of height $h$, where $n = \alpha h$. To each vertex except the root attach an independent and identically distributed continuous random variable. We are concerned with whether there is a simple (that is, non-backtracking) path from the root to a leaf whose labels only increase. Nowak and Krug [7] called this accessibility percolation and showed that $P(\text{there exists an increasing path}) \to 0$ as $n \to \infty$ if $\alpha < 1/e$, whereas if $\alpha > 1$ then there exists some $p > 0$ depending on $\alpha$ such that $P(\text{there exists an increasing path}) > p$. We give a complete characterisation in terms of $\alpha$, showing that there is a phase transition at $\alpha = 1/e$.

Theorem 1. For $\alpha > 1/e$, $P(\text{there exists an increasing path}) \to 1$ as $h \to \infty$.

In fact we will show that for any $\alpha > 1/e$, there exist $\delta > 0$ and $\eta > 0$ such that

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

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 genotype or disappear from the population. If the rate of selection is stronger than the rate of mutation, only mutations which give rise to a fitter genotype survive. Therefore, 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 [6, 8, 9].

∗Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK.
Email: mattiroberts@gmail.com.
†Statistical Laboratory, University of Cambridge, Cambridge, CB3 0WB, UK.
Email: lzz20@statslab.cam.ac.uk
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 [4, 5], 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 initially 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, 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. We might also be able to fine-tune our methods to gain information about the finer behaviour near the critical point $\alpha = 1/e$, but this would be highly technical work and seems unlikely to offer further insight into the model.

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 [2,3]. A selectively accessible path in this setting is a path of minimal length on increasing labels from $(0, \ldots, 0)$ to $(1, \ldots, 1)$. Both papers consider the effect of varying the fitness at the zero vertex on the number of accessible paths. Hegarty and Martinsson obtain the threshold for the phase transition of the existence of increasing paths as $n \to \infty$.

Berestycki, Brunet and Shi 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 also consider another model for the relationship between genotype and fitness, known as the Rough Mount Fuji model in the 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 [7].

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 $n \geq 1$,

$$2 < \frac{n!}{\sqrt{n(n/e)^n}} < 3.$$
and
\[ D_\varepsilon = \left\{ (x_1, \ldots, x_h) \in [0,1]^h : x_j \geq \varepsilon + (1 - \varepsilon) \left( \frac{j - 1}{h} \right) \forall j \right\}. \]

Define
\[ N_\varepsilon = \sum_{u \in P} 1_{\{X(u) \in I \cap D_\varepsilon\}}, \]
and
\[ N = \sum_{u \in P} 1_{\{X(u) \in I\}}. \]

It is clear that \( N_\varepsilon \leq N \). We will attempt to show that \( \mathbb{P}(N_\varepsilon \geq 1) \) is not too small when \( \alpha > \frac{1}{\varepsilon} \) and \( \alpha(1 - \varepsilon) > 1 \).

## 2 Second moment bound

We break the second moment into a sum over \( k \)-forks:

To this end, for \( k = 0, \ldots, h \), let
\[ N_\varepsilon^2(k) = \sum_{u,v \in P: a(u,v)=k} 1_{\{X(u),X(v) \in I \cap D_\varepsilon\}}. \]

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 \( U = (U_1, \ldots, U_h) \) and \( V = (V_1, \ldots, V_h) \) each be a sequence of i.i.d. \( U[0,1] \) random variables such that \( U_j = V_j \) for all \( j \leq k \) 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 \( \varepsilon \) is a uniform
[\varepsilon, 1] \text{ random variable, we have for } k = 2, \ldots, h - 1,
\begin{align*}
\mathbb{E}[N^2(\alpha)] &= n^k \cdot n(n-1) \cdot n^{2h-2k} \cdot \mathbb{P}(U, V \in I \cap D_x) \\
&= \left(\frac{n-1}{n}\right) n^{2h-k} \mathbb{P}(U, V \in I \cap D_x | U, V \in C_x) \mathbb{P}(U, V \in C_x) \\
&= \left(\frac{n-1}{n}\right) (\alpha h)^{2h-k}(1-\varepsilon)^{2h-k} \mathbb{P}(U, V \in I \cap D_0).
\end{align*}

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

The curve \( x^{k-1}(1-x)^{2h-2k} \) is decreasing on \( x > (k-1)/(2h-k+1) \), so since \( (k-1)/(2h-k+1) < (k-1)/h \),
\begin{align*}
\int_{k-1}^{1} \frac{x^{k-1}}{(k-1)!} \cdot \frac{(1-x)^{2h-2k}}{(h-k)!^2} \, dx &\leq \frac{((k-1)/h)^{k-1}}{(k-1)!} \cdot \frac{((h-k+1)/h)^{2h-2k}}{(h-k)!^2}.
\end{align*}

Putting these estimates together and then applying Stirling’s approximation, we obtain that for \( k = 2, \ldots, h - 1 \),
\begin{align*}
\mathbb{E}[N^2(\alpha)] &\leq (\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} \\
&\leq (\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{\varepsilon}{8} \cdot \frac{(\alpha(1-\varepsilon)e)^{2h-k}h}{(k-1)^{1/2}(h-k+1)}
\end{align*}

Similarly,
\begin{align*}
\mathbb{E}[N^2(1)] &\leq n^{2h-1} \frac{(1-\varepsilon)^{2h-1}}{(h-1)!^2} \\
&\leq \frac{\varepsilon}{4} (\alpha(1-\varepsilon)e)^{2h-1}.
\end{align*}

Thus if \( \alpha(1-\varepsilon)e > 1 \), then for some constant \( c \),
\begin{align*}
\mathbb{E}[N^2] &\leq \mathbb{E}[N_\varepsilon] + \mathbb{E}[N_\varepsilon]^2 + \frac{\varepsilon}{4} (\alpha(1-\varepsilon)e)^{2h-1} + \sum_{k=2}^{h-1} \frac{\varepsilon}{8} \cdot \frac{(\alpha(1-\varepsilon)e)^{2h-k}h}{(k-1)^{1/2}(h-k+1)} \\
&\leq \mathbb{E}[N_\varepsilon] + \mathbb{E}[N_\varepsilon]^2 + c(\alpha(1-\varepsilon)e)^{2h}.
\end{align*}
3 First moment bound

To bound $E[N_r]$, we shall need the following lemma.

**Lemma 2.** Let $U_1, \ldots, U_j$ be i.i.d. $U[0, 1]$ random variables. Then

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

**Proof.** Let

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

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

$$I_i = \int_{1/j+1}^{v_i} \int_{1/j+1}^{v_{i-1}} \ldots \int_{1/j+1}^{v_1} (i-1)! - \frac{v_i^{i-2}}{(j+1)(i-2)!} dv_i \ldots dv_j.$$

Note that

$$p = \int_{1/j+1}^{v_j} \int_{1/j+1}^{v_{j-1}} \ldots \int_{1/j+1}^{v_2} 1 dv_1 \ldots dv_j = \int_{1/j+1}^{v_j} \int_{1/j+1}^{v_{j-1}} \ldots \int_{1/j+1}^{v_3} v_2 - \frac{1}{j+1} dv_2 \ldots dv_j = I_2.$$

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

$$I_i = \int_{1/j+1}^{v_i} \int_{1/j+1}^{v_{i-1}} \ldots \int_{1/j+1}^{v_1} \frac{v_i^{i-1}}{(i-1)!} - \frac{v_{i+1}^{i-2}}{(j+1)(i-2)!} dv_i \ldots dv_j$$

$$= \int_{1/j+1}^{v_i} \int_{1/j+1}^{v_{i+2}} \ldots \int_{1/j+1}^{v_1} \frac{v_i}{i!} - \frac{v_{i+1}^{i-1}}{(j+1)(i-1)!} dv_{i+1} \ldots dv_j$$

$$= \int_{1/j+1}^{v_i} \int_{1/j+1}^{v_{i+2}} \ldots \int_{1/j+1}^{v_1} \frac{v_i^{i-1}}{i!} - \frac{v_{i+1}^{i-1}}{(j+1)(i-1)!} dv_{i+1} \ldots dv_j = I_{i+1}.$$

Therefore

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

$$= \frac{1}{j!} - \frac{1}{(j+1)(j-1)!} = \frac{j+1-j}{(j+1)!} = \frac{1}{(j+1)!}$$

as claimed. \qed

Now,

$$E[N_r] = n^h P(U \in I \cap D_c) = n^h P(U \in I \cap D_c | U \in C_c) P(U \in C_c) = (\alpha h (1-\varepsilon))^{h} P(U \in I \cap D_0).$$

But by Lemma 2

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

Applying Stirling’s approximation once more, we obtain

$$E[N_r] \geq \frac{(\alpha (1-\varepsilon) e)^h}{3h^{3/2}}.$$
4 Proof of Theorem 1

By the Paley-Zygmund inequality,
\[ \mathbb{P} \left( N_\varepsilon \geq \frac{\mathbb{E}[N_\varepsilon]}{2} \right) \geq \frac{\mathbb{E}[N_\varepsilon]^2}{4\mathbb{E}[N_\varepsilon]}. \]

From our bounds on the first and second moments of \( N_\varepsilon \), we get for some \( \delta > 0 \),
\[ \mathbb{P}(N_\varepsilon \geq \exp(\delta h)) \geq c'h^{-3} \tag{1} \]
for some constant \( c' > 0 \).

To complete the proof, we will consider the first four levels of the tree separately from the rest. We require the following form of Hoeffding’s inequality.

**Lemma 3.** Suppose that \( \varepsilon \in (0, 1) \) and \( 0 \leq a \leq 1 - \varepsilon/4 \). If \( A_1, \ldots, A_r \) are i.i.d. \( U[0, 1] \) random variables, then
\[ \mathbb{P}(\# \{j : A_j \in [a, a + \varepsilon/4]\} \leq r\varepsilon/8) \leq \exp(-\varepsilon^2 r/16). \]

For \( j = 1, 2, 3, 4 \), let \( M_j \) be the set of vertices \( v \) at the \( j \)th level of the tree such that \( v_i \in [(i - 1)\varepsilon/4, i\varepsilon/4) \) for each \( i = 1, \ldots, j \).

**Lemma 4.**
\[ \mathbb{P}(\#M_4 \leq n\varepsilon^4/8^4) \leq 4\exp(-\varepsilon^5 n/8192). \]

**Proof.** At level 1, there are \( n \) vertices; thus by Lemma 3, the probability that fewer than \( n\varepsilon/8 \) of them have label in \( [0, \varepsilon/4) \) is at most \( \exp(-\varepsilon^2 n/16) \). That is,
\[ \mathbb{P}(\#M_1 \leq n\varepsilon/8) \leq \exp(-\varepsilon^2 n/16). \]

At level 2, given that \( \#M_1 > n\varepsilon/8 \), there are at least \( n^2\varepsilon/8 \) vertices whose parent had label in \( [0, \varepsilon/4) \), and the probability that fewer than \( n^2\varepsilon^2/8^2 \) of these have label in \( [\varepsilon/4, \varepsilon/2) \) is (again by Lemma 3) at most \( \exp(-\varepsilon^3 n^2/2 \cdot 8^2) \). That is,
\[ \mathbb{P}(\#M_2 \leq n^2\varepsilon^2/8^2 \mid \#M_1 > n\varepsilon/8) \leq \exp(-\varepsilon^3 n^2/2 \cdot 8^2). \]

Similarly,
\[ \mathbb{P}(\#M_3 \leq n^3\varepsilon^3/8^3 \mid \#M_2 > n^2\varepsilon^2/8^2) \leq \exp(-\varepsilon^4 n^3/2 \cdot 8^3) \]
and
\[ \mathbb{P}(\#M_4 \leq n^4\varepsilon^4/8^4 \mid \#M_3 > n^3\varepsilon^3/8^3) \leq \exp(-\varepsilon^5 n^4/2 \cdot 8^4). \]

Summing these estimates gives the result. \( \square \)

To complete the proof of Theorem 1 note that
\[ \mathbb{P}(N \leq \exp(\delta h)) \leq \mathbb{P}(\#M_4 \leq n^4\varepsilon^4/8^4) + \mathbb{P}(N \leq \exp(\delta h), \#M_4 > n^4\varepsilon^4/8^4). \]

Suppose that \( u \in M_4 \), and consider the subtree of height \( h - 4 \) rooted at the vertex \( u_4 \). In order that \( N \leq \varepsilon^4 \), it must hold that there are no more than \( \varepsilon^4 h \) paths in this subtree that have labels ordered and greater than \( \varepsilon \). But we know from (1), since \( n/(h - 4) \geq n/h = \alpha \), that the probability of this event is at most \( 1 - c'h^{-3} \). Thus, applying also Lemma 1
\[ \mathbb{P}(N \leq \exp(\delta h)) \leq 4\exp(-\varepsilon^5 n/8192) + (1 - c'h^{-3}) n^4\varepsilon^4/8^4 \leq \exp(-\eta h) \]
for some \( \eta > 0 \), which proves Theorem 1.
Acknowledgements. The first author is partly supported by EPSRC grant EP/K007440/1. This collaboration began during a 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.

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, 6479

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

[3] P. Hegarty and A. Martinsson (2013+). On the existence of accessible paths in various models of fitness landscapes. Preprint, [arXiv:1210.4798](https://arxiv.org/abs/1210.4798).

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

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

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

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

[8] 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.

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