A Metropolis-type Optimization Algorithm on the Infinite Tree

David Aldous *
Department of Statistics
University of California
Berkeley CA 94720
http://www.stat.berkeley.edu/users/aldous

March 21, 1997

Abstract

Let $S(v)$ be a function defined on the vertices v of the infinite binary tree. One algorithm to seek large positive values of S is the Metropolis-type Markov chain (X_n) defined by

$$P(X_{n+1} = w | X_n = v) = \frac{1}{3} \frac{e^{b(S(w)-S(v))}}{1 + e^{b(S(w)-S(v))}}$$

for each neighbor w of v, where b is a parameter ("1/temperature") which the user can choose. We introduce and motivate study of this algorithm under a probability model for the objective function S, in which S is "tree-indexed simple random walk", that is the increments $\xi(e) = S(w) - S(v)$ along parent-child edges $e = (v, w)$ are independent and $P(\xi = 1) = p$, $P(\xi = -1) = 1 - p$. This algorithm has a "speed" $r(p, b) = \lim_n n^{-1} ES(X_n)$. We study the speed via a mixture of rigorous arguments, non-rigorous arguments and Monte Carlo simulations, and compare with a deterministic greedy algorithm which permits rigorous analysis. Formalizing the non-rigorous arguments presents a challenging problem. Mathematically, the subject is in part analogous to recent work of Lyons-Pemantle-Peres (1995,1996) on the speed on random walk on Galton-Watson trees. A key feature of the model is existence of a critical point p_{crit} below which the problem is infeasible; we study behavior of algorithms as $p \downarrow p_{\text{crit}}$.

Preliminary version. See homepage for updates.

*Research supported by N.S.F. Grant DMS96-22859

1 Introduction

Section 1 describes the model, results and conjectures. Discussion of background algorithmic issues and of background statistical physics methods is in sections 2.1 and 5.1.

Figure 1 illustrates a function $S(v)$ defined on the vertices v of the infinite binary tree $\mathcal{T}^\infty$. Our convention is to make the root have degree 3 (rather than 2, the convention in the theory of algorithms), but the choice of convention is not important. Note that $S(v)$ takes both positive and negative values. Note also that specifying a function $S(v)$ with $S(\text{root}) = 0$ is equivalent to specifying a function $\xi(v,w)$ on directed edges (v,w) satisfying $\xi(w,v) = -\xi(v,w)$, the equivalence being via

$$S(w) - S(v) = \xi(v,w) \text{ for all edges } (v,w).$$

We study the (imprecise) question

> What is a good algorithm for finding large positive values of $S(\cdot)$?

One can invent many algorithms, but the following two seem fundamental.

The greedy algorithm. Suppose we have examined vertices root $= v_0, v_1, \ldots, v_n$. Consider the subset of those vertices which have some child which has not been examined; from that subset, choose a vertex v for which $S(v)$ is maximal and then choose some previously-unexamined child of v to be the next vertex v_{n+1} to be examined.

The Metropolis algorithm. Fix a parameter $b \geq 0$. Let the sequence of (not distinct) vertices examined be the Markov chain (X_n) with $X_0 = \text{root}$ and with transition probabilities

$$P(X_{n+1} = w | X_n = v) = \frac{1}{3} \frac{e^{b(S(w)-S(v))}}{1 + e^{b(S(w)-S(v))}} \tag{1}$$

for each neighbor w of v.

Our term "Metropolis algorithm" in the present context is rather non-standard, because the term is properly used in the context of simulating a stationary distribution $\pi(x) \propto \exp(bS(x))$ on a finite set. A connection between the finite and infinite settings will be discussed in section 2, but for the moment keep in mind the idea that studying *transient* (in the Markov chain sense) behavior of the infinite-state chain is intended as a toy model for the pre-equilibrium (*transient*, in the engineer's sense) behavior of randomized optimization algorithms on large finite sets.

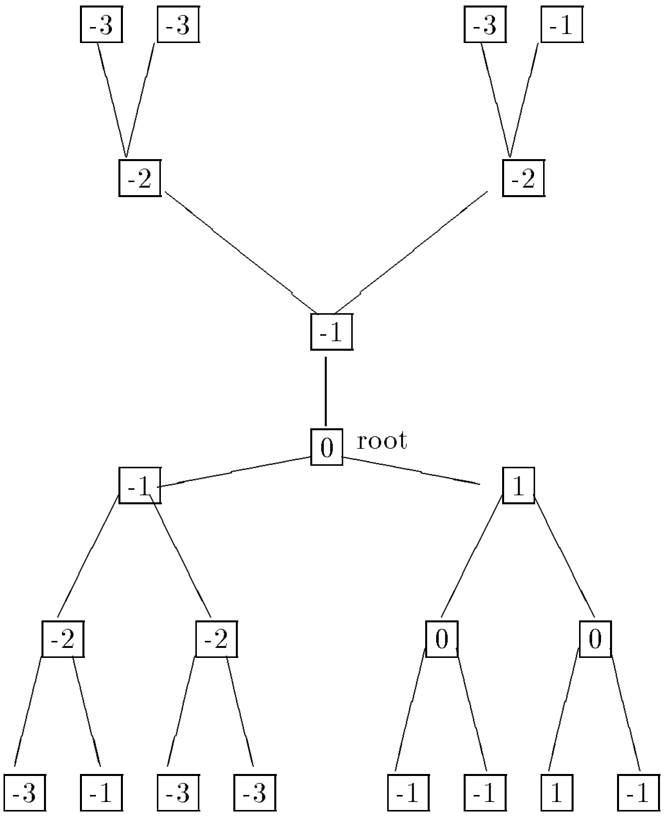

figure 1

We called our original question "imprecise". The algorithms themselves are precisely specified (up to arbitrary tie-breaking and selection of children conventions in the greedy algorithm, where the exact conventions are unimportant). What is imprecise is the criterion for an algorithm to be "good". Without assumptions on the objective function S there seems no hope of comparing algorithms. So we introduce the simplest possible probability model, in which the increments $\xi(v,w)$ across edges (v,w) directed away from the root are modeled as independent random variables with common distribution ξ such that

$$P(\xi = 1) = p, \qquad P(\xi = -1) = 1 - p. \tag{2}$$

Here p is a parameter outside our control. For $p > 1/2$ the "oblivious" algorithm which simply follows a prescribed ray in the tree will find large positive values of S, so we restrict to the case $p < 1/2$. Much of our analysis should extend to the setting where the increment distributions ξ have more general distributions depending on a parameter p. But since the ± 1-valued case already presents enough difficulty, we stick with the special setting, with occasional remarks about "the general ξ setting". We call this model *tree-indexed random walk*: see Pemantle [20] for a survey of theoretical probability results about such processes.

The "stationarity" inherent in this tree-indexed random walk model suggests that for any reasonable algorithm ALG which generates a deterministic or random sequence of vertices (root $= X_0, X_1, X_2, \ldots$) to be examined, there should be some asymptotic "rate" $r(p, \mathrm{ALG})$ such that

$$\lim_{n \to \infty} n^{-1} S(X_n) = r(p, \mathrm{ALG}) \text{ a.s.} \tag{3}$$

Note that $r(p, \mathrm{ALG})$ may be positive, negative or zero. Of course from the algorithmic viewpoint we are interested in the *maximum* value found, but (3) immediately implies

$$\lim_{n \to \infty} n^{-1} \max_{m \leq n} S(X_m) = \max(0, r(p, \mathrm{ALG})) \text{ a.s.}$$

and so it is enough to consider $r(p, \mathrm{ALG})$.

Standard results on branching random walk (see Biggins [2] for a recent treatment) imply a rather different linear rate result:

$$\lim_{d \to \infty} d^{-1} \max_{v \in H(d)} S(v) = a(p) \text{ a.s., and there exists a path}$$

$$\text{root}, v_1, v_2, \ldots \text{ such that } \lim_{d \to \infty} d^{-1} S(v_d) = a(p) \text{ a.s.} \tag{4}$$

where $H(d)$ is the set of $3 \cdot 2^{d-1}$ vertices at depth d, and where $a(p)$ is the solution of

$$\inf_{\theta} \left(pe^{\theta} + (1-p)e^{-\theta} - \tfrac{1}{2} e^{\theta a(p)} \right) = 0.$$

And there is a critical value

$$p_{\mathrm{crit}} = \tfrac{1 - \sqrt{3/4}}{2} \approx 0.06699$$

such that $a(p) < 0$ for $p < p_{\mathrm{crit}}$ and $a(p) > 0$ for $p > p_{\mathrm{crit}}$. This result has an immediate negative implication: if $p < p_{\mathrm{crit}}$ then there are only finitely many vertices v with $S(v) > 0$ and so no algorithm can have $r(p, \mathrm{ALG}) > 0$. The converse, that for $p > p_{\mathrm{crit}}$ there exists some algorithm with $r(p, \mathrm{ALG}) > 0$, is not obvious from the *statement* of (4) but, as explained in [1], does indeed follow fairly easily from the *proof* of (4).

Thus our setting provides a toy model in which to study and compare general purpose optimization algorithms: one can ask whether a specific algorithm ALG satisfies

$$r(p, \mathrm{ALG}) > 0 \text{ for all } p > p_{\mathrm{crit}}$$

and one can compare algorithms by comparing the values of $r(\cdot)$. It was shown in Aldous [1] that $r(p, \mathrm{GREEDY})$ can be expressed in terms of the solution of a fixed-point identity for distributions and thereby computed numerically – see figure 3 later. A minor purpose of this paper is to derive (rigorously) the asymptotic behavior around the critical point.

Proposition 1 $r(p, \mathrm{GREEDY}) = \exp\left(-c(p - p_{\mathrm{crit}})^{-1/2} + O(1)\right)$ *as* $p \downarrow p_{\mathrm{crit}}$, *where* $c \approx 1.11$ *is given by the explicit formula (13).*

The proof comprises section 3.

The major purpose of this paper is to initiate study of the Metropolis algorithm (1) in our tree-indexed random walk model. Recall that the Metropolis algorithm involves a parameter b (= "1/temperature"), and as at (3) write $r(p, b)$ for the rate

$$\lim_{n \to \infty} n^{-1} S(X_n) = r(p, b) \text{ a.s.}$$

associated with the Metropolis algorithm. We do not know any theoretical result which enables $r(p, b)$ to be calculated numerically. Figure 2 shows values of $r(0.2, b)$ for $0 \leq b \leq 3$ obtained by Monte Carlo simulation. Note that $p = 0.2$ is about 3 times p_{crit}. The qualitative shape of the function $b \to r(p, b)$ seen in simulations is similar for all $p_{\mathrm{crit}} < p < 1/2$: the function starts negative at $b = 0$, increases until reaching a positive maximum, then decreases to 0 as $b \to \infty$. The maximum rate

$$r^*(p) \equiv \max_b r(p, b) \tag{5}$$

is about 10^{-3} for $p = 0.2$ – by comparison, $r(0.2, \mathrm{GREEDY}) \approx 0.044$ is much larger. Figure 3 compares $r(p, \mathrm{GREEDY})$ (calculated numerically) with the maximal Metropolis rate $r^*(p)$ (found by simulation).

Can we give theoretical arguments which explain the simulation results? As to figure 3, the answer is "no". The simulations show a remarkably good fit to the curve

$$r^*(p) = \text{constant} \times (r(p, \text{GREEDY}))^2$$

and so for the record we make

Conjecture 2 *As $p \downarrow p_{\text{crit}}$,*

$$r^*(p) = (r(p, \text{GREEDY}))^{2+o(1)}.$$

Proving this seems well beyond the reach of current mathematics, though in section 4 we point out the same "squared" relationship in an analogous context. What is within reach? Mathematically, our Metropolis chain is analogous to the well-studied topic RWIRE (Random Walk in Random Environment) [7] and in particular to recent work of Lyons-Pemantle-Peres [16, 17] studying speed of random walk on Galton-Watson trees. Applying known methods gives

Theorem 3 *Fix $1/2 > p > p_{\text{crit}}$ and $b \geq 0$. Write $b_0(p) = \log \frac{1-p}{p} > 0$. Then the limit rate*

$$\lim_{n\to\infty} n^{-1} S(X_n) = r(p,b) \ a.s. \tag{6}$$

exists, and

(A) $r(p,0) = \frac{2p-1}{6} < 0$.
(B) $r(p, b_0(p)) = 0$.
(C) For $b = b_0(p)$ there is a variance rate

$$\sigma^2(p, b_0(p)) = \lim_t t^{-1} \text{var } S(X_t) \geq 0.$$

(D) $r(p,b) \leq 0$ for $0 \leq b \leq b_0(p)$ and $r(p,b) \geq 0$ for $b_0(p) \leq b < \infty$.
(E) $r(p,b) \to 0$ as $b \to \infty$.

To avoid technicalities, we shall give (section 5.2) only an informal treatment of Theorem 3 emphasizing calculations, but there would be no real difficulty in rephrasing our arguments rigorously. In contrast, in section 5.3 we give calculations (which do seem difficult to make rigorous) for (F,G) below, and reasons to believe (H).

Conjecture 4 *In the setting of Theorem 3, write* $r_b(p,b) = \frac{d}{db} r(p,b)$*. Then*
(F) $r_b(p,0) = \frac{3-8p+8p^2}{12} > 0$.
(G) $r_b(p,b_0(p)) = \frac{1}{2}\sigma^2(p,b_0(p))$.
(H) $\sigma^2(p,b_0(p)) > 0$.

The point is that (G,H) would imply

$$r^*(p) \equiv \sup_b r(p,b) > 0, \text{ for each } p > p_{\mathrm{crit}}. \tag{7}$$

See section 2.1 for interpretation.

The remaining sections of the paper are largely independent of each other. Section 2 elaborates the conceptual connection between our setup and randomized optimization over a large finite set. Section 3 gives the proof of Proposition 1. Section 4 uses a formula of Lyons-Pemantle-Peres [16] to exhibit a "squared" relationship (analogous to Conjecture 2) in the context on random walk on near-critical Galton-Watson trees. Section 5 outlines the proof of Theorem 3 and arguments in support of Conjecture 4.

2 Finite optimization algorithms

2.1 General remarks

The *universality paradigm* in statistical physics asserts that, in a system with a phase transition at a critical value θ_{crit} of a parameter θ, one expects statistics of the system to scale as $(\theta - \theta_{\mathrm{crit}})^\alpha$ near the critical point, where the *scaling exponent* α depends on the statistic but not on the details of the model. Our motivation for this paper was to investigate whether analogs exist in the context of randomized optimization algorithms. A proof of Conjecture 2 would be an appealing starting point for such an area of research.

Diaconis and Saloff-Coste [4] survey what is rigorously known about Metropolis algorithms, from the usual viewpoint of sampling from a given distribution rather than our viewpoint as a randomized optimization algorithm. As noted by Jerrum and Sorkin [9], despite the intuitively appealing story behind simulated annealing, there is no known interesting example where it can be proved that varying the temperature parameter improves performance of the Metropolis optimization algorithm. And as noted by Juels [10], there is no known interesting example where it can be proved that either algorithm improves on more elementary "randomized hillclimbing" algorithms.

We regard our tree-indexed random walk model of the graph and objective function as a caricature of an optimization problem on a large finite graph. In typical such problems, the neighborhood size increases with the problem size, but once a reasonably good value of the objective function S has ben found, most neighbors offer undesirable changes in S, so that an "effective neighborhood size" (i.e. moves with a non-vanishing chance of being accepted) can be regarded as bounded. Moreover on large graphs the Metropolis chain seems unlikely to return to a state in the short term except by retracing steps. Thus the idea of mimicing an optimization problem on a large graph by a problem on a bounded-degree tree (we chose binary merely for simplicity) isn't unreasonable; of course what is artificial is to model the increments of the objective function as independent random variables.

From the viewpoint of this caricature, if (7) were false it would imply an unsuspected weakness in Metropolis-type algorithms: there would be simple explicit optimization problems (see section 2.2) where Metropolis does much worse than greedy. Assuming (7) true suggests the following "practical" procedure. Suppose we do a long run of the Metropolis scheme on a finite problem with a fixed parameter b, and that the observed values of S look like a stationary process (call this *metastability*: note we are *not* assuming the Metropolis chain reaches its global stationary distribution). Can we predict whether repeating a run with $b' > b$ will be an improvement (e.g. compared to repeating with the same value of b)? For each state v there is a vector $(x_i(v), i \geq 1)$ of increments

$$\{S(w) - S(v) : w \text{ a neighbor of } v\}$$

arranged in decreasing order, say. Take a sample of states $(v_j, 1 \leq j \leq J)$ found in the parameter-b run, search exhaustively their neighbors $(v_{j,i}, i \geq 1)$, and calculate $\hat{\psi}(\theta) = J^{-1} \sum_j \sum_i \exp(\theta(S(v_{j,i}) - S(v_j)))$. Now predict

Using a larger value of b will be an improvement iff $\inf_{\theta > 0} \hat{\psi}(\theta) > 1$.

The point is that $\hat{\psi}(\theta)$ is an estimate of $\psi(\theta) = E \sum_i \exp(\theta x_i(V))$, where V has the metastable distribution attained by the run. In the caricature we assume the environment is such that the increments across edges at a single vertex are distributed as $(x_i(V), i \geq 1)$ independently over vertices. Then the classical large deviation analysis (4) shows that paths with positive rate of growth of $S(\cdot)$ exist if $\inf_\theta \psi(\theta) > 0$. By metastability, our parameter-b run has rate zero, i.e. the increments over edges satisfy the general-ξ analog of our "balance" condition $b = b_0(p)$ in (B). But given that paths exist with positive rate of growth of $S(\cdot)$, (7) suggests that the Metropolis algorithm with slightly large value of b will find such paths.

2.2 A random graph model

In this section we explain how our infinite tree-indexed random walk model for an optimization problem can be viewed precisely as a limit of a certain random model for a *finite* optimization problem.

Fix large $H \geq 1$ and $0 < p < 1/2$. Write $n_e(-1,0) = n_e(H, H+1) = 0$ and

$$n_e(h, h+1) = 3\lfloor \left(\tfrac{1-p}{p}\right)^{H-h} \rfloor, \ 0 \leq h \leq H-1.$$

$$n_v(h) = (n_e(h-1,h) + n_e(h,h+1))/3, \ 0 \leq h \leq H.$$

We construct a 3-regular graph with vertices in "levels" $h = 0, 1, \ldots, H$, with $n_v(h)$ vertices at level h and with $e(h, h+1)$ edges linking level h with level $h+1$ ($0 \leq h < H$), and with no other edges. The construction mimics the usual construction (e.g. [8] p. 374) of a random 3-regular graph. Start with the vertices arranged in levels, each with 3 "handles". Connect a pair of randomly-chosen handles in levels H and $H-1$, and continue connecting distinct random pairs until all handles in level H have been used. Then connect randomly-chosen handles in levels $H-1$ and $H-2$, and so on. See figure 4.

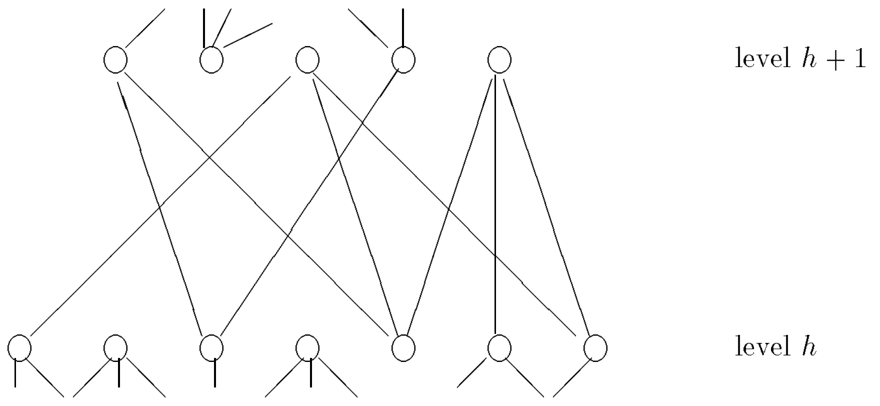

figure 4

The construction yields a random graph $\mathcal{G}^{(H)}$. The graph may have multiple edges or small components disconnected from the giant component, but that doesn't affect the $H \to \infty$ asymptotics of interest to us. For each vertex v of $\mathcal{G}^{(H)}$, define $S^{(H)}(v) =$ level of v. Fix b and write $(Y_t^{(H)}, t \geq 0)$ for the Metropolis chain on $\mathcal{G}^{(H)}$ with objective function $S^{(H)}$ and parameter b. Write

$$\bar{S}_t^{(H)} = S^{(H)}(Y_t^{(H)}) - S^{(H)}(Y_0^{(H)}), \ t \geq 0. \tag{8}$$

The point of the construction is

Lemma 5 *Let $h_H \to \infty$, $H - h_H \to \infty$ as $H \to \infty$. Let $(Y_t^{(H)}, t \geq 0)$ be the Metropolis chain on $(\mathcal{G}^{(H)}, S^{(H)})$ with initial state $Y_0^{(H)}$ uniform on level h_H. Then*

$$(\bar{S}_t^{(H)}, t \geq 0) \ \stackrel{d}{\to} \ (S(X_t), t \geq 0) \tag{9}$$

where (X_t) is the Metropolis chain associated with the infinite tree-indexed random walk (2), started at the root.

Here convergence is convergence of finite-dimensional distributions. The assertion of the lemma is intuitively clear: at a vertex at level h not near the "boundary" levels 0 and H, there are 3 edges, and the relative chances of an edge going to level $h+1$ or level $h-1$ are $p/(1-p)$ to 1, so the absolute chances are p and $1-p$. The details of the proof are uninteresting.

The Metropolis chain on the finite graph $(\mathcal{G}^{(H)}, S^{(H)})$ has stationary distribution

$$\pi(v) = \frac{e^{bS^{(H)}(v)}}{\sum_h n_v(h) e^{bh}}. \tag{10}$$

Under stationary we may write $\bar{S}_t^{(H)} = \sum_{u=1}^t \xi(X_{u-1}, X_u)$, or in words

(*) $\bar{S}_t^{(H)}$ is a sum of antisymmetric functionals along a stationary reversible chain.

Since $n_v(h) \propto \left(\frac{p}{1-p}\right)^h$, if we choose the special value $b = b_0(p)$ such that $e^b \times p/(1-p) = 1$, then for large H

$$\pi\{v : S^{(H)}(v) = h\} \approx 1/H, \ 0 \leq h \leq H.$$

In words, the *level* of the stationary chain is approximately uniform on $[0, H]$. So, for the special value $b = b_0(p)$, Lemma 5 applies to the *stationary* chain, implying (see section 5.1) that property (*) remains true for $S(X_t)$. This is not true for general b. The limit (9) for the stationary chains exists, but the limit environment is not the infinite-tree-indexed random walk, but instead is a modified tree with leaves.

3 Near-critical behavior of the greedy algorithm

3.1 Background

The main result of Aldous [1] was that, for the tree-indexed random walk model with general distribution ξ, the asymptotic rate for GREEDY is

$$r = E(\xi + Y)^+$$

where

$$Y = \sup_{\text{branches } (v_i)} \inf_i S(v_i) \leq 0 \tag{11}$$

and ξ is independent of Y. Specializing to the case where ξ is ± 1-valued, the rate becomes

$$r(p, \text{GREEDY}) = pP(Y = 0). \tag{12}$$

In the special case, the S-values at successive depths of the tree form a binary branching random walk on the integers, as follows. Start at time 0 with a single particle at position 0. At each step, replace each particle by two children, each independently placed at the parent's position plus one (with probability p) or at the parent's position minus one (with probability $1 - p$).

Modify this branching random walk by killing any child placed at position -1. Let $\rho(p) = P(Y = 0)$ be the probability that the modified process survives forever. While it is routine to set up equations which in principle determine the value of $\rho(p)$, it is not so routine to extract from these equations the behavior of $\rho(p)$ near the critical point p_{crit}. We will take a different approach to prove the following result, which by (12) implies Proposition 1.

Theorem 6

$$-\log \rho(p) = c(p - p_{\text{crit}})^{-1/2} + O(1) \ \textit{as } p \downarrow p_{\text{crit}}$$

$$\textit{where } c = \frac{\pi \log(\frac{1}{4p_{\text{crit}}})}{4\sqrt{1 - 2p_{\text{crit}}}} \approx 1.11. \tag{13}$$

It turns out that Theorem 6 can be proved by analyzing the simple recurrence relation (15), and we present this proof in section 3.2. This proof is elementary, and indeed could almost be given in a "mathematics for the analysis of algorithms" course in the spirit of [6].

One can also consider the more general setting where instead of ± 1-valued variables we have a one-parameter family $(\xi^{(p)})$ of distributions on R. Suppose the family is stochastically increasing with p and suppose there is a critical value p_{crit} defined by

$$\inf_{\theta>0} E\exp(\theta\xi^{(p_{\text{crit}})}) = 1/2.$$

Defining $\rho(p)$ as the non-extinction probability when particles entering $(-\infty, 0)$ are removed, one can argue informally that, under suitable regularity conditions, Theorem 6 should remain true with

$$c = \pi\theta^* \sqrt{\frac{E(\xi^{(p_{\text{crit}})})^2 \exp(\theta^*\xi^{(p_{\text{crit}})})}{2\frac{\partial}{\partial p} E\exp(\theta^*\xi^{(p)})|_{p=p_{\text{crit}}}}} \tag{14}$$

where $\theta^* = \arg\min_\theta E\exp(\theta\xi^{(p_{\text{crit}})})$.

Kesten [11] studied in great detail some analogous questions about branching Brownian motion with drift. His main results are stated for fixed supercritical or critical drift, but undoubtedly a result for Brownian motion analogous to Theorem 6 can be extracted from the technical estimates in [11]. Making a rigorous proof of (14) in the general setting involves issues analogous to justifying smoothness of solutions of renewal-type equations, and these seem different from the technical issues in [11]. Existing work in the discrete-time context seems limited to determining critical values (e.g. Biggins et al [3]). As observed in [1] and [3], the fact that $\rho(p) > 0$ iff $p > p_{\text{crit}}$ is a simple consequence of the standard large deviation analysis of the right-most walker in branching random walk.

3.2 Analysis of a recursion

Fix $p > p_{\text{crit}}$. Recall the definition (12) of Y, and write $a_n = P(Y \le -n)$. Then $a_0 = 1, a_1 = 1 - \rho(p)$, and by conditioning on the values of ξ on the two edges at the root we get the recursion

$$a_n = (pa_{n+1} + (1-p)a_{n-1})^2, \quad n \ge 1. \tag{15}$$

Rearranging,

$$a_{n+1} = \tfrac{1}{p}\sqrt{a_n} - \tfrac{1-p}{p}a_{n-1}, \ n \ge 1. \tag{16}$$

The key idea is to study instead the linear difference equation

$$b_{n+1} = \frac{1}{2p}b_n - \frac{1-p}{p}b_{n-1}, \quad n \ge 1 \tag{17}$$

with $b_0 = 0, b_1 = \rho(p)$. Note that (b_n) is the "linearization" of $1 - a_n$ for small $1 - a_n$. In what follows, asymptotics are always as $p \downarrow p_{\text{crit}}$, and we will sometimes assume $p - p_{\text{crit}}$ is "sufficiently small" in nonasymptotic assertions. We use "big-O" notation: $d(p) = \Omega(e(p))$ means $e(p) = O(d(p))$, and $d(p) = \Theta(e(p))$ means $d(p) = O(e(p))$ and $d(p) = \Omega(e(p))$.

It is elementary that the solution to the recurrence (17) is

$$b_n = \rho(p)\frac{z^n - \bar{z}^n}{z - \bar{z}} = \rho(p)r^{n-1}\frac{\sin n\varepsilon}{\sin \varepsilon}, \ n \geq 1 \tag{18}$$

where $(z, \bar{z}) = (re^{i\varepsilon}, re^{-i\varepsilon})$ are the solutions of $x^2 - \frac{1}{2p}x + \frac{1-p}{p} = 0$, which are

$$z, \bar{z} = \frac{1 \pm \sqrt{1 - 16p(1-p)}}{4p}.$$

The critical value p_{crit} is such that $16p_{\text{crit}}(1 - p_{\text{crit}}) = 1$, and a brief calculation gives

$$r = r(p) = \frac{1}{4p_{\text{crit}}} + O(p - p_{\text{crit}}) \tag{19}$$

$$\varepsilon = \varepsilon(p) = 4\sqrt{1 - 2p_{\text{crit}}}(p - p_{\text{crit}})^{1/2} + o(1). \tag{20}$$

Define $N = N(p)$ by

$$N = \min\{n : n\varepsilon(p) \geq \pi\}. \tag{21}$$

Our goal is to show

$$\rho(p)r^N = \Theta(1). \tag{22}$$

Granted (22),

$$-\log \rho(p) = N \log r + O(1) = \frac{\pi}{\varepsilon(p)} \log r(p) \ + O(1)$$

and Theorem 6 follows from (19,20).

The heuristic explanation of (22) is as follows. The approximation of $1 - a_n$ by b_n should hold as long as these quantities are $o(1)$. The sequence b_n increases until, for $n = N - \Theta(1)$, it reaches a maximum value $\Theta(\rho(p)r^N)$ and then decreases and becomes negative. So the natural place for the approximation to break down is at $n = N - \Theta(1)$, so this should be the first time that $b_n = \Theta(1)$.

To start the rigorous analysis, define

$$N_0 = \min\{n \geq 1 : b_n < 2(1-p)b_{n-1}\}.$$

Note that $b_n > 0$ for $1 \le n < N_0$. Let $c_0 = 0$ and for $1 \le n < N_0$ define c_n by

$$1 - a_n = b_n c_n \tag{23}$$

so that $c_1 = 1$. The recursions (16,17) for (a_n) and (b_n) imply a recursion for (c_n), which after elementary manipulations becomes

$$c_{n+1} = \frac{1 - \sqrt{1 - b_n c_n} - (1-p) b_{n-1} c_{n-1}}{\frac{1}{2} b_n - (1-p) b_{n-1}}, \ 1 \le n < N_0.$$

Subtracting c_n gives

$$c_{n+1} - c_n = \frac{1 - \sqrt{1 - b_n c_n} - \frac{1}{2} b_n c_n + (1-p) b_{n-1} (c_n - c_{n-1})}{\frac{1}{2} b_n - (1-p) b_{n-1}}, 1 \le n < N_0. \tag{24}$$

Using the inequality

$$1 - \sqrt{1-x} - \tfrac{1}{2} x \ge 0 \text{ for } 0 \le x \le 1,$$

(24) implies inductively that

$$c_n - c_{n-1} \ge 0; \quad c_n \ge 1, \quad 1 \le n < N_0. \tag{25}$$

In particular, $c_{N_0 - 1} \ge 1$, and since $a_n \ge 0$ for all n we have from (23) that

$$b_{N_0 - 1} \le 1. \tag{26}$$

From the exact formula (18) for b_n we have

$$\frac{b_n}{2(1-p) b_{n-1}} = \frac{1}{8p(1-p)} \frac{\sin n\varepsilon}{\sin(n-1)\varepsilon}.$$

Since $\frac{1}{8p(1-p)} \to 2$ it follows easily from the definition of N_0 that

$$-o(\varepsilon) \le \pi - N_0 \varepsilon \le (1 + o(1))\varepsilon$$

$$N - N_0 = O(1); \quad \frac{\sin(N_0 - 1)\varepsilon}{\sin \varepsilon} = \Theta(1). \tag{27}$$

So by the exact formula for b_n

$$b_{N_0 - 1} = \Theta(\rho(p) r^N)$$

which by (26) gives

$$\rho(p) r^N = O(1). \tag{28}$$

To get a bound in the other direction, define (e_n) by

$$1 - a_n = b_n + e_n. \tag{29}$$

So $e_0 = e_1 = 0$. The recursions (16,17) lead to the recursion

$$e_{n+1} = \frac{1}{p}\left(1 - \frac{1}{2}b_n - \sqrt{1 - b_n - e_n}\right) - \frac{1-p}{p}e_{n-1}, \ n \geq 1. \tag{30}$$

Define

$$N_1 = \min\left\{n \geq 1 : b_{n+1}/b_n < \sqrt{\tfrac{2}{3p_{\mathrm{crit}}}}\ \right\}.$$

Since $\frac{1}{4p_{\mathrm{crit}}} > \sqrt{\frac{2}{3p_{\mathrm{crit}}}} > 2(1 - p_{\mathrm{crit}})$ we see that (for sufficiently small $p - p_{\mathrm{crit}}$) we have $N_1 \leq N_0$. Arguing as for (27),

$$N - N_1 = O(1); \quad \frac{\sin(N_1 - 1)\varepsilon}{\sin\varepsilon} = \Theta(1). \tag{31}$$

By (25) we have

$$e_n \geq 0; \ 1 \leq n < N_1. \tag{32}$$

Define

$$N_2 = \min\{n : b_n > 1/10\}.$$

We shall show by induction that

$$e_n \leq b_n^2, \quad 1 \leq n \leq \min(N_1, N_2). \tag{33}$$

We use the calculus bound

$$1 - \tfrac{1}{2}x - \sqrt{1 - x - x^2} \leq \tfrac{2}{3}x^2, \quad 0 \leq x \leq 1/10.$$

So if (33) holds for a particular $n < \min(N_1, N_2)$ then

$$\begin{aligned}
e_{n+1} &\leq \tfrac{1}{p}\left(1 - \tfrac{1}{2}b_n - \sqrt{1 - b_n - e_n}\right) \\
&\leq \tfrac{1}{p_{\mathrm{crit}}}\left(1 - \tfrac{1}{2}b_n - \sqrt{1 - b_n - b_n^2}\right) \text{ by (33)} \\
&\leq \tfrac{1}{p_{\mathrm{crit}}}\tfrac{2}{3}b_n^2 \text{ by the calculus bound, since } n < N_2 \\
&\leq b_{n+1}^2 \text{ since } n < N_1
\end{aligned}$$

establishing (33). Next, by considering subsequences of p's we may suppose either

(a) $b_{\min(N_1,N_2)} \to 0$; or
(b) $b_{\min(N_1,N_2)} = \Theta(1)$.
Suppose case (a) holds, in which case $b_{N_1} \to 0$. For fixed $k \geq 1$ we have $|b_{N_1+k}| = O(b_{N_1})$. By (33) we have $e_{N_1}/b_{N_1} \to 0$ and then, using (30) and induction on k,

$$\frac{|e_{N_1+k}|}{b_{N_1}} \to 0, \text{ for fixed } k \geq 0.$$

But for some $k = \Theta(1)$ we have $b_{N_1+k} = \Omega(-b_{N_1})$ and hence $b_{N_1+k} + e_{N_1+k}$ is negative, which is impossible by (29). Thus case (b) must hold. But (b_n) is increasing on $n \leq N_1$, and by (31) $\frac{\sin N_1 \varepsilon}{\sin \varepsilon} = \Theta(1)$, so

$$\Theta(1) = b_{\min(N_1,N_2)} \leq b_{N_1} = \Theta(\rho(p) r^{N_1})$$

which by (31) leads to

$$\rho(p) r^N = \Omega(1). \tag{34}$$

Now (34) and (28) imply (22), establishing Theorem 6.

4 Speeds of random walks and depth-first search on Galton-Watson trees

Let $\mathcal{T}$ be a supercritical Galton-Watson tree with offspring distribution $\mathbf{p} = (p_i, i = 0, 1, 2, \ldots)$, conditioned on non-extinction. Write $D(v)$ for the depth of vertex v. A simpler analog of seeking algorithms to find large values of $S(v)$ in the tree-indexed random walk model is to seek algorithms to find large values of $D(v)$ in this Galton-Watson setting. As in the former case, one feels that for any reasonable algorithm ALG which generates a deterministic or random sequence of vertices (root $= X_0, X_1, X_2, \ldots$) to be examined, there should be some asymptotic "speed"

$$\lim_{n\to\infty} n^{-1} D(X_n) = s(\mathbf{p}, \mathrm{ALG}) \text{ a.s.} \tag{35}$$

The natural greedy algorithm is *depth-first search*: having examined vertices $v_0, v_1, \ldots, v_n$, choose from that set a maximal-depth vertex v^* with some child not in that set, and let v_{n+1} be a child of v^*. It is elementary to give a expression (41) for $s(\mathbf{p}, \mathrm{DEPTH-FIRST})$. The analog of the Metropolis chain (1) is simple symmetric random walk (X_n). Lyons-Pemantle-Peres ([16] page 601) establish the (non-obvious) formula

$$s(\mathbf{p}, \mathrm{RANDOM-WALK}) = \sum_{i=0}^{\infty} \frac{i-1}{i+1} p_i \frac{1-q^{i+1}}{1-q^2} \tag{36}$$

where $q < 1$ is the extinction probability, i.e. the solution of

$$q = \sum_{i=0}^{\infty} p_i q^i. \tag{37}$$

We consider a family, parametrized by the mean $\lambda \in [1, \lambda_0]$, of probability distributions $(p_i(\lambda); i = 0, 1, 2, 3, \ldots)$ on non-negative integers. Write

$$\begin{aligned} \lambda = m(\lambda) &= E\xi_\lambda \\ \sigma^2(\lambda) &= E\xi_\lambda(\xi_\lambda - 1) \\ \kappa^3(\lambda) &= E\xi_\lambda(\xi_\lambda - 1)(\xi_\lambda - 2) \end{aligned}$$

where ξ_λ has distribution $(p_i(\lambda))$. The behaviors of the speeds as $\lambda \downarrow 1$, that is as the trees approach criticality, are as follows.

Corollary 7 *Suppose* $\mathbf{p}(\lambda) \to \mathbf{p}(1)$ *as* $\lambda \downarrow 1$, *and suppose* $\sigma^2(1) > 0$ *and* $\sup_\lambda \kappa^3(\lambda) < \infty$. *Write* $s(\lambda) = s(\mathbf{p}(\lambda), \mathrm{RANDOM-WALK})$ *and* $s_*(\lambda) = s(\mathbf{p}(\lambda), \mathrm{DEPTH-FIRST})$. *Then*
(a) $s(1) = 0,\ s'(1) = 0,\ s''(1) = \frac{2}{3\sigma^2(1)}$.
(b) $s_*(1) = 0,\ s'_*(1) = \frac{2}{\sigma^2(1)}$.

In other words, depth-first search has speed $O(\lambda - 1)$ whereas random walk has speed $O\left((\lambda - 1)^2\right)$. A verbal explanation of the difference is given after the proof.

Proof of Corollary 7. We first give the routine-but-tedious calculus to derive (a) from (36).

Step 1. Let $f(x, y)$ be smooth and let $y = y(x)$ be the solution of $f(x, y) = 0$. Differentiating once and twice gives

$$0 = f_x + y' f_y$$

$$0 = f_{xx} + 2y' f_{xy} + y'' f_y + (y')^2 f_{yy}$$

which rearranges to

$$y' = \frac{-f_x}{f_y} \tag{38}$$

$$y'' = -\frac{f_{xx}}{f_y} + \frac{2 f_x f_{xy}}{f_y^2} - \frac{f_x^2 f_{yy}}{f_y^3}. \tag{39}$$

Step 2. Writing $y = 1 - q$, formula (37) for y in terms of λ can be written as the solution of $f(\lambda, y) = 0$ where

$$f(\lambda, y) = y^{-1}\left(\sum_i p_i(\lambda)(1-y)^i \; - (1-y)\right).$$

Computing partial derivatives at $\lambda = 1$,

$$f_\lambda = -m' = -1; \quad f_{\lambda\lambda} = -m'' = 0$$

$$f_y = \frac{1}{2}\sigma^2; \quad f_{yy} = -\frac{1}{3}\kappa^3$$

$$f_{\lambda y} = \frac{1}{2}(\sigma^2)'.$$

Substituting into (38,39), the derivatives of $y(\lambda)$ at λ_0 are

$$y' = \frac{2}{\sigma^2} \tag{40}$$

$$y'' = -\frac{4(\sigma^2)'}{\sigma^4} + \frac{8\kappa^3}{3\sigma^6}.$$

Step 3. Write

$$t_i(y) = \frac{1}{i+1}\,\frac{1-(1-y)^{i+1}}{1-(1-y)^2}.$$

Then the derivatives w.r.t. y at $y = 0$ are

$$t_i = \tfrac{1}{2} \; t_i' = -\tfrac{i-1}{4} \; t'' = \tfrac{i(i-1)}{6} + \frac{1}{4}.$$

Step 4. We are interested in the speed

$$s(\lambda) = \sum_i (i-1)p_i(\lambda)t_i(y(\lambda)).$$

The first derivative at $\lambda = 1$ is

$$\begin{aligned} s' &= m'\,\tfrac{1}{2} + \sum_i (i-1)p_i y'\left(-\tfrac{i-1}{4}\right) \\ &= \tfrac{1}{2} - \tfrac{\sigma^2 y'}{4} \\ &= 0. \end{aligned}$$

The second derivative is

$$s'' = m'' \tfrac{1}{2} + 2\sum_i (i-1)p_i' y' \left(-\tfrac{i-1}{4}\right) + \sum_i (i-1)p_i y'' \left(-\tfrac{i-1}{4}\right) + \sum_i (i-1)p_i (y')^2 \left(\tfrac{i(i-1)}{6} + \tfrac{1}{4}\right)$$

$$= 0 - \frac{y'(\sigma^2)'}{2} - \frac{y''\sigma^2}{4} + \frac{(y')^2}{6}\sum_i p_i i(i-1)^2 + (y')^2 \cdot 0$$

$$= \frac{m''}{2} - \frac{m'(\sigma^2)'}{\sigma^2} - \frac{m''}{2} + \frac{m'(\sigma^2)'}{\sigma^2} - \frac{2(m')^2\kappa^3}{3\sigma^4} + \frac{4(m')^2}{6\sigma^4}(\kappa^3 + \sigma^2)$$

$$= \frac{2(m')^2}{3\sigma^2} = \frac{2}{3\sigma^2}.$$

This establishes (a). The analysis of depth-first search involves some elementary and well-known properties of Galton-Watson trees. A *backbone* vertex is one with an infinite line of descendants. Clearly we have

$$s(\mathbf{p}, \mathrm{DEPTH-FIRST}) = \frac{1}{1+\alpha\beta} \tag{41}$$

where α is the mean number of non-backbone children of a backbone vertex that are examined before encountering a backbone vertex; and β is the mean total progeny in the Galton-Watson tree conditioned to become extinct. One may derive formulas for α and β in terms of $\mathbf{p}$ as follows. The Galton-Watson tree conditioned to become extinct has offspring distribution

$$\hat{p}_i = \frac{p_i q^i}{q}$$

with mean

$$\hat{m} = \sum_i i\hat{p}_i,$$

and so $\beta = 1/(1-\hat{m})$. For a backbone vertex, the chance of a particular sequence such as (N, N, B, N, B) of n non-backbone children and $b \geq 1$ backbone children is $p_{n+b}q^n(1-q)^b/(1-q)$, and α can be expressed in terms of this distribution.

In the $\lambda \downarrow 1$ setting of the corollary, one can directly see the limiting behavior of $\alpha(\lambda)$. In the limit, a backbone vertex has exactly one backbone child, and the total number of children is the *size-biased* distribution

$$P(\bar{\xi} = i) = ip_i(1), \ i \geq 1.$$

So

$$\alpha(1) = \tfrac{1}{2}(E\bar{\xi} \ -1) = \frac{1}{2}\left(\frac{E\xi^2}{E\xi} - 1\right) = \frac{\sigma^2(1)}{2}.$$

So by (41), proving (b) reduces to proving

$$1 - \hat{m}(\lambda) \sim \lambda - 1.$$

This is routine:

$$\begin{aligned} m(\lambda) - \hat{m}(\lambda) &= \sum_i i p_i(\lambda)(1 - q^{i-1}(\lambda)) \\ &\sim \sum_i i p_i(\lambda)\ (i-1)(1-q(\lambda)) \\ &\sim (1-q(\lambda))\sigma^2(1) \\ &\sim 2(\lambda - 1) \end{aligned}$$

because $\frac{dq(\lambda)}{d\lambda} = \frac{-2}{\sigma^2(1)}$ at $\lambda = 1$ by (40).

Remarks. Here is an informal explanation of the corollary. When the mean number of offspring is $1+\delta$, the backbone is a branching process with chance $\Theta(\delta)$ of two children, and the side branches have mean size $\Theta(1/\delta)$. Depth-first search spends time $\Theta(1/\delta)$ in branches before taking a step down the backbone, so its speed is $\Theta(\delta)$. Random walk also spends mean time $\Theta(1/\delta)$ in each visit to a branch, that is between successive steps on the backbone. And random walk restricted to the backbone has drift rate $\Theta(\delta)$, so to reach level L requires $\Theta(L/\delta)$ steps on the backbone, which means $\Theta(\frac{L}{\delta} \times \frac{1}{\delta})$ steps in total.

5 Analysis of the Metropolis algorithm

5.1 Background

As mentioned in the introduction, the Metropolis chain is analogous to RWIRE (Random Walk in Random Environment, which on Z^d is a well-studied topic with statistical physics motivation [7]) and in particular to recent work of Lyons-Pemantle-Peres [16, 17] studying speed of random walk on Galton-Watson trees. So we shall view the objective function $S(v)$ as defining a "random environment" on the infinite binary tree. Note that the Metropolis chain has an "antisymmetric" character, so in general is different from the "symmetric" random walk on a random electrical network (see [15] for an example on trees).

In the context of biased random walk on Galton-Watson trees, Lyons-Pemantle-Peres ([18] Question 2.1) noted that, while monotonicity of speed as a function of bias seems intuitively obvious, and one might expect some simple coupling proof, there is no known proof. Analogously, monotonicity of the optimal Metropolis rate $r^*(p)$ at (5) as a function of p seems intuitively obvious, but we do not see a proof.

In analysis of RWIRE a central role is played by the notion of "**E**nvironment as seen by the **W**alker". Given a graph G and a function $s(\cdot)$ on vertices, for each vertex v_* ("position of walker") define $\mathbf{EW}(v_*)$ to be the graph rooted at v_*, with the function $s^*(v) = s(v) - s(v_*)$. So $\mathbf{EW}(v)$ takes values in the space $\mathcal{E}$ of all possible objective functions $(s(v) : v \in \mathcal{T}^\infty)$. Thus associated with the Metropolis chain $Y_t^{(H)}$ on the finite random graphs $(\mathcal{G}^{(H)}, S^{(H)})$ in section 2.2, or with the Metropolis chain X_t on the infinite binary tree with objective function $S(\cdot)$, are processes $(\mathbf{EW}(Y_t^{(H)}))$ and $(\mathbf{EW}(X_t))$ describing the environment as seen by the walker. Lemma 5 extends to prove weak convergence

$$(\mathbf{EW}(Y_t^{(H)}), t \geq 0) \ \stackrel{d}{\to} \ (\mathbf{EW}(X_t), t \geq 0) \tag{42}$$

under the same assumptions. (We skip over technical issues such as embedding the state space of $\mathbf{EW}(Y_t^{(H)})$ into $\mathcal{E}$.) If we take $(Y_t^{(H)})$ to be the stationary process then the process $\mathbf{EW}(Y_t^{(H)})$ is stationary and reversible. Taking $H \to \infty$ limits as in section 2.2, we see that for the special value $b = b_0(p)$ the process $\mathbf{EW}(X_t)$ is stationary and reversible. For general b, though the process $\mathbf{EW}(X_t)$ does have a stationary distribution $\Xi = \Xi_{p,b}$ (see section 5.2), the stationary distribution is not the initial distribution $\mathbf{EW}(X_0)$ and the stationary process is not reversible.

The most interesting part of our analysis is the non-rigorous "differentiation with respect to the parameter b" argument in section 5.3. Such arguments are part of the statistical physics toolkit, but in our context seem very hard to make rigorous. The closest rigorous argument I know is that in Lebowitz - Rost [13], in the context of small perturbations of Brownian motion.

As we shall see, general results imply existence of the limit variance rate $\sigma^2(p, b_0(p))$, but there are no general results to distinguish whether the limit is positive or zero (i.e. whether the chain is *diffusive* or *subdiffusive*). Such questions have been well-studied in the context of a tagged particle in the symmetric exclusion process, which is subdiffusive in the one-dimensional nearest-neighbor case and diffusive otherwise ([14] sec. 8.4). By analogy with this and other examples of RWIRE one expects diffusive behavior in our setting, but I do not see any simple proof strategy.

Recall a helpful way to view the Metropolis chain: from the current vertex v pick a uniform random neighbor w as a "proposed" move, and "accept" the move with probability $p_{\mathrm{acc}}(\xi(v,w))$, where

$$p_{\mathrm{acc}}(x) = \frac{e^{bx}}{1+e^{bx}}. \tag{43}$$

5.2 Outline proof of Theorem 3

Writing a rigorous proof requires substantial investment in notation and technical background, so we shall just outline the major points. For easy reference we restate the theorem.

Fix $1/2 > p > p_{\mathrm{crit}}$ *and* $b \geq 0$. *Write* $b_0(p) = \log \frac{1-p}{p} > 0$. *Then the limit rate*

$$\lim_{n\to\infty} n^{-1} S(X_n) = r(p,b) \ a.s. \tag{44}$$

exists, and

(A) $r(p,0) = \frac{2p-1}{6} < 0$.

(B) $r(p,b_0(p)) = 0$.

(C) For $b = b_0(p)$ *there is a variance rate*

$$\sigma^2(p,b_0(p)) = \lim_t t^{-1} \mathrm{var}\ S(X_t) \geq 0.$$

(D) $r(p,b) \leq 0$ *for* $0 \leq b \leq b_0(p)$ *and* $r(p,b) \geq 0$ *for* $b_0(p) \leq b < \infty$.

(E) $r(p,b) \to 0$ *as* $b \to \infty$.

Existence of $r(p,b)$ is obvious when $b = 0$, so let us consider only the case $b > 0$.

Lemma 8 *For almost all realizations of the environment, the chain is transient.*

Proof. Fix the environment $(S(v))$. In the well-known correspondence [5] between reversible Markov chains and electrical networks, the Metropolis chain corresponds to the network where a parent-child edge (v,w) has conductance (i.e. 1/ resistance)

$$a(v,w) = e^{bS(v)} p_{\mathrm{acc}}(\xi(v,w)) = e^{bS(w)} p_{\mathrm{acc}}(\xi(w,v)).$$

The chain is transient iff there exists a unit flow $\mathbf{f} = (f(v,w))$ from the root to infinity such that $\sum_{(v,w)} f^2(v,w)/a(v,w) < \infty$. By considering the unit flow along a single ray (v_i), to prove transience it suffices to exhibit a ray satisfying $\sum_i e^{-bS(v_i)} < \infty$. This in turn holds if the ray satisfies $\liminf_i i^{-1} S(v_i) > 0$. But rays with this property exist by (4).

The remainder of the proof of (44) uses only "soft" arguments (cf. [17] section 3), which we outline very briefly. Consider a version of the environment in which a finite number of ξ-values are fixed arbitrarily. By transience, the chain visits new vertices infinitely often. For the chains (X_t^1) and (X_t^2) associated with two such versions of the environment, is it easy to exhibit a 'local shift-coupling" in which for each k

$$\rho_k \ \mathbf{EW}(X_t^1) = \rho_k \ \mathbf{EW}(X_{t+\Sigma}^2) \text{ for all } T_k \le t < \infty$$

where Σ is a random time-shift and ρ_k denotes the restriction of the environment to within distance k from the root. By considering subsequential weak limits of time-averaged distributions $\mathbf{EW}(X_t)$ one obtains a stationary distribution Ξ for $\mathbf{EW}(X_t)$. Then the existence of local shift-couplings implies that the stationary process implies that the stationary process is ergodic and identifies $r(p,b)$ as the ergodic rate.

(A) For $b = 0$ the Metropolis chain is simple symmetric random walk on the tree (paying no attention to the S-values), with chance 1/2 of holding. It is elementary that the depth $D(X_n)$ grows at rate 1/6, and so $S(X_n)$ grows at rate $\frac{1}{6} \times E\xi = \frac{1}{6}(2p-1)$.

(B,C) These follow from the fact that for $b = b_0(p)$ the process $(\mathbf{EW}(X_t))$ is stationary and reversible. The point is that there is a representation

$$S(X_{t+1}) - S(X_t) = \psi(\mathbf{EW}(X_{t+1}) - \mathbf{EW}(X_t)) \tag{45}$$

where ψ is antisymmetric, i.e. $\psi(\eta_1, \eta_2) = -\psi(\eta_2, \eta_1)$. So the stationary reversible property immediately implies $ES(X_{t+1}) = ES(X_t)$, giving (B). Moreover it is well known and easy that (45) implies a subadditivity property

$$\text{var } (S(X_{2t})) \le 2 \text{ var } (S(X_t))$$

from which (C) follows, with $\sigma^2(p, b_0(p)) \le ES^2(X_1) \le 1$. In fact general theory [12, 19] gives a central limit theorem

$$n^{-1/2} S(X_t) \ \xrightarrow{d} \ \text{Normal}(0, \sigma^2(p, b_0(p))).$$

(D) We use Lemma 5, which represents $(S(X_t))$ as a weak limit of $(S^{(H)}(Y_t^{(H)}))$. We will do the case $b < b_0(p)$: the other case is similar. To get a contradiction, suppose $r(p,b) > 0$. Choose m_0 so that $P(\inf_t S(X_t) < -m_0) < 1/5$. Then for all sufficiently large t,

$$P\left(S(X_t) \geq \tfrac{t}{2} r(p,b), \min_{0 \leq u \leq u} S(X_u) \geq -m_0\right) \geq 3/5.$$

Fix such a t. Using Lemma 5, for sufficiently large H, the Metropolis chain $Y_u^{(H)}$ started from a uniform random vertex at level h (for arbitrary $H/3 \leq h \leq 2H/3$) satisfies

$$P\left(S^{(H)}(Y_t^{(H)}) - h \geq \tfrac{t}{2} r(p,b)\right) \geq 2/5.$$

Thus the stationary distribution π of $Y^{(H)}$ satisfies

$$\pi\{v : S^{(H)}(v) \geq \tfrac{H}{3} + \tfrac{t}{2} r(p,b)\} \geq \tfrac{2}{5} \pi\{v : \tfrac{H}{3} \leq S^{(H)}(v) \leq \tfrac{2H}{3}\}.$$

But for large t and H this contradicts the fact (10) that $\pi\{v : S^{(H)}(v) = l\} \propto \alpha^l$ where $\alpha = \frac{pe^b}{1-p} < 1$.

(E) follows from (C) and the very crude bound

$$r(p,b) \leq (1-p)^{-2}(1+e^b)^{-1}.$$

Consider the times $U_1, U_2, \ldots$ at which the chain reaches a previously-unvisited vertex via a $+1$ edge. With chance $(1-p)^2$ both outward edges from $X(U_m)$ have ξ-value -1, in which case the walk stays at $X(U_m)$ for mean time $1+e^b$, and so $E(U_{m+1} - U_m | X(t), t \leq U_m) \geq (1-p)^2(1+e^b)$. By renewal theory, $\liminf_m m^{-1} U_m \geq (1-p)^2(1+e^b)$ a.s. Since $S(X_t) \leq \max\{m : U_m \leq t\}$, the stated bound follows.

5.3 Regarding Conjecture 4

We show how (F,G) may be obtained by an argument involving (47) differentiating an ergodic limit w.r.t. a parameter. I have no idea how to make this argument rigorous.

Give the environment S the stationary distribution Ξ of $\mathbf{EW}(X_t)$. Write $w \sim v$ if (v,w) is an edge, and write $\mathcal{L}(v) = \sigma(\xi(v,w) : w \sim v)$. Define

$$Q(v) = \lim_n E_v(S(X_n) - S(v) - nr(p,b) | \mathcal{L}(v)). \qquad (46)$$

The fundamental idea is

$$r_b(p,b) = \frac{1}{3} \sum_{w \sim \text{root}} E \, \frac{dp_{\text{acc}}(\xi(\text{root},w))}{db} \, (\xi(\text{root},w) + Q(w) - Q(\text{root})). \quad (47)$$

To argue this, consider the effect of replacing b by $b+db$. At a typical step, with chance $dp_{\text{acc}}(\xi(\text{root},w))$ the $b+db$ process makes a step to a neighbor w while the b process makes no move. Conditional on the entire environment $S(\cdot)$, the long-run effect of the different step is $\xi(\text{root},w) + \tilde{Q}(w) - \tilde{Q}(\text{root})$, where $\tilde{Q}$ is the analog of (46) obtained by conditioning on the entire environment. Thus the overall effect is given by (47) with $\tilde{Q}$ in place of Q, which is equivalent to (47).

Of course, (47) is only useful when we know the stationary distribution Ξ, which is only for $b=0$ and $b=b_0(p)$. We first consider the simpler case $b=0$, where the Metropolis chain is simple random walk with holding probability $1/2$. The stationary distribution of the environment is as follows. Take the tree-indexed random walk, pick a uniform random ray from the root to infinity, and for each edge (v,w) in that ray replace $\xi(v,w)$ by $-\xi(v,w)$. Because

$$\frac{dp_{\text{acc}}(x)}{db} = \frac{x}{4} \text{ at } b=0$$

(47) gives

$$r_b(p,0) = \frac{1}{12} \sum_{w \sim \text{root}} E\xi(\text{root},w)(\xi(\text{root},w) + Q(w) - Q(\text{root}))$$

$$= \frac{1}{12} \sum_{v \sim \text{root}} (1 + E[\xi(\text{root},w) E(Q(w) - Q(\text{root}) | \xi(\text{root},w))]). \quad (48)$$

The key fact is that for any edge (v,w),

$$E(Q(v)|\xi(v,w)) = \frac{1}{3}(\xi(v,w) - E\xi(v,w)).$$

This holds because the final exit of the walk from v is equally likely to be along any of the three edges at v, and the ξ-values are independent over edges and independent of the walk. So for $w \sim \text{root}$,

$$E(Q(w)|\xi(\text{root},w)) = E(Q(w)|\xi(w,\text{root}))$$

$$= \frac{1}{3}(\xi(w,\text{root}) - E\xi(w,\text{root})) = -\frac{1}{3}(\xi(\text{root},w) - E\xi(\text{root},w)).$$

So

$$E(Q(w) - Q(\mathrm{root})|\xi(\mathrm{root}, w)) = -\frac{2}{3}(\xi(\mathrm{root}, w) - E\xi(\mathrm{root}, w)).$$

So

$$E\xi(\mathrm{root}, w)E(Q(w) - Q(\mathrm{root})|\xi(\mathrm{root}, w)) = -\frac{2}{3}\mathrm{var}\,\xi(\mathrm{root}, w) = -\frac{2}{3}(1-(2p-1)^2)$$

and substituting into (48) establishes (F).

Now consider the case $b = b_0(p) = \log\frac{1-p}{p}$. Here the stationary distribution Ξ for **EW** is just the tree-indexed random walk itself. In particular, by symmetry we can reformulate (47) in terms of a single neighbor w of the root:

$$r_b(p, b_0(p)) = E\ \frac{dp_{\mathrm{acc}}(\xi(\mathrm{root}, w))}{db}\ (\xi(\mathrm{root}, w) + Q(w) - Q(\mathrm{root}))$$

where

$$Q(\mathrm{root}) = \lim_n E_{\mathrm{root}}(S(X_n)|\xi(\mathrm{root}, w))$$
$$Q(w) = \lim_n E_w(S(X_n) - S(w)|\xi(\mathrm{root}, w)).$$

Now

$$\frac{dp_{\mathrm{acc}}(x)}{db} = \frac{xe^{bx}}{(1+e^{bx})^2}$$

and so

$$r_b(p, b_0(p)) = E\frac{\xi(\mathrm{root}, w)e^{b\xi(\mathrm{root},w)}(\xi(\mathrm{root}, w) + Q(w) - Q(\mathrm{root}))}{(1+e^{b\xi(\mathrm{root},w)})^2} \quad (49)$$

where $b = b_0(p)$. Now $Q(w)$ does not have the same unconditional distribution as $Q(\mathrm{root})$, but it is easy to see the relationship between conditional distributions:

$$\mathrm{dist}(Q(w)|\xi(w, \mathrm{root}) = x) = \mathrm{dist}(Q(\mathrm{root})|\xi(\mathrm{root}, w) = x).$$

Since the likelihood ratio is

$$L(x) \equiv P(\xi(w, \mathrm{root}) = x)/P(\xi(\mathrm{root}, w) = x) = e^{bx}$$

we have for any function ϕ

$$EQ(w)\phi(\xi(\mathrm{root}, w)) = EQ(w)\phi(-\xi(w, \mathrm{root})) = EQ(\mathrm{root})\phi(-\xi(\mathrm{root}, w))e^{-b\xi(\mathrm{root},w)}.$$

In particular, the term of (49) involving $Q(w)$ becomes

$$EQ(w)\frac{\xi(\text{root},w)e^{b\xi(\text{root},w)}}{(1+e^{b\xi(\text{root},w)})^2} = -E\frac{Q(\text{root})\xi(\text{root},w)}{(1+e^{b\xi(\text{root},w)})^2}.$$

So (49) reduces to

$$r_b(p,b_0(p)) = \frac{e^b}{(1+e^b)^2} - E\frac{Q(\text{root})\xi(\text{root},w)}{1+e^{b\xi(\text{root},w)}}.$$

Using the definition of $b_0(p)$, the first term equals $p(1-p)$ and also

$$\frac{\xi(\text{root},w)}{1+\xi(\text{root},w)} = \frac{\xi(\text{root},w)+2p-1}{2},$$

so since $EQ(\text{root}) = 0$ we find

$$r_b(p,b_0(p)) = p(1-p) - \frac{1}{2}EQ(\text{root})\xi(\text{root},w). \tag{50}$$

We do not a formula for $Q(\text{root})$, so this does not enable us to actually calculate the derivative. But the asymptotic variance rate for partial sums of a stationary mean-zero sequence (η_i) can be written as

$$E\eta_0^2 + 2\lim_t E\left(\eta_0\sum_{i=1}^t \eta_i\right).$$

Applying this to $\eta_i = S(X_i) - S(X_{i-1})$,

$$\sigma^2 = \sigma^2(p,b_0(p)) = E\eta_0^2 + 2\lim_t E\eta_0 S(X_t). \tag{51}$$

Here $\eta_0 = -S(X_{-1})$, where by reversibility X_{-1} is the position after a step of the chain independent (conditional on the environment) of later steps. To study this, we may suppose the attempted move is to w, and write $\xi = \xi(\text{root},w)$ and A for the event that the move is accepted. Then $\eta_0 = -\xi 1_A$. Writing as before

$$Q(\text{root}) = \lim_n E_{\text{root}}(S(X_n)|\xi(\text{root},w))$$

formula (51) becomes

$$\sigma^2 = Ep_{\text{acc}}(\xi) - 2EQ(\text{root})\xi 1_A.$$

Now for $b = b_0(p)$ we have

$$p_{\mathrm{acc}}(\xi) = \frac{1 + (1-2p)\xi}{2}.$$

So $E p_{\mathrm{acc}}(\xi) = 2p(1-p)$ and

$$EQ(\mathrm{root})\xi 1_A = EQ(\mathrm{root})\xi p_{\mathrm{acc}}(\xi) = EQ(\mathrm{root})\frac{\xi + (1-2p)}{2}.$$

Since $EQ(\mathrm{root}) = 0$, the formula reduces to

$$\sigma^2 = 2p(1-p) - EQ(\mathrm{root})\xi.$$

Comparing with (50) gives (G).

Simulation results. For the reader unimpressed by the argument for (G), we mention that it is supported by simulation evidence for $p = 0.2$. Examining the slope in figure 2 gives $r_b(0.2, b_0(0.2)) \approx 7 \times 10^{-3}$. And direct simulations of the chain with $p = 0.2,\ b = b_0(0.2)$ give $\sigma^2(0.2, b_0(0.2)) \approx 14 \times 10^{-3}$.

References

[1] D.J. Aldous. Greedy search on the binary tree with random edge-weights. *Combinatorics, Probability and Computing*, 1:281–293, 1992.

[2] J.D. Biggins. The growth and spread of the general branching random walk. *Ann. Appl. Probab.*, 5:1008–1024, 1995.

[3] J.D. Biggins, B.D. Lubachevsky, A. Shwartz, and A. Weiss. A branching random walk with a barrier. *Ann. Appl. Probab.*, 1:573–581, 1991.

[4] P. Diaconis and L. Saloff-Coste. What do we know about the Metropolis algorithm? *J. Comput. System Sci.*, xxx:xxx, 1997.

[5] P.G. Doyle and J.L. Snell. *Random Walks and Electrical Networks.* Mathematical Association of America, Washington DC, 1984.

[6] D.H. Greene and D.E. Knuth. *Mathematics for the Analysis of Algorithms.* Birkhauser, 1981.

[7] B.D. Hughes. *Random Walks and Random Environments. Volume 2. Random Environments.* Oxford University Press, 1996.

[8] S. Janson. Random regular graphs: Asymptotic distributions and contiguity. *Combin. Probab. Comput.*, 4:369–405, 1995.

[9] M. Jerrum and G. Sorkin. Simulated annealing for graph bisection. In *Proc. 34th IEEE Symp. Found. Comp. Sci.*, pages 94–103, 1993.

[10] A. Juels. *Topics in Black-Box Combinatorial Optimization*. PhD thesis, C.S. Dept, U.C. Berkeley, 1996.

[11] H. Kesten. Branching Brownian motion with absorption. *Stochastic Process. Appl.*, 7:9–47, 1978.

[12] C. Kipnis and S.R.S. Varadhan. Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Comm. Math. Phys.*, 104:1–19, 1986.

[13] J.L. Lebowitz and H. Rost. The Einstein relation for the displacement of a test particle in a random environment. *Stochastic Process. Appl.*, 54:183–196, 1994.

[14] T.M. Liggett. *Interacting Particle Systems*. Springer-Verlag, 1985.

[15] R. Lyons and R. Pemantle. Random walk in a random environment and first-passage percolation on trees. *Ann. Probab.*, 20:125–136, 1992.

[16] R. Lyons, R. Pemantle, and Y. Peres. Ergodic theory on Galton-Watson trees: Speed of random walk and dimension of harmonic measure. *Ergodic Theory Dynamical Systems*, 15:593–619, 1995.

[17] R. Lyons, R. Pemantle, and Y. Peres. Biased random walks on Galton-Watson trees. *Probab. Th. Rel. Fields*, 106:249–264, 1996.

[18] R. Lyons, R. Pemantle, and Y. Peres. Unsolved problems concerning random walks on trees. In K. Athreya and P. Jagers, editors, *Classical and Modern Branching Processes*, pages 223–238. Springer-Verlag, 1996.

[19] A. De Masi, P. A. Ferrari, S. Goldstein, and W. D. Wick. An invariance principle for reversible Markov processes: Applications to random motions in random environments. *J. Stat. Phys.*, 55:787–855, 1989.

[20] R. Pemantle. Tree-indexed processes. *Statistical Science*, 10:200–213, 1995.