Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2010, Vol. 46, No. 4, 1080–1092
DOI: 10.1214/09-AIHP345

ANNALES DE L'INSTITUT HENRI POINCARÉ
PROBABILITÉS ET STATISTIQUES
www.imstat.org/aihp

Entropy of random walk range

Itai Benjamini[a], Gady Kozma[a], Ariel Yadin[b] and Amir Yehudayoff[c]

[a]*Faculty of Mathematics and Computer Science, The Weizmann Institute of Science, Rehovot 76100, Israel.*
E-mails: itai.benjamini@weizmann.ac.il; gady.kozma@weizmann.ac.il
[b]*DPMMS, Centre for Mathematical Sciences, University of Cambridge, Cambridge CB3 0WB, UK. E-mail: A.Yadin@statslab.cam.ac.uk*
[c]*School of Mathematics, Institute for Advanced Study, Princeton, NJ 08540, USA. E-mail: amir.yehudayoff@gmail.com*

Received 27 April 2009; revised 30 September 2009; accepted 2 October 2009

Abstract. We study the entropy of the set traced by an n-step simple symmetric random walk on $\mathbb{Z}^d$. We show that for $d \geq 3$, the entropy is of order n. For $d = 2$, the entropy is of order $n/\log^2 n$. These values are essentially governed by the size of the boundary of the trace.

Résumé. Nous étudions l'entropie de la trace d'une marche aléatoire simple et symétrique de longueur n sur $\mathbb{Z}^d$. Nous montrons que si $d \geq 3$, cette entropie est d'ordre n, tandis que pour $d = 2$ elle est d'ordre $n/\log^2 n$. Ces valeurs proviennent essentiellement de la taille de la frontière de la trace.

MSC: 82C41; 94A17

Keywords: Random walk; Entropy

1. Introduction

A natural observable of a random walk is its *range*, the set of positions it visited. In this note we study the entropy of this range, which is roughly the number of bits of information needed in order to describe it. We calculate the entropy of the range of a random walk on $\mathbb{Z}^d$, $d \in \mathbb{N}$, up to constant factors.

1.1. *Main result*

Let $S(0), \ldots, S(n)$ be a simple symmetric random walk on $\mathbb{Z}^d$, $d \in \mathbb{N}$, of length n. Define the *range* of the random walk to be

$$R(n) = \big\{S(0), S(1), \ldots, S(n)\big\},$$

the set of vertices visited by the walk.

In this note we study the entropy of $R(n)$ as a function of n (for formal definition of entropy, see Section 2.1). We calculate the value of the entropy, $H(R(n))$, up to constants, precisely:

Theorem 1. *For $d = 2$ there exist constants $c_2, C_2 > 0$ such that for all $n \in \mathbb{N}$,*

$$c_2 \frac{n}{\log^2(n)} \leq H\big(R(n)\big) \leq C_2 \frac{n}{\log^2(n)},$$

and for $d \geq 3$ there exist constants $c_d, C_d > 0$ such that for all $n \in \mathbb{N}$,

$$c_d n \leq H\big(R(n)\big) \leq C_d n.$$

The proof of Theorem 1 is organized as follows: We first prove the lower bound which is easier and follows directly from estimates on the size of the boundary of the range; in two dimensions the boundary of the range of the walk is of order $n/\log^2 n$, and in higher dimensions it is linear in n. This is done in Section 2.2. We then show the upper bounds. The proof for dimensions greater than two is in Section 2.3. The proof for dimension two, which requires a certain renormalization argument, appears in Section 2.4. An interesting feature of the procedure is that at each step of the renormalization process, the number of "active" boxes is not determined by examining the previous renormalization step, but rather globally. For a more detailed discussion of this procedure see Section 2.4.2.

The one-dimensional case is not difficult.

Exercise
In the case $d = 1$, there exist constants $c_1, C_1 > 0$ such that for all $n \in \mathbb{N}$,

$$c_1 \log n \leq H\big(R(n)\big) \leq C_1 \log n.$$

2. Entropy of random walk

2.1. *Entropy*

Here we provide some background on entropy. Let X be a random variable taking values in an arbitrary finite set Ω. For $x \in \Omega$, let $p(x)$ be the probability that $X = x$. The *entropy* of X is defined as $H(X) = \mathbb{E}[-\log p(X)]$ (all logarithms in this note are base 2). For two random variables X and Y, the *conditional entropy* of X conditioned on Y is defined as $H(X|Y) = H(X, Y) - H(Y)$.

Proposition 2. *The following relations hold*:

(i) $0 \leq H(X) \leq \log |\Omega|$.
(ii) *For every function* f, $H(f(X)|X) = 0$.
(iii) $H(X) \leq H(Y) + H(X|Y)$.

For more information on entropy and for proofs of these properties see, e.g., [1], Chapter 2.

2.2. *Lower bound*

Notation
By $\mathbb{P}_z$ and $\mathbb{E}_z$ we denote the probability measure and expectation of the random walk conditioned on $S(0) = z$. We denote $\mathbb{P} = \mathbb{P}_0$ and $\mathbb{E} = \mathbb{E}_0$. Let $z, w \in \mathbb{Z}^d$ and $A \subset \mathbb{Z}^d$. Denote by $\operatorname{dist}(z, w)$ the graph distance between z and w in $\mathbb{Z}^d$. Denote $\operatorname{dist}(z, A) = \inf\{\operatorname{dist}(z, a)\colon\ a \in A\}$. We write $z \sim w$ if $\operatorname{dist}(z, w) = 1$, and $z \sim A$ if $\operatorname{dist}(z, A) = 1$. The *inner boundary* of A is defined as

$$\partial A = \big\{z \in A\colon\ z \sim \mathbb{Z}^d \setminus A\big\}.$$

Let $p_n(A) = \mathbb{P}[R(n) = A]$.

Lemma 3. *For every $A \subset \mathbb{Z}^d$,*

$$p_n(A) \leq \left(1 - \frac{1}{2d}\right)^{|\partial A| - 1}.$$

Proof. Let $T_0 = 0$ and define inductively for $j \geq 1$,

$$T_j = \inf\bigl\{t \geq T_{j-1} + 1\colon\ S(t) \in \partial A\bigr\}.$$

By the strong Markov property, for any $0 < j < |\partial A|$,

$$\mathbb{P}\bigl[S(T_j + 1) \notin A \mid S(0), \ldots, S(T_j), T_j < \infty\bigr] \geq \frac{1}{2d}.$$

The event $A \subseteq R(n)$ implies that $T_j \leq n$ for all $j \leq |\partial A|$. The event $R(n) \subseteq A$ implies that $S(T_j + 1) \in A$ for all $j \leq |\partial A| - 1$. Let E_j be the event that $S(T_j + 1) \in A$ and $T_{j+1} < \infty$. Thus,

$$\mathbb{P}\bigl[R(n) = A\bigr] \leq \mathbb{P}\left[\bigcap_{j=1}^{|\partial A|-1} E_j\right] \leq \prod_{j=1}^{|\partial A|-1} \mathbb{P}[E_j \mid E_1, \ldots, E_{j-1}] \leq \left(1 - \frac{1}{2d}\right)^{|\partial A|-1}.$$

□

Lemma 3 shows that in order to bound the entropy of the random walk trace from below, it is enough to bound the expected value of the size of the inner boundary of the random walk trace from below. More precisely, we have:

Corollary 4. $H(R(n)) \geq -\log(1 - \frac{1}{2d}) \cdot \mathbb{E}[|\partial R(n)| - 1]$.

The following lemma gives the lower bound for the entropy of the random walk trace.

Lemma 5. *For any $d \geq 2$, there exists a constant $c_d > 0$ such that for all $n \in \mathbb{N}$,*

$$H\bigl(R(n)\bigr) \geq \begin{cases} c_2 \frac{n}{\log^2(n)}, & d = 2, \\ c_d n, & d \geq 3. \end{cases}$$

Proof. By Corollary 4, it suffices to show that

$$\mathbb{E}\bigl[\bigl|\partial R(n)\bigr|\bigr] \geq \begin{cases} c_2 \frac{n}{\log^2(n)}, & d = 2, \\ c_d n, & d \geq 3 \end{cases}$$

for some constants $c_d > 0$. For $z \in \mathbb{Z}^d$, define $T_z = \inf\{t \geq 0\colon\ S(t) = z\}$. By Lemma 19.1 of [4], and by the transience of the random walk for $d \geq 3$, there exist constants $c_d > 0$ such that for any $z \sim w \in \mathbb{Z}^d$,

$$\mathbb{P}_z[T_w > n] \geq \begin{cases} \frac{c_2}{\log n}, & d = 2, \\ c_d, & d \geq 3. \end{cases}$$

Denote the right-hand side of the above inequality by $f_d(n)$. Using the strong Markov property at time T_z, for any $z \sim w \in \mathbb{Z}^d$,

$$\mathbb{P}\bigl[z \in \partial R(n)\bigr] \geq \mathbb{P}[T_z \leq n, T_w > n] \geq f_d(n)\mathbb{P}[T_z \leq n].$$

This proves the lemma, since

$$\mathbb{E}\bigl[\bigl|\partial R(n)\bigr|\bigr] \geq f_d(n) \sum_{z \in \mathbb{Z}^d} \mathbb{P}[T_z \leq n] = f_d(n)\mathbb{E}\bigl[\bigl|R(n)\bigr|\bigr],$$

and since

$$\mathbb{E}\bigl[\bigl|R(n)\bigr|\bigr] \geq \begin{cases} c_2' \cdot \frac{n}{\log n}, & d = 2, \\ c_d' n, & d \geq 3 \end{cases}$$

for some constants $c_d' > 0$ (see, e.g., Theorem 20.1 in [4]). □

2.3. *Upper bound*

We now show that the lower bounds on the entropy of the random walk trace given by Lemma 5 are correct up to a constant. The transient case is much simpler than the two-dimensional case.

Proposition 6. *For $d \geq 3$, there exists a constant $C_d > 0$ such that for all $n \in \mathbb{N}$,*

$$H\big(R(n)\big) \leq C_d \cdot n.$$

Proof. Let $\Omega = \{A \subset \mathbb{Z}^d\colon\ p_n(A) > 0\}$. By clause (i) of Proposition 2 it suffices to prove that $|\Omega| \leq (2d)^n$. This follows from the fact that the number of possible n-step trajectories in $\mathbb{Z}^d$ starting at 0 is $(2d)^n$. □

2.4. *Two dimensions*

We now turn to the two-dimensional case, which is more elaborate.

For $z \in \mathbb{Z}^2$, we denote by $\|z\|$ the L^2-norm of z. Define

$$T_{z,r} = \inf\big\{t \geq 0 : \big\|S(t) - z\big\| \leq r\big\},$$

and $T_r = T_{\vec{0},r}$. Also set

$$\tau_{z,r} = \inf\big\{t \geq 0\colon\ \big\|S(t) - z\big\| \geq r\big\},$$

and $\tau_r = \tau_{\vec{0},r}$.

2.4.1. *Probability estimates*

We begin with some classical probability estimates regarding the random walk on $\mathbb{Z}^2$, which we include for the sake of completeness.

Lemma 7. *There exists a constant $C > 0$ such that for all $n \in \mathbb{N}$,*

$$\mathbb{E}\Big[\max_{0\leq k\leq n} \big\|S(k)\big\|^2\Big] \leq Cn.$$

Proof. Let $S(k) = (X(k), Y(k))$, so $\|S(k)\|^2 = |X(k)|^2 + |Y(k)|^2$. Doob's maximal inequality (see, e.g., [5], Chapter II) on the martingale $X(k)$ tells us that

$$\mathbb{E}\Big[\max_{0\leq k\leq n} \big|X(k)\big|^2\Big] \leq 4\mathbb{E}\big[\big|X(n)\big|^2\big].$$

The martingale $|X(k)|^2 - k/2$ tells us that $\mathbb{E}[|X(n)|^2] = n/2$, which completes the proof, since $X(k)$ and $Y(k)$ have the same distribution. □

Lemma 8. *There exist constants $c_1, c_2 > 0$ such that for all $n \in \mathbb{N}$ and $\lambda > 0$,*

$$\mathbb{P}\Big[\max_{1\leq j\leq n} \big\|S(j)\big\| \geq \lambda\Big] \leq c_1 \cdot \exp\left(-c_2 \frac{\lambda^2}{n}\right).$$

Proof. This is a consequence of Theorem 2.13 in [4]. □

Lemma 9. *There exists a constant $c > 0$ such that the following holds. Let $T = T_{\vec{0},0}$. Then, for $z \in \mathbb{Z}^2$ and $r \geq 2\|z\|$,*

$$\mathbb{P}_z[T \leq \tau_r] \geq \frac{c \log(r/\|z\|)}{\log r}.$$

Proof. Let $a:\mathbb{Z}^2 \to [0,\infty)$ be the *potential kernel* defined in Chapter 1.6 of [2]. That is, $a(0)=0$, $a(\cdot)$ is harmonic in $\mathbb{Z}^2 \setminus \{0\}$, and there exist constants $c_1, c_2 > 0$ such that for any $z \in \mathbb{Z}^2 \setminus \{0\}$, $a(z) = c_1 \log \|z\| + c_2 + \mathrm{O}(\|z\|^{-2})$. Since $a(\cdot)$ is harmonic in $\mathbb{Z}^2 \setminus \{0\}$, if $r > \|z\|$ then $a(S(t))$ is a martingale up to time $T' = \min\{T, \tau_r\}$. Thus,

$$a(z) = \big(1 - \mathbb{P}_z[T \le \tau_r]\big) \cdot \mathbb{E}_z\big[a\big(S\big(T'\big)\big) \mid T > \tau_r\big],$$

which implies

$$\mathbb{P}_z[T \le \tau_r] \ge 1 - \frac{c_1 \log \|z\| + c_2 + \mathrm{O}(\|z\|^{-2})}{c_1 \log r + c_2 + \mathrm{O}(r^{-2})}.$$

□

We also need an upper bound.

Lemma 10. *There exists a constant $C > 0$ such that for every $z \in \mathbb{Z}^2$ and r, R such that $1 \le r \le \frac{1}{2}\|z\| \le \frac{1}{4}R$,*

$$\mathbb{P}_z[T_r \le \tau_R] \le C \cdot \frac{\log(R/\|z\|)}{\log(R/r)}.$$

Proof. Using the potential kernel from the proof of Lemma 9 with the stopping time $\min\{T_r, \tau_R\}$, there exists a constant $c_1 > 0$ such that

$$\begin{aligned}\mathbb{P}_z[T_r \le \tau_R] &\le \frac{c_1(\log R - \log\|z\|) + \mathrm{O}(R^{-1} + \|z\|^{-2})}{c_1(\log R - \log r) + \mathrm{O}(r^{-2})} \\ &\le C \cdot \frac{\log(R/\|z\|)}{\log(R/r)}\end{aligned}$$

for some constant $C > 0$. □

Lemma 11. *For any $0 < \alpha < 1$, there exists a constant $C > 0$ such that the following holds. Let $z \in \mathbb{Z}^2$ such that $\|z\| \ge 1/\alpha$. Then for any $n \in \mathbb{N}$ such that $n > \|z\|^4$,*

$$\mathbb{P}_z[T_{\alpha\|z\|} \ge n] \le \frac{C}{\log(n/\|z\|^4)}.$$

Proof. By adjusting the constant, we can assume without loss of generality that $n/\|z\|^4$ is large enough. Let $r = \alpha\|z\|$ and $R = n^{1/4}$. Using the potential kernel from the proof of Lemma 9 with the stopping time $T' = \min\{T_r, \tau_R\}$,

$$\mathbb{P}_z[T_r \ge \tau_R] \le \frac{c_1 \log(\|z\|/r) + \mathrm{O}(r^{-1})}{c_1 \log(R/r) + \mathrm{O}(r^{-1})} \le \frac{C_1}{\log(n/\|z\|^4)} \tag{2.1}$$

for some constant $C_1 = C_1(\alpha) > 0$ independent of z and n. Also, considering the martingale $\|S(t)\|^2 - t$ up to time τ_R shows that $\mathbb{E}_z[\tau_R] \le (R+1)^2$. Thus, by Markov's inequality,

$$\mathbb{P}_z[\tau_R > n] \le \frac{4}{\sqrt{n}}. \tag{2.2}$$

Equations (2.1) and (2.2) together prove the proposition, since

$$\mathbb{P}_z[T_r \ge n] \le \mathbb{P}_z[T_r \ge \tau_R] + \mathbb{P}_z[\tau_R > n].$$

□

Lemma 12. *There exists a constant $C > 0$ such that for all $n \in \mathbb{N}$ and $1 \le r \le \frac{1}{2}\sqrt{n}$ the following holds. Let $z \in \mathbb{Z}^2$ be such that $\|z\| \ge \sqrt{n}$. Then,*

$$\mathbb{P}_z[T_r \le n] \le \frac{C}{\log(n/r^2)}.$$

Proof. For $m \geq 1$, let A_m be the event $\{\tau_{m\|z\|} < T_r \leq \tau_{(m+1)\|z\|} \leq n\}$. The family $\{A_m\}$ consists of pairwise disjoint events, and

$$\mathbb{P}_z[T_r \leq n] \leq \sum_{m=1}^{\infty} \mathbb{P}[A_m].$$

For every $m \geq 1$, using the strong Markov property at time $\tau_{m\|z\|}$,

$$\mathbb{P}_z[A_m] \leq \mathbb{P}_z[\tau_{m\|z\|} \leq n] \cdot \max\bigl\{\mathbb{P}_x[T_r \leq \tau_{(m+1)\|z\|}]\colon\ m\|z\| \leq \|x\| \leq m\|z\| + 1\bigr\}.$$

By Lemma 8, there exist constants $C_1, c_2 > 0$ such that

$$\mathbb{P}_z[\tau_{m\|z\|} \leq n] \leq \mathbb{P}_z\Bigl[\max_{1\leq j\leq n} \bigl\|S(j)\bigr\| \geq m\|z\| - \|z\|\Bigr] \leq C_1 \exp\bigl(-c_2 m^2\bigr).$$

By Lemma 10, for any $x \in \mathbb{Z}^2$ such that $m\|z\| \leq \|x\| \leq m\|z\| + 1$,

$$\mathbb{P}_x[T_r \leq \tau_{(m+1)\|z\|}] \leq \mathbb{P}_x[T_r \leq \tau_{2(m\|z\|+1)}] \leq \frac{c_3}{\log(n/r^2)}$$

for some constant $c_3 > 0$. Summing over all $m \geq 1$,

$$\mathbb{P}_z[T_r \leq n] \leq \frac{c_3}{\log(n/r^2)} \sum_{m=1}^{\infty} c_1 \exp\bigl(-c_2 \cdot m^2\bigr).$$

□

2.4.2. *Upper bound in two dimensions*

The general scheme of the proof is quite simple. Here is how we efficiently store the information (this implies that the entropy cannot be too big). Divide the relevant area, which is roughly a $\sqrt{n} \times \sqrt{n}$ box, to smaller blocks. Each of these blocks can have three states: empty (most common), full, or partial (least common). Record the state of all blocks. The empty and full blocks need not be considered any further. The partial blocks are further divided into sub-blocks, and recorded recursively. We will show that this scheme requires an average of $\approx n/\log^2 n$ bits, which will finish the proof.

Hence the crucial question is: how many partial blocks are there? Consider $k \times k$ blocks. For a block to be partial, the random walk needs to first hit it (this "costs" $1/\log n$), and then escape before covering the block completely. It is well known that the cover time of a $k \times k$ square is order $k^2 \log^2 k$, that is, it requires about $\log^2 k$ "visits" to the block (if you are unfamiliar with the notion of a visit, examine the proof of Lemma 15, a visit is the time interval $[\tau_j, \sigma_j]$). Thus, the random walk needs to "escape" from our block, this costs another factor of $1/\log n$, but it has about $\log^2 k$ "attempts" to do so. To conclude, the final probability of a typical block to be partial is order $\log^2 k/\log^2 n$.

This argument is spelt out in Lemmas 15 and 17 below. Lemma 15 contains the calculation above except for the very first $1/\log n$ term, since there the starting point is close to the block. Lemma 17 contains the full calculation, with the main problem being estimating separately blocks which are close by and far away (Lemma 13 helps with the far away blocks).

Let us use this opportunity to repeat a point already made in the Introduction. Our blocks have a typical hierarchical structure, with the blocks of level j contained in blocks of level $j+1$. Naïvely, at level j the blocks we are interested in are blocks partially covered by the random walk, which are contained in partial block at level $j+1$, which are themselves contained in partial blocks at level $j+2$, etc. So it seems that estimating the number of such blocks requires going through this hierarchy. But it does not, because once a block is partially covered by the random walk, we get this property automatically for all its super-blocks. We can thus estimate the number of partially covered blocks in level j directly. This simplifies the proof significantly.

We move to the details of the proof. For $z \in \mathbb{Z}^2$ and $k \in \mathbb{N}$, let

$$Q(z,k) = \bigl\{z + (j, j')\colon\ -k \leq j, j' \leq k\bigr\};$$

i.e., $Q(z,k)$ is the square of side length $2k+1$ centered at z. For a path $x(0), x(1), \ldots, x(n)$ in $\mathbb{Z}^2$, we denote by $x[s,t]$ the path $x(s), x(s+1), \ldots, x(t)$.

Lemma 13. *There exist constants $c, C > 0$ such that for all $n, k \in \mathbb{N}$ such that $k \le n^{1/4}$, and all $z \in \mathbb{Z}^d$ such that $\|z\| \ge 5\sqrt{n}$,*

$$\mathbb{P}\big[R(n) \cap Q(z,k) \ne \varnothing\big] \le \frac{C}{\log n} \cdot \exp\left(-c\frac{\|z\|^2}{n}\right).$$

Proof. Let $\lambda = \|z\| - 2\sqrt{n}$. Let T be the first time the walk $S(\cdot)$ started at 0 hits $Q(z,k)$. Then $\tau_\lambda < T_{z,2k} < T$. By Lemmas 8 and 12,

$$\begin{aligned}\mathbb{P}\big[R(n) \cap Q(z,k) \ne \varnothing\big] &\le \mathbb{P}[\tau_\lambda \le n] \cdot \max\big\{\mathbb{P}_x[T_{z,2k} \le n]\colon\ \lambda \le \|x\| \le \lambda + 1\big\} \\ &\le \mathbb{P}\Big[\max_{1\le j\le n} \big\|S(j)\big\| \ge \lambda\Big] \cdot \frac{c_1}{\log n} \\ &\le \frac{c_2}{\log n} \cdot \exp\left(-c_3\frac{\|z\|^2}{n}\right)\end{aligned}$$

for some constants $c_1, c_2, c_3 > 0$. □

Lemma 14. *There exists a constant $C > 0$ such that the following holds. For all $n, k \in \mathbb{N}$ such that $k \le n^{1/4}$, and all $z \in \mathbb{Z}^d$ such that $1 \le \|z\| < 5\sqrt{n}$,*

$$\mathbb{P}\big[R(n) \cap Q(z,k) \ne \varnothing\big] \le C \cdot \frac{\log(10\sqrt{n}/\|z\|)}{\log n}.$$

Proof. By adjusting the constant, we can assume without loss of generality that $\|z\| \ge 3k$. Let $Q = Q(z,k)$. Define $\sigma_0 = 0$, and for $i \ge 1$, define

$$\sigma_i = \tau_{10^i\sqrt{n}} = \inf\big\{t \ge 0\colon\ \big\|S(t)\big\| \ge 10^i\sqrt{n}\big\}.$$

The event $\{R(n) \cap Q \ne \varnothing\}$ is contained in the event

$$\big\{S[0,\sigma_1] \cap Q \ne \varnothing\big\} \cup \bigcup_{i\ge 1}\big\{S[\sigma_i,\sigma_{i+1}] \cap Q \ne \varnothing, \sigma_i \le n\big\}.$$

Since $3k \le \|z\| < 5\sqrt{n}$, we have that on the event $\{S[0,\sigma_1] \cap Q \ne \varnothing\}$, the random walk started at 0 enters the ball of radius $2k$ around z before exiting the ball of radius $20\sqrt{n}$ around z. Translating by minus z we get by Lemma 10 that there exists a constant $C_1 > 0$ such that

$$\mathbb{P}\big[S[0,\sigma_1] \cap Q \ne \varnothing\big] \le \mathbb{P}_{-z}[T_{2k} \le \tau_{20\sqrt{n}}] \le C_1 \cdot \frac{\log(10\sqrt{n}/\|z\|)}{\log n}.$$

Fix $i \ge 1$. By Lemma 8,

$$\mathbb{P}[\sigma_i \le n] \le \mathbb{P}\Big[\max_{0\le j\le n} \big\|S(j)\big\| \ge 10^i\sqrt{n}\Big] \le C_2 \cdot \exp\big(-C_3 \cdot 10^{2i}\big)$$

for some constants $C_2, C_3 > 0$. Using Lemma 10 again,

$$\mathbb{P}\big[S[\sigma_i,\sigma_{i+1}] \cap Q \ne \varnothing \mid \sigma_i \le n\big] \le \frac{C_4}{\log n}$$

for some constant $C_4 > 0$. Therefore,

$$\mathbb{P}\big[R(n) \cap Q \ne \varnothing\big] \le C_1 \cdot \frac{\log(10\sqrt{n}/\|z\|)}{\log n} + \frac{C_2 \cdot C_4}{\log n}\sum_{i\ge 1}\exp\big(-C_3 \cdot 10^{2i}\big).$$

□

We have reached the main geometric lemma.

Lemma 15. *There exists a constant $C > 0$ such that the following holds. Let $n, k \in \mathbb{N}$, let $Q = Q(0,k)$ and let $z \sim Q$. Then,*

$$\mathbb{P}_z\big[\partial R(n) \cap Q \neq \varnothing\big] \leq C \cdot \frac{\log^2 k}{\log n}.$$

Proof. Without loss of generality assume that $\log^2 k \leq \log n$. Define $Q^+ = Q(0, k+1)$. So Q^+ contains the union of Q with all vertices that are adjacent to Q. Define $\tau_0 = 0$, and inductively

$$\sigma_j = \inf\big\{t \geq \tau_j \colon \big\|S(t)\big\| \geq 10k\big\},$$
$$\tau_{j+1} = \inf\big\{t \geq \sigma_j \colon S(t) \in Q^+\big\}.$$

If $Q^+ \subseteq R(n)$ then $\partial R(n) \cap Q = \varnothing$. Thus, it suffices to bound from above the probability of the event $\{Q^+ \not\subset R(n)\}$. It is convenient to choose $m = \lceil \log k \cdot \log n \rceil$, as will soon be clear. Set $V_j = \{\sigma_{j+1} - \sigma_j \geq \frac{n}{2m}\}$ and $U_j = \{Q^+ \not\subset R(\sigma_j)\}$. We prove the following inclusion of events

$$\big\{Q^+ \not\subset R(n)\big\} \subseteq \{\sigma_0 \geq n/2\} \cup U_m \cup \bigcup_{j=0}^{m-1} (U_j \cap V_j). \tag{2.3}$$

Assume that the event on the right-hand side of (2.3) does not occur; i.e., assume that $\sigma_0 < n/2$, that $\overline{U_m}$, and that for all $0 \leq j \leq m-1$, $\overline{U_j} \cup \overline{V_j}$. Let $J = \min\{0 \leq j \leq m \colon \overline{U_j}\}$. Consider the following cases:

- Case 1: $J = 0$. Then $Q^+ \subset R(\sigma_0)$. Since $\sigma_0 < n/2$, we get that $Q^+ \subset R(n)$.
- Case 2: $J > 0$. Since we assumed that $\overline{U_m}$, we know that $1 \leq J \leq m$. By the assumption $\bigcap_{j=0}^{m-1} (\overline{U_j} \cup \overline{V_j})$, we have that $\sigma_{j+1} - \sigma_j < n/2m$, for all $0 \leq j \leq J-1$. Since we assumed that $\sigma_0 < n/2$, we get that

$$\sigma_J = \sigma_0 + \sum_{j=0}^{J-1} \sigma_{j+1} - \sigma_j < n.$$

But J was chosen so that $\overline{U_J}$ occurs, so $Q^+ \subset R(\sigma_J) \subset R(n)$. This proves (2.3).

Fix $j \geq 0$. Since $\|S(t) - z\|^2 - t$ is a martingale, we have that $\mathbb{E}_z[\sigma_j - \tau_j \mid \mathcal{F}(\tau_j)] \leq C_1 k^2$ for some constant $C_1 > 0$. Using Markov's inequality,

$$\mathbb{P}_z\left[\sigma_j - \tau_j \geq \frac{n}{4m} \;\middle|\; \mathcal{F}(\tau_j)\right] \leq \frac{C_2 m k^2}{n} \tag{2.4}$$

for some constant $C_2 > 0$. By Lemma 11, there exists a constant $C_3 > 0$ such that

$$\mathbb{P}_z\left[\tau_{j+1} - \sigma_j \geq \frac{n}{4m} \;\middle|\; \mathcal{F}(\sigma_j)\right] \leq \frac{C_3}{\log n}. \tag{2.5}$$

The two inequalities, (2.4) and (2.5), imply that

$$\mathbb{P}_z\big[V_j \mid \mathcal{F}(\sigma_j)\big] \leq \frac{C_4}{\log n} \tag{2.6}$$

for some constant $C_4 > 0$. Using Lemma 9, there exists a universal constant $C_5 > 0$ such that for any $x \in Q^+$,

$$\mathbb{P}_z\big[x \in S[\tau_j, \sigma_j] \mid \mathcal{F}(\tau_j)\big] \geq \frac{C_5}{\log k}.$$

Thus,

$$\begin{aligned}\mathbb{P}_z[U_j] &= \mathbb{P}_z\big[Q^+ \not\subset R(\sigma_j)\big] \le \min\big\{1, \big|Q^+\big| \cdot (1 - C_5/\log k)^{j+1}\big\} \\ &\le \min\big\{1, C_6 k^2 \exp\big(-C_5(j+1)/\log k\big)\big\} \end{aligned} \tag{2.7}$$

for some constant $C_6 > 0$. Plugging (2.4), (2.6) and (2.7) into (2.3) yields

$$\begin{aligned}\mathbb{P}_z\big[Q^+ \not\subset R(n)\big] &\le \mathbb{P}_z[\sigma_0 \ge n/2] + \mathbb{P}_z[U_m] + \sum_{j=0}^{K} \mathbb{P}_z[U_j \cap V_j] + \sum_{j>K} \mathbb{P}_z[U_j \cap V_j] \\ &\le C_7\left(\frac{k^2}{n} + n^{-C_8} + \sum_{j=0}^{K} \frac{1}{\log n} + \sum_{j>K} \frac{k^2 \exp(-C_5(j+1)/\log k)}{\log n}\right) \le \frac{C_9 \log^2 k}{\log n}, \end{aligned} \tag{2.8}$$

where $K = \lceil 4\log^2 k / C_5 \rceil$ and $C_7, C_8, C_9 > 0$ are constants. □

Definition 16. *Define $\Lambda(k) = \{(2k+1)z\colon\ z \in \mathbb{Z}^2\}$. The collection $\{Q(z,k)\}_{z\in\Lambda(k)}$ consists of disjoint squares that cover $\mathbb{Z}^2$. For $k, n \in \mathbb{N}$ and $z \in \mathbb{Z}^2$, define $I(z,k,n)$ to be the indicator function of the event $\{\partial R(n) \cap Q(z,k) \ne \varnothing\}$. Define*

$$M(k,n) = \sum_{z\in\Lambda(k)} I(z,k,n),$$

i.e. the number of squares that intersect $\partial R(n)$.

Lemma 17. *There exists a constant $C > 0$ such that for every $k, n \in \mathbb{N}$,*

$$\mathbb{E}\big[M(k,n)\big] \le C \cdot \max\left\{1, \frac{n}{k^2} \cdot \frac{\log^2 k}{\log^2 n}\right\}.$$

Proof. Fix $k, n \in \mathbb{N}$. For $z \in \mathbb{Z}^2$, the event $\{\partial R(n) \cap Q(z,k) \ne \varnothing\}$ implies the event

$$\left\{\max_{0\le j\le n} \big\|S(j)\big\| \ge \|z\| - \sqrt{2}(k+1)\right\}.$$

We start with an a priori bound. Using Lemma 7, there exist constants $C_1, C_2 > 0$ so that

$$\begin{aligned}\mathbb{E}\big[M(n,k)\big] &\le \sum_{z\in\Lambda(k)} \mathbb{P}\Big[\|z\| \le \max_{0\le j\le n} \big\|S(j)\big\| + \sqrt{2}(k+1)\Big] \\ &\le \mathbb{E}\Big[\Big|\Big\{z \in \Lambda(k)\colon\ \|z\| \le \max_{0\le j\le n} \big\|S(j)\big\| + \sqrt{2}(k+1)\Big\}\Big|\Big] \\ &\le C_1 \cdot \max\Big\{1, k^{-2} \cdot \mathbb{E}\Big[\max_{0\le j\le n} \big\|S(j)\big\|^2\Big]\Big\} \le C_2 \cdot \max\left\{1, \frac{n}{k^2}\right\}, \end{aligned}$$

where the third inequality holds as for every $R > 0$, the size of $\{z \in \Lambda(k)\colon\ \|z\| \le R\}$ is at most a constant times $\max\{1, k^{-2} \cdot R^2\}$. Thus, we can assume without loss of generality that $k < k+1 \le (n - \sqrt{n})^{1/4} \le n^{1/4}$.

Let

$$\tau_Q(z) = \inf\big\{t \ge 0\colon\ S(t) \in Q(z, k+1)\big\}$$

(we use that fact that $Q(z,k+1)$ is 'bigger' than $Q(z,k)$), and let

$$J(z,k,n) = \mathbf{1}_{\{\tau_Q(z) \le n - \sqrt{n}\}} \cdot I(z,k,n).$$

For all $z \in \Lambda(k)$, a.s.

$$I(z,k,n) \le \mathbf{1}_{\{n-\sqrt{n}<\tau_Q(z)\le n\}} + J(z,k,n).$$

Summing over all $z \in \Lambda(k)$, a.s.

$$M(n,k) \le 4\sqrt{n} + \sum_{z\in\Lambda(k)} J(z,n,k). \tag{2.9}$$

By the strong Markov property at time $\tau_Q(z)$ and Lemma 15, there exists a constant $C_3 > 0$ such that a.s.

$$\mathbb{P}\big[\partial R(n) \cap Q(z,k) \ne \varnothing \mid \tau_Q(z) \le n - \sqrt{n}\big] \le C_3 \cdot \frac{\log^2 k}{\log n}. \tag{2.10}$$

By Lemma 14, as $k < n^{1/4}$, there exists a constant $C_4 > 0$ such that for all $z \in \mathbb{Z}^d$ with $1 \le \|z\| < 5\sqrt{n}$,

$$\mathbb{P}\big[\tau_Q(z) \le n - \sqrt{n}\big] \le C_4 \cdot \frac{\log(10\sqrt{n}/\|z\|)}{\log n},$$

which implies

$$\mathbb{P}\big[J(z,k,n)\big] \le C_5 \cdot \frac{\log^2 k}{\log n} \cdot \frac{\log(10\sqrt{n}/\|z\|)}{\log n} \tag{2.11}$$

for some constant $C_5 > 0$.

Denote $\Gamma = 5\sqrt{n}/(2k+1)$. Summing over all $z \in \Lambda(k)$ such that $2 \le \|z\| < 5\sqrt{n}$,

$$\sum_{\substack{z\in\Lambda(k)\\ 2\le\|z\|<5\sqrt{n}}} \log\big(10\sqrt{n}/\|z\|\big) \le \sum_{\substack{x,y\in\mathbb{Z}\\ 2\le x^2+y^2<\Gamma^2}} \log\big(2\Gamma/\sqrt{x^2+y^2}\big) \le C_6\Gamma \sum_{2\le x\le\Gamma} \log(2\Gamma/x) \le C_7\Gamma^2 \tag{2.12}$$

for some constants $C_6, C_7 > 0$. Plugging (2.12) into (2.11), and summing over all $z \in \Lambda(k)$ such that $\|z\| < 5\sqrt{n}$, we get

$$\sum_{z\in\Lambda(k):\ \|z\|<5\sqrt{n}} \mathbb{P}\big[J(z,k,n)\big] \le C_8 \cdot \frac{\log^2 k}{\log^2 n} \cdot \frac{n}{k^2} \tag{2.13}$$

for some constant $C_8 > 0$. In addition, by Lemma 13, there exist constants $C_9, C_{10} > 0$ such that for every $z \in \Lambda(k)$ such that $\|z\| \ge 5\sqrt{n}$,

$$\mathbb{P}\big[\tau_Q(z) \le n - \sqrt{n}\big] \le \frac{C_9}{\log n} \cdot \exp\left(-C_{10}\frac{\|z\|^2}{n}\right),$$

which implies, using (2.10),

$$\mathbb{P}\big[J(z,k,n)\big] \le C_{11} \cdot \frac{\log^2 k}{\log^2 n} \cdot \exp\left(-C_{10}\frac{\|z\|^2}{n}\right)$$

for some constant $C_{11} > 0$. Summing over all $z \in \Lambda(k)$ such that $\|z\| \ge 5\sqrt{n}$,

$$\sum_{z\in\Lambda(k):\ \|z\|\ge5\sqrt{n}} \mathbb{P}\big[J(z,k,n)\big] \le C_{11} \cdot \frac{\log^2 k}{\log^2 n} \sum_{z\in\Lambda(k):\|z\|\ge5\sqrt{n}} \exp\left(-C_{10}\frac{\|z\|^2}{n}\right) \le C_{12} \cdot \frac{\log^2 k}{\log^2 n} \cdot \frac{n}{k^2} \tag{2.14}$$

for some constant $C_{12} > 0$. The lemma follows by (2.9), (2.13) and (2.14). □

For $k < n \in \mathbb{N}$, let $\partial(k,n)$ be the vector $(I(z,k,n))_{z\in\Lambda(k)\cap[-2n,2n]^2}$. Note that

$$M(k,n) = \sum_{z\in\Lambda(k)} I(z,k,n) = \sum_{z\in\Lambda(k)\cap[-2n,2n]^2} I(z,k,n).$$

Lemma 18. *Let $k,\ell,n \in \mathbb{N}$ and let $k' = (2\ell+1)k+\ell$. Then,*

$$H\big(\partial(k,n) \mid \partial\big(k',n\big)\big) \le \mathbb{E}\big[M\big(k',n\big)\big]\cdot(2\ell+1)^2.$$

Proof. For any $z' \in \Lambda(k')$, the square $Q(z',k')$ is of side length $2k'+1 = (2\ell+1)(2k+1)$, and so $Q(z',k')$ is a union of $(2\ell+1)^2$ disjoint squares from the collection $\{Q(z,k)\}_{z\in\Lambda(k)}$.

If $Q(z,k) \subset Q(z',k')$, then $I(z,k,n) \le I(z',k',n)$. Thus, conditioned on the vector $\partial(k',n)$, there are at most $2^{M(k',n)\cdot(2\ell+1)^2}$ possibilities for the vector $\partial(k,n)$. By clause (i) of Proposition 2, and by the definition of conditional entropy, $H(\partial(k,n) \mid \partial(k',n)) \le \mathbb{E}[M(k',n)\cdot(2\ell+1)^2]$. □

Lemma 19. *There exists a constant $C_2 > 0$ such that for all n,*

$$H\big(R(n)\big) \le C_2 \frac{n}{\log^2(n)}.$$

Proof. Since the vector $\partial(0,n)$ determines $R(n)$, clauses (ii) and (iii) of Proposition 2 yield that $H(R(n)) \le H(\partial(0,n))$.

Set $k_0 = 0$, and for $j \ge 0$, define inductively $k_{j+1} = 3k_j + 1$. For every $j \ge 1$, since $3k_j \le k_{j+1} \le 4k_j$, it holds that $\frac{\log k_j}{k_j} \le 9j3^{-j}$. Let $m > 0$ be the smallest j such that $k_j > n$. The vector $\partial(k_m,n)$ is constant, and so its entropy is zero. By Lemmas 17 and 18, for $0 \le j \le m-1$, there exist universal constants $c_2, c_3 > 0$ such that

$$H\big(\partial(k_j,n) \mid \partial(k_{j+1},n)\big) \le c_3 \cdot \max\left\{1, \frac{n}{\log^2 n}\cdot\frac{(j+1)^2}{9^{j+1}}\right\}.$$

Using clause (iii) of Proposition 2, there exists a constant $C > 0$ such that

$$H\big(\partial(0,n)\big) \le \sum_{j=0}^{m-1} H\big(\partial(k_j,n) \mid \partial(k_{j+1},n)\big) + H\big(\partial(k_m,n)\big) \le C\cdot\frac{n}{\log^2 n}.$$

□

Remark 20. *The proof of Lemma 19 shows that provided one can calculate the various conditional probabilities (e.g., with unlimited computational power), one can sample the range of a random walk using only order $n/\log^2 n$ bits.*

3. Concluding remarks and problems for further research

3.1. *Extracting entropy*

Lemma 5 shows that the entropy of $R(n)$ in two dimensions is at least $c_2 n/\log^2 n$. It is interesting to note that one can extract order of $n/\log^2 n$ almost uniformly distributed random bits, by observing a sample of the range. We sketch the construction.

Consider the two configurations that appear in Fig. 1. The walk can only enter the configuration from the "bridge" on the right, so the situation is symmetric to vertical flips. This symmetry implies that conditioned on outside of the configuration, both have the same probability of occurring. Thus, any occurrence of such a configuration in the range of the random walk gives an independent bit, e.g., setting the bit to be 1 if the right configuration occurs, and 0 if the left configuration occurs. Considerations similar to those raised in the proofs above show that the expected number of such configurations is of order $n/\log^2 n$.

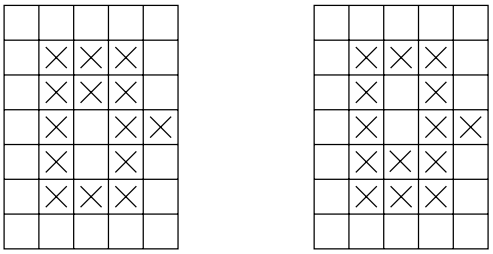

Fig. 1. Two symmetric configurations. ×'s are vertices occupied by the range.

3.2. *Intersection equivalence*

Consider the $n \times n$ square centered at 0 in $\mathbb{Z}^2$, and consider the following procedure. Divide the square into 4 squares of side length $n/2$. Retain each of the squares with probability $1/2$, independently. Continue inductively: at level k, divide each remaining square of side length $n2^{-(k-1)}$ into 4 squares of side length $n2^{-k}$, and retain each one with probability $k/(k+1)$ independently.

This procedure produces a random subset of the $n \times n$ square, denote this set by $Q(n^2)$. In [3], Peres shows that the sets $Q(n^2)$ and $R(n^2)$ are *intersection equivalent*; that is, there exist constants $c, C > 0$ such that for any set $A \subset \mathbb{Z}^2$,

$$c \leq \frac{\mathbb{P}[Q(n^2) \cap A \neq \varnothing]}{\mathbb{P}[R(n^2) \cap A \neq \varnothing]} \leq C$$

from a random starting point. The entropy $H(Q(n^2))$ is of order $n^2/\log^2(n)$, as is $H(R(n^2))$. Note that intersection equivalence does not imply or follow from equal entropy. See [3] for more details.

3.3. *Open questions*

Let G be an infinite graph, and let $\{S(n)\}_{n\geq 0}$ be a simple random walk on G. Let $R(n) = \{S(0), S(1), \ldots, S(n)\}$ be the range of the walk at time n. Let $H(n)$ be the entropy of $R(n)$.

Our results above suggest the following natural question.

- How small can $H(n)$ be in transient graphs? It is possible to construct (spherically symmetric) trees that are transient but have $H(n) = \mathrm{O}(\log^2 n)$. Is it possible to get a smaller entropy?

Note added in proof

David Windisch has some new results on this topic, including an answer to a question posed in previous versions of this paper, see [6].

Acknowledgements

We wish to thank Elchanan Mossel for indicating the configuration in the construction in Section 3.1, and to Eric Shellef for helpful discussions. The material is based upon work supported by the National Science Foundation under agreement No. DMS-0835373. Any opinions, findings and conclusions or recommendations expressed in the material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991. MR1122806

[2] G. F. Lawler. *Intersections of Random Walks*. Springer, New York, 1996. MR1117680

[3] Y. Peres. Intersection-equivalence of Brownian paths and certain branching processes. *Comm. Math. Phys.* **177** (1996) 417–434. MR1384142
[4] P. Révész. *Random Walk in Random and Non-Random Environments*. World Scientific, Hackensack, NJ, 2005. MR2168855
[5] D. Revuz and M. Yor. *Continuous Martingales and Brownian Motion*. Springer, Berlin, 1991. MR1083357
[6] D. Windisch. Entropy of random walk range on uniformly transient and on uniformly recurrent graphs. Preprint. Available at http://arxiv.org/abs/1001.0355.