Ballot theorems for random walks with finite variance

L. Addario-Berry, B.A. Reed

March 5, 2008

Abstract

We prove an analogue of the classical ballot theorem that holds for any mean zero random walk with positive but finite variance. Our result is best possible: we exhibit examples demonstrating that if any of our hypotheses are removed, our conclusions may no longer hold.

1 Introduction

The classical ballot theorem, proved by Bertrand (1887), states that in an election where one candidate receives p votes and the other receives $q < p$ votes, the probability that the winning candidate is in the lead throughout the counting of the ballots is precisely

$$\frac{p-q}{p+q},$$

assuming no one order for counting the ballots is more likely than another. Viewed as a statement about random walks, Bertrand's ballot theorem states that given a symmetric simple random walk S and integers n, k with $0 < k \leq n$ and with n and k of the same parity,

$$\mathbf{P}\left\{S_i > 0 \ \forall\ 0 < i < n | S_n = k\right\} = \frac{k}{n}.$$

The standard approach to extending Bertrand's ballot theorem is most easily explained by first transforming the statement, letting $S'_i = i - S_i$ for $i = 1, 2, \ldots, n$. S' is an *increasing* random walk, and the classical ballot theorem states that

$$\mathbf{P}\left\{S'_i < i \ \forall\ 0 < i < n \mid S'_n = n - k\right\} = \frac{k}{n}.$$

One may then ask: for what other increasing stochastic processes does the same result hold? This question has been well-studied; much of the seminal work on the subject was done by

Takács (1962a,b,c, 1963, 1964a,b, 1967). The most general result to date is due to Kallenberg (1999) (see also (Kallenberg, 2003, Chapter 11)).

If, rather than transforming S into an increasing stochastic process, on takes the fact that $S_n/\sqrt{n}$ converges in distribution to a normal random variable as a starting point, a different generalization of Bertrand's ballot theorem emerges. It turns out that if $\mathbf{E}X = 0$ and $0 < \mathbf{E}\{X^2\} < \infty$, then for a simple random walk S with step size X,

$$\mathbf{P}\{S_i > 0 \; \forall \; 1 \le i \le n | S_n = k\} = \Theta\left(\frac{k}{n}\right) \text{ for all } k \text{ with } 0 < k = O(\sqrt{n}). \tag{1}$$

(This is a slight misrepresentation; when we consider random variables that do not live on a lattice, the right conditioning will in fact be on an event such as $\{k \le S_n < k+1\}$ or something similar. For the moment, we ignore this technicality, presuming for the remainder of the introduction that X is integer-valued and $\mathbf{P}\{X = 1\} > 0$, say.)

Furthermore, and (to the authors) more surprising, it turns out that this result is essentially best possible. We provide examples which demonstrate that if either $\mathbf{E}\{X^2\} = \infty$ or $k \neq O(\sqrt{n})$, no equation such as (1) can be expected to hold. The philosophy behind these examples can be explained yet another perspective on ballot-style results. One may ask: what are sufficient conditions on the structure of a multiset $\mathcal{S}$ of n numbers summing to some value k to ensure that, in a uniformly random permutation of the set, all partial sums are positive with probability of order k/n? This latter perspective is philosophically closely tied to work of Andersen (1953, 1954), Spitzer (1956) and others on the amount of time spent above zero by a conditioned random walk and on related questions. Our procedure for constructing the examples showing that our result is best possible is:

1. construct a multiset $\mathcal{M}$ (whose elements sum to k, say) and for for which the bounds predicted by the ballot theorem fail;
2. find a random variable X and associated random walk S for which, given that $S_n = k$, the random multiset $\{X_1, \ldots, X_n\}$ is likely close to $\mathcal{M}$ in composition.

The details involved in carrying out this procedure end up being somewhat involved.

Outline

In Section 2 we prove our positive result, the formalization of (1). Section 3 contains the examples which show that (1) may fail to hold if either $\mathbf{E}\{X^2\} = \infty$ or if $k \neq O(\sqrt{n})$. Finally, in Section 4 we briefly discuss directions in which the current work might be extended and related potential avenues of research.

2 A general ballot theorem

Throughout this section, X is a mean zero random variable with $0 < \mathbf{E}\{X^2\} < \infty$ and S is a simple random walk with step size X. Before stating our ballot theorem, we introduce a small amount of terminology. We say X is a *lattice random variable* with period $d > 0$ if there is a constant z such that $dX - z$ is an integer random variable and d is the smallest positive real number for which this holds; in this case, we say that the set $\mathbb{L}_X = \{(n+z)/d : n \in \mathbb{Z}\}$ is *the lattice of* X. A real number A is *acceptable* for X if $A \geq 1/d$ (if X is a lattice random variable), or $A > 0$ (otherwise).

Theorem 1. *For any A which is acceptable for X, there is $C > 0$ depending only on X and A such that for all n, for all $k > 0$,*

$$\mathbf{P}\{k \leq S_n < k + A, S_i > 0 \ \forall\ 0 < i < n\} \leq \frac{C\max\{k,1\}}{n^{3/2}},$$

and for all k with with $0 < k \leq \sqrt{n}$,

$$\mathbf{P}\{k \leq S_n < k + A, S_i > 0 \ \forall\ 0 < i < n\} \geq \frac{\max\{k,1\}}{Cn^{3/2}},$$

Before proving Theorem 1, we collect a handful of results which we will use in the course of the proof. First, we will use a rather straightforward result on how "spread out" sums of independent identically distributed random variables become. The version we present is a simplification of Theorem 1 in Kesten (1972):

Theorem 2. *For any family of independent identically distributed real random variables $X_1, X_2, \ldots$ with positive, possibly infinite variance and associated partial sums $S_1, S_2, \ldots$, there is $c > 0$ depending only on the distribution of X_1 such that for all n,*

$$\sup_{x \in \mathbb{R}} \mathbf{P}\{x \leq S_n \leq x + 1\} \leq c/\sqrt{n}.$$

We will also use the following lemma, Lemma 3.3 from (Pemantle and Peres, 1995):

Lemma 3. *For $h \geq 0$, let T_h be the first time t that $S_t < -h$. Then there are constants c_1, c_2, c_3 such that for all n:*

(i) for all h with $0 \leq h \leq \sqrt{n}$, $\mathbf{P}\{T_h \geq n\} \geq c_1 \cdot \max\{h,1\}/\sqrt{n}$;

(ii) for all h with $0 \leq h \leq \sqrt{n}$, $\mathbf{E}\{S_n^2 | T_h > n\} \leq c_2 n$; and

(iii) for all $h \geq 0$, $\mathbf{P}\{T_h \geq n\} \leq c_3 \cdot \max\{h,1\}/\sqrt{n}$.

(In (Pemantle and Peres, 1995), (ii) was only proved with $h = 0$, but an essentially identical proof yields the above formulation.) We will use the following easy corollary of Lemma 3 in proving the lower bound of Theorem 1.

Corollary 4. *There exists $\epsilon > 0$ such that for all n and all h with $0 \le h \le \sqrt{n}$,*

$$\mathbf{E}\left\{S_n^2 \mid T_h > n, S_n \ge \epsilon\sqrt{n}\right\} \le 3c_2 n.$$

Proof. Choose $\epsilon > 0$ such that for all n, $\mathbf{P}\{S_n \ge \epsilon\sqrt{n}\} > 1/3$. By the FKG inequality, $\mathbf{P}\{S_n \ge \epsilon\sqrt{n} \mid T_0 > n\} \ge 1/3$. We thus have

$$\begin{aligned}\mathbf{E}\left\{S_n^2 \mid T_h > n, S_n \ge \epsilon\sqrt{n}\right\} &= \frac{\mathbf{E}\left\{S_n^2 \mathbf{1}_{[S_n \ge \epsilon\sqrt{n}]} \mid T_0 > n\right\}}{\mathbf{P}\left\{S_n \ge \epsilon\sqrt{n} \mid T_0 > n\right\}} \\ &\le 3\mathbf{E}\left\{S_n^2 \mathbf{1}_{[S_n \ge \epsilon\sqrt{n}]} \mid T_0 > n\right\} \le 3\mathbf{E}\left\{S_n^2 \mid T_0 > n\right\},\end{aligned}$$

and Lemma 3 (ii) completes the proof. □

Finally, in proving the lower bound of Theorem 1, we will use a local central limit theorem. The following is a weakening of Theorem 1 from Stone (1965).

Theorem 5 (Stone (1965))**.** *Fix any $c > 0$. If $\mathbf{E}X = 0$ and $0 < \mathbf{E}\{X^2\} < \infty$ then for any $h > 0$, if X is non-lattice then for all x with $|x| \le c\sqrt{n}$,*

$$\mathbf{P}\{x \le S_n < x + h\} = (1 + o(1))\frac{h \cdot e^{-x^2/(2n\mathbf{E}\{X^2\})}}{\sqrt{2\pi\mathbf{E}\{X^2\}n}},$$

and if X is lattice then for all $x \in \mathbb{L}_X$ with $|x| \le c\sqrt{n}$,

$$\mathbf{P}\{S_n = x\} = (1 + o(1))\frac{e^{-x^2/(2n\mathbf{E}\{X^2\})}}{\sqrt{2\pi\mathbf{E}\{X^2\}n}}.$$

In both cases, $o(1) \to 0$ as $n \to \infty$ uniformly over all x in the allowed range.

Proof of Theorem 1. We first remark that when X is a lattice random variable with period d, if $A = 1/d$ then $[k, k+A)$ contains precisely one element from the lattice of X. To shorten the formulas in the proof, we assume that X is indeed lattice, that $A = 1/d$, and that k is in the lattice of X, so that $k \le S_n < k + A$ if and only if $S_n = k$. The proof of the more general formulation requires only a line-by-line rewriting of what follows.

Throughout this proof, c_1, c_2, and c_3 are the constants from Lemma 3 and ϵ is the constant from Corollary 4. We first prove the upper bound. We assume that $n \ge 4$. Let S^r be the random walk with $S_0^r = 0$ and, for i with $0 \le i < n$, with $S_r^{i+1} = S_i^r - X_{n-i}$. Define T_0 as in Lemma 3 and let T_k^r be the minimum of n and the first time t that $S_t^r \le -k$.

In order that $S_n = k$ and $S_i > 0$ for all $0 < i < n$, it is necessary that

- $T_0 > \lfloor n/4 \rfloor$,

- $T^r_{k+A} > \lfloor n/4 \rfloor$, and
- $S_n = k$.

By the independence of disjoint sections of the random walk and by Lemma 3 (iii), we have

$$\begin{aligned} \mathbf{P}\left\{T_0 > \lfloor n/4 \rfloor, T^r_k > \lfloor n/4 \rfloor\right\} &= \mathbf{P}\left\{T_0 > \lfloor n/4 \rfloor\right\} \mathbf{P}\left\{T^r_k > \lfloor n/4 \rfloor\right\} \\ &\leq \frac{c_3 \max\{k, 1\}}{(\sqrt{\lfloor n/4 \rfloor})^2} < \frac{8c_3 \max\{k, 1\}}{n}. \end{aligned} \tag{2}$$

We next rewrite the condition $S_n = k$ as

$$S_{\lceil 3n/4 \rceil} - S_{\lfloor n/4 \rfloor} = k - S_{\lfloor n/4 \rfloor} - S^r_{\lfloor n/4 \rfloor}.$$

Since $S_{\lceil 3n/4 \rceil} - S_{\lfloor n/4 \rfloor}$ is independent of $S_{\lfloor n/4 \rfloor}$, of $S^r_{\lfloor n/4 \rfloor}$, and of the events $\{T_0 > \lfloor n/4 \rfloor\}$ and $\{T^r_{k+A} > \lfloor n/4 \rfloor\}$, we have

$$\mathbf{P}\left\{S_n = k \mid T_0 > \lfloor n/4 \rfloor, T^r_k > \lfloor n/4 \rfloor\right\} \leq \sup_{r \in \mathbb{R}} \mathbf{P}\left\{S_{\lceil 3n/4 \rceil} - S_{\lfloor n/4 \rfloor} = r\right\}.$$

By Theorem 2, we thus have

$$\mathbf{P}\left\{S_n = k \mid T_0 > \lfloor n/4 \rfloor, T^r_k > \lfloor n/4 \rfloor\right\} \leq \frac{c}{\sqrt{\lceil 3n/4 \rceil - \lfloor n/4 \rfloor}} < \frac{2c}{\sqrt{n}}, \tag{3}$$

where c is the constant from Theorem 2. Combining (2) and (3) proves the upper bound.

Fix α with $0 < \alpha < 1/2$ so that $4(1-2\alpha)\mathbf{E}\left\{X^2\right\} < \epsilon^2$, where ϵ is the constant from Corollary 4. In order that $S_n = k$ and $S_i > 0$ for all $0 < i < n$, it is sufficient that the following events occur:

E_1: $T_0 > \lfloor \alpha n \rfloor$ and $\epsilon\sqrt{n} \leq S_{\lfloor \alpha n \rfloor} \leq \sqrt{6c_2 n}$,

E_2: $T^r_k > \lfloor \alpha n \rfloor$ and $\epsilon\sqrt{n} \leq S^r_{\lfloor \alpha n \rfloor} \leq \sqrt{6c_2 n}$,

E_3: $\min_{\lfloor \alpha n \rfloor \leq i \leq \lceil (1-\alpha)n \rceil} S_i - S_{\lfloor \alpha n \rfloor} > -S_{\lfloor \alpha n \rfloor}$, and

E_4: $S_n = k$.

By the independence of disjoint sections of the random walk, we therefore have

$$\mathbf{P}\left\{S_n = k, S_i > 0 \;\forall\; 0 < i < n\right\} \geq \mathbf{P}\left\{E_1\right\} \cdot \mathbf{P}\left\{E_2\right\} \cdot \mathbf{P}\left\{E_3, E_4 \mid E_1, E_2\right\}. \tag{4}$$

By Lemma 3 (i), we have $\mathbf{P}\left\{T_0 > \lfloor \alpha n \rfloor\right\} \geq c_1/\sqrt{n}$, so by the FKG inequality we see that

$$\mathbf{P}\left\{T_0 > \lfloor \alpha n \rfloor, S_{\lfloor \alpha n \rfloor} \geq \epsilon\sqrt{n}\right\} \geq \frac{c_1}{3\sqrt{n}}. \tag{5}$$

By applying Corollary 4 and Markov's inequality, we also have

$$\mathbf{P}\left\{S_{\lfloor \alpha n\rfloor} > \sqrt{6c_2 n}\ \middle|\ T_0 > \lfloor \alpha n\rfloor, S_{\lfloor \alpha n\rfloor} \geq \epsilon\sqrt{n}\right\} \leq \frac{\mathbf{E}\left\{S_n^2 \mid T_0 > \lfloor \alpha n\rfloor, S_{\alpha n} \geq \epsilon n\right\}}{6c_2 n} \leq \frac{1}{2}, \quad (6)$$

and (5) and (6) together imply that

$$\mathbf{P}\left\{E_1\right\} \geq \frac{c_1}{6\sqrt{n}}. \quad (7)$$

An identical argument shows that $\mathbf{P}\left\{E_2\right\} \geq c_1 \cdot \max\{k, 1\}/(6\sqrt{n})$. By (4), to prove the lower bound it thus suffices to show that there is $\gamma > 0$ not depending on n or on our choice of k and such that $\mathbf{P}\left\{E_3, E_4 \mid E_1, E_2\right\} \geq \gamma/\sqrt{n}$; we now turn to establishing such a bound.

Let $m = \lceil (1-\alpha)n\rceil - \lfloor \alpha n\rfloor$. For $1 \leq i \leq m$, let

$$L_i = S_{\lfloor \alpha n\rfloor + i} - S_{\lfloor \alpha n\rfloor}, \text{ and let } R_i = S_{\lceil (1-\alpha)n\rceil - i} - S_{\lceil (1-\alpha)n\rceil},$$

so in particular $L_m = -R_m$. Next, rewrite the event E_4 as

$$L_m = k - S_{\lfloor \alpha n\rfloor} - S^r_{\lfloor \alpha n\rfloor}.$$

By the independence of disjoint sections of the random walk, we then have

$$\mathbf{P}\left\{E_3, E_4 \mid E_1, E_2\right\} \geq \inf_{p,q \in [\epsilon\sqrt{n}, \sqrt{6c_2 n}] \cap \mathbb{L}_X} \left\{\mathbf{P}\left\{L_m = -p + (k+q), \min_{1\leq i\leq m} L_i > -p\right\}\right\}. \quad (8)$$

For p and q as in (8), we define the following shorthand:

- $A_{p,q}$ is the event that $L_m = -p + (k+q)$, and
- B_p is the event that $\min_{1\leq i\leq m} L_i > -p$.

We bound $\mathbf{P}\left\{A_{p,q}, B_p\right\}$ by first writing

$$\mathbf{P}\left\{A_{p,q}, B_p\right\} \geq \mathbf{P}\left\{A_{p,q}\right\} - \mathbf{P}\left\{A_{p,q}, \overline{B_p}\right\}. \quad (9)$$

By Theorem 5, for all k with $0 \leq k \leq \sqrt{n}$ and all $p, q \in [\epsilon\sqrt{n}, \sqrt{6c_2 n}] \cap \mathbb{L}_X$,

$$\mathbf{P}\left\{A_{p,q}\right\} = (1 + o(1))\frac{e^{-(k+q-p)^2/(2\sigma^2 m)}}{\sqrt{2\pi\sigma^2 m}}, \quad (10)$$

where $o(1) \to 0$ as $n \to \infty$, uniformly over all k, p, and q as above. To bound $\mathbf{P}\left\{A_{p,q}, \overline{B_p}\right\}$ from above, we first further divide the events $A_{p,q}$, B_p. Let $m' = \lfloor n/2\rfloor - \lfloor \alpha n\rfloor$. Observe that $A_{p,q}$ occurs if and only if $R_m = -(k+q) + p$. Similarly, if $A_{p,q}$ occurs, then for $\overline{B_p}$ to occur one of the following events must occur: either

1. $\min_{1\leq i\leq m'} L_i \leq -p$ (we call this event C_p); or

2. $\min_{1\le i\le m-m'} R_i \le -(k+q)$ (we call this event D_q).

Thus,

$$\mathbf{P}\left\{A_{p,q}, \overline{B_p}\right\} \le \mathbf{P}\left\{C_p, L_m = -p+(k+q)\right\} + \mathbf{P}\left\{D_q, R_m = -(k+q)+p\right\}. \tag{11}$$

By Kolmogorov's inequality and our choice of α,

$$\mathbf{P}\left\{C_p\right\} \le \frac{\mathbf{E}\left\{L_m^2\right\}}{p^2} = \frac{\mathbf{E}\left\{X^2\right\}\cdot mm}{p^2} \le \frac{\mathbf{E}\left\{X^2\right\}((1-2\alpha)n+1)}{\epsilon^2 n} < \frac{1}{4}, \tag{12}$$

for all n sufficiently large. Furthermore, since for any i with $1 \le i \le m'$, $L_m - L_i$ is a sum of $m - i \ge m - m' \ge \lceil (1-\alpha)n \rceil - \lfloor n/2 \rfloor$ copies of X, by the independence of disjoint sections of the random walk and by Theorem 5,

$$\begin{aligned}
\mathbf{P}\left\{L_m = -p+(k+q) \mid L_i \le -p\right\} &= (1+o(1))\frac{e^{-(k+q)^2/(2\sigma^2(m-i))}}{\sqrt{2\pi\sigma^2(m-i)}} \\
&\le (1+o(1))\frac{e^{-(k+q-p)^2/(2\sigma^2 m)}}{\sqrt{2\pi\sigma^2 m}},
\end{aligned} \tag{13}$$

where $o(1) \to 0$ as $n \to \infty$, uniformly over all i, k, p, and q as above. It follows by Bayes' formula that

$$\begin{aligned}
\mathbf{P}\left\{C_p, L_m = -p+(k+q)\right\} &\le \mathbf{P}\left\{C_p\right\} \cdot \max_{1\le i\le m} \mathbf{P}\left\{L_m = -p+(k+q) \mid L_i \le -p\right\} \\
&\le (1+o(1))\frac{e^{-(k+q-p)^2/(2\sigma^2 m)}}{4\sqrt{2\pi\sigma^2 m}}.
\end{aligned} \tag{14}$$

A similar calculation shows that

$$\mathbf{P}\left\{D_q, R_m = -(k+q)+p\right\} \le (1+o(1))\frac{e^{-(k+q-p)^2/(2\sigma^2 m)}}{4\sqrt{2\pi\sigma^2 m}}, \tag{15}$$

and combining (9), (10), (14) and (15), since m and n have the same order we see that

$$\mathbf{P}\left\{A_{p,q}, B_p\right\} \ge (1+o(1))\frac{e^{-(k+q-p)^2/(2\sigma^2 m}}{2\sqrt{2\pi\sigma^2 m}} \ge \frac{\gamma}{\sqrt{n}}, \tag{16}$$

for all k, p and q in the allowed ranges, for some γ not depending on k, p, q, or n. Together (8) and (16) establish the required lower bound on $\mathbf{P}\left\{E_3, E_4 \mid E_1, E_2\right\}$ and complete the proof. □

3 Counterexamples

In Section 3.2 we exhibit a random walk S with mean zero step, finite variance step size X and show that with $k = n$, $\mathbf{P}\left\{S_n > 0 \;\forall\; 0 < i < n | S_n = k\right\}$ is not $\Theta(k/n) = \Theta(1)$ as the ballot theorem would suggest. This example may be easily modified to show that we can not in general expect a result of the form $\mathbf{P}\{S_n > 0 \ \forall\ 0 < i < n | S_n = k\} = \Theta(k/n)$ for any $k = \Omega(\sqrt{n}\log n)$, and we believe that with a little effort it should be possible to make this approach work for any $k = \omega(\sqrt{n})$.

In Section 3.3 we exhibit a random walk S with step size X for which X is an integer random variable with period 1, and for which $\mathbf{E}X = 0$, $\mathbf{E}\left\{X^{3/2-\epsilon}\right\} < \infty$ for all $\epsilon > 0$, but such that $\mathbf{P}\{S_n > 0 \ \forall\ 0 < i < n \mid S_n = \sqrt{n}\}$ is not $\Theta(n^{-1/2})$; the same idea can be easily modified to yield a random variable X with $\mathbf{E}X^\alpha < \infty$ for any fixed $\alpha < 2$ and for which the ballot theorem can be seen to fail even in the range $S_n = O(\sqrt{n})$.

The ideas behind our examples are most easily explained from the multiset mentioned in the introduction. The "underlying multiset" $\mathcal{S}$ for our example showing that the condition $k = O(n)$ is necessary consists of $(n-1)/2$ elements of value 1, the same number of elements of value -1's, and a single element of value n. The elements of $\mathcal{S}$ sum to n, and in order for all partial sums to stay positive, it is necessary and sufficient that the partial sums not containing the element n stay positive (as all partial sums containing n are certainly positive). Denoting the elements of $\mathcal{S}$ by $x_1, \ldots, x_n$ and letting σ be a uniformly random permutation of $\{1, \ldots, n\}$, the index i for which $x_{\sigma(i)} = n$ is uniform among $\{1, \ldots, n\}$. Letting $S_{\sigma(i)} = \sum_{j=1}^{i} x_{\sigma(j)}$ for $0 < i \leq n$, we may thus write

$$\begin{aligned}\mathbf{P}\left\{S_{\sigma(i)} > 0 \ \forall\ 0 < i < n\right\} &= \frac{1}{n} \cdot \sum_{j=1}^{n} \mathbf{P}\left\{S_{\sigma(i)} > 0 \ \forall\ 0 < i < n | x_{\sigma(j)} = n\right\} \\ &= \frac{1}{n} \cdot \sum_{j=1}^{n} \mathbf{P}\left\{S_{\sigma(i)} > 0 \ \forall\ 0 < i < j | x_{\sigma(j)} = n\right\}. \qquad (17)\end{aligned}$$

For a symmetric simple random walk S', it is well-known (see, e.g., Feller (1968a), Lemma III.3.1) that for integers $j > 0$, $\mathbf{P}\left\{S'_i > 0 \ \forall\ 0 < i < j\right\} = O(j^{-1/2})$. Making the hopefully plausible leap of faith that the same bound holds for $\mathbf{P}\left\{S_{\sigma(i)} > 0 \ \forall\ 0 < i < j | x_{\sigma(j)=n}\right\}$, (17) then yields that

$$\mathbf{P}\left\{S_{\sigma(i)} > 0 \ \forall\ 0 < i < n\right\} = \frac{1}{n} \cdot \sum_{j=1}^{n} O\left(\frac{1}{\sqrt{j}}\right) = O\left(\frac{1}{\sqrt{n}}\right),$$

not $\Theta(1)$ as the ballot theorem would suggest. Turning this intuition into the example of Section 3.2 is a matter of finding a random walk S with mean zero step size $X \in \mathcal{D}$ for which, given that $S_n = n$, the set $\{X_1, \ldots, X_n\}$ very likely looks like the set $\mathcal{S}$ above, i.e., there is a single index i for which $X_i = n$, and for all other j, X_j is "small".

For the example showing that the condition $\mathbf{E}\{X^2\} < \infty$, "underlying multiset" we are thinking of consists of roughly $(n - n^{1/4})/2$ elements of value 1, the same number of elements of value $+1$, $(n^{1/4}+1)/2$ elements of value $\sqrt{n}$, and $(n^{1/4}-1)/2$ elements of value $-\sqrt{n}$. These elements sum to $\sqrt{n}$.

For all partial sums in a uniformly random permutation of this multiset to stay positive, it is necessary that the partial sums stay positive until an element of value $\sqrt{n}$ is sampled – this should occur after about $n^{3/4}$ elements have been sampled, so the intuition given by a symmetric simple random walk suggests that the partial sums stay positive until this time with probability of order $O(n^{-(3/4)\cdot(1/2)}) = O(n^{-3/8})$.

In order that the partial sums stay positive, it is also essentially necessary that the "sub-random walk" consisting of the partial sums *of only elements of absolute value* $\sqrt{n}$ stays positive – for if this "sub-random walk" becomes extremely negative then it is very unlikely that the original partial sums stay positive. Dividing through by $n^{1/2}$ we can view this "sub-random walk" as a symmetric simple random walk S' of length $n^{1/4}$, conditioned on having $S'_{n^{1/4}} = 1$. By the ballot theorem, the probability such a random walk stays positive is $O(1/n^{1/4})$. Combining the bounds of the this paragraph and the previous paragraph as though the two events were independent (which, though clearly false, gives the correct intuition) suggests that the original partial sums should stay positive with probability $O(1/n^{3/8+1/4}) = O(1/n^{5/8})$, not $\Theta(n^{-1/2})$ as the ballot theorem would suggest.

Before we turn to the details of these examples, we first spend a moment gathering two easy lemmas that we will use in the course of their explanation.

3.1 Two Easy Lemmas

The first lemma bounds the probability that a random walk stays above zero until some time m, and is a simplification of Feller (1968b), Theorem XII.7.12a.

Lemma 6. *Given a random walk S with step size X, if X is symmetric then for integers $m > 0$,*

$$\mathbf{P}\{S_i > 0 \ \forall\ 0 < i \leq m\} = \Theta\left(\frac{1}{\sqrt{m}}\right)$$

The second lemma is an easy extension of classical Chernoff (1952) bounds to a setting in which the number of terms in the binomial is random. The classical Chernoff bounds (see, e.g., (2.5) and (2.6) in Janson et al. (2000) for a modern reference), state: given a binomial random variable $\mathrm{Bin}(n,p)$ with mean $\mu = np$, for all $c > 0$,

$$\mathbf{P}\{\mathrm{Bin}(n,p) > (1+c)\mu\} \leq e^{-c^2\mu/(2(1+c/3))} \tag{18}$$

$$\mathbf{P}\{\mathrm{Bin}(n,p) < (1-c)\mu\} \leq e^{-c^2\mu/2} \tag{19}$$

The following lemma follows from the Chernoff bounds by straightforward applications of Bayes' formula:

Lemma 7. *Let m be a positive integer, let $0 < q < 1$, and let U be distributed as* $\mathrm{Bin}(m,q)$. *Let v be a positive real number and let $V_1, V_2, \ldots$ be i.i.d. random variables taking values $\pm v$,* *each with probability* $1/2$*. Finally, let* $Y = V_1 + \ldots + V_U$*. The for all* $t > 0$*,*

$$\begin{aligned} \mathbf{P}\{Y > t\} &\leq \exp\left\{-\frac{t^2}{8mq + \frac{4tv}{3}}\right\} + \exp\left\{-\frac{mq}{3}\right\}, \quad \textit{and} \\ \mathbf{P}\{Y < -t\} &\leq \exp\left\{-\frac{t^2}{8mq}\right\} + \exp\left\{-\frac{mq}{3}\right\}. \end{aligned}$$

Proof. As $\mathbf{E}U = mq$, by (18) we have

$$\begin{aligned} \mathbf{P}\{Y > t\} &\leq \sum_{u=1}^{\lfloor 2mq \rfloor} \mathbf{P}\{Y > t | U = u\} \mathbf{P}\{U = u\} + \mathbf{P}\{U > 2mq\} \\ &\leq \sup_{u \leq 2mq} \mathbf{P}\{Y > t | U = u\} + \mathbf{P}\{U > 2mq\} \\ &\leq \sup_{u \leq 2mq} \mathbf{P}\{Y > t | U = u\} + e^{-\frac{mq}{3}} \end{aligned} \tag{20}$$

Given that $U = u$, $(Y + uv)/2v$, which we denote Y', is distributed as $\mathrm{Bin}(u, 1/2)$. Furthermore, in this case $Y > t$ if and only if

$$Y' > \frac{u}{2} + \frac{t}{2v} = \frac{u}{2}\left(1 + \frac{t}{uv}\right).$$

It follows by (18) that

$$\begin{aligned} \mathbf{P}\{Y > t | U = u\} &\leq \exp\left\{-\frac{u}{2}\left(\frac{t}{uv}\right)^2\left(2 + \frac{2t}{3uv}\right)^{-1}\right\} \\ &= \exp\left\{-\frac{t^2}{4uv^2 + \frac{4tv}{3}}\right\} \\ &< \exp\left\{-\frac{t^2}{8mq + \frac{4tv}{3}}\right\}, \end{aligned}$$

where in the last inequality we use that $u \leq 2mq$ and $v \geq 1$. As u was arbitrary, combining this inequality with (20) yields the desired bound on $\mathbf{P}\{Y > t\}$. The bound on $\mathbf{P}\{Y < -t\}$ is proved identically. □

3.2 Optimality for normal random walks

Let f be the tower function: $f(0) = 1$ and $f(k+1) = 2^{f(k)}$ for integers $k \geq 0$. We define a random variable X as follows:

$$X = \begin{cases} \pm 1, & \text{each with probability } \frac{1}{2} \cdot \left(1 - \sum_{i=0}^{\infty} \frac{1}{f(k)^4}\right) \\ \pm f(k), & \text{each with probability } \frac{1}{2f(k)^4}, \text{ for } k = 1, 2, \ldots \end{cases}$$

and let S be a random walk with steps distributed as X. Clearly $\mathbf{E}X = 0$, and it is easily checked that $\mathbf{E}\{X^2\} < 2$. We will show that when $n = f(k)$ for positive integers k, $\mathbf{P}\{S_i > 0 \; \forall 0 < i < n | S_n = n\}$ is $O(1/\sqrt{n})$, so in particular, for such values of n this probability is not $\Theta(1)$ as the ballot theorem would suggest.

For $i \geq 0$ we let N_i be the number of times t with $1 \leq t \leq n$ that $|X_t| = f(i)$; clearly $\sum_{i=0}^{\infty} N_i = n$. For $S_n = n$ to occur, it suffices that for some t with $1 \leq t \leq n$, $X_t = n$ and $S_n - X_t = 0$; therefore

$$\mathbf{P}\{S_n = n\} \geq \sum_{t=1}^{n} \mathbf{P}\{X_t = n, S_n - X_t = 0\} \geq \sum_{t=1}^{n} \frac{\mathbf{P}\{S_n - X_t = 0\}}{2n^4}.$$

As $\mathbf{E}X^2 < \infty$ and, for all $1 \leq t \leq n$, $S_n - X_t$ is simply a sum of $n - t$ independent copies of X, by Theorem 5 we know that $\mathbf{P}\{S_n - X_t = 0\} = \Theta(n^{-1/2})$ uniformly over $1 \leq i \leq n$. It follows that

$$\mathbf{P}\{S_n = n\} = \Omega\left(\frac{1}{n^{7/2}}\right). \tag{21}$$

We next bound the probability that $S_n = n$ and $S_t > 0$ for all $0 < t < n$; we denote this conjunction of events E. Our aim is to show that $\mathbf{P}\{E\} = O(n^{-4})$, which together with (21) will establish our claim that $\mathbf{P}\{S_i > 0 \; \forall \; 0 < i < n | S_n = n\}$ is $O(1/\sqrt{n})$. Recalling that $f(k) = n$, we write

$$\mathbf{P}\{E\} \;=\; \mathbf{P}\{E, N_k = 0\} + \mathbf{P}\{E, N_k \geq 1\}. \tag{22}$$

It is easy to show using Chernoff bounds that $\mathbf{P}\{E, N_k = 0\} = O(n^{-6})$ (we postpone this step for the moment). From this fact and from (22), we therefore have

$$\mathbf{P}\{E\} = \mathbf{P}\{E, N_k \geq 1\} + O\left(n^{-6}\right),$$

from which it will follow that $\mathbf{P}\{E\} = O(n^{-4})$ if we can show that $\mathbf{P}\{E, N_k \geq 1\} = O(n^{-4})$. We first do so, then justify our assertion that $\mathbf{P}\{E, N_k = 0\} = O(n^{-6})$.

For E and $\{N_k \geq 1\}$ to occur, one of the following events must occur.

- For some t with $1 \leq t \leq \lfloor n/2 \rfloor$, $X_t = \pm n$, $S_i > 0$ for all $0 < i < t$, and $S_n = n$. We denote these events B_t, for $1 \leq t \leq \lfloor n/2 \rfloor$, and remark that they are *not* disjoint.
- $S_i > 0$ for all $0 < i < \lfloor n/2 \rfloor$, and for some t with $\lfloor n/2 \rfloor < t \leq n$, $X_t = \pm n$ and $S_n = n$. We denote these events D_t for $\lfloor n/2 \rfloor \leq t \leq n$; again, they are not disjoint.

We first bound the probabilities of the events B_t, $1 \leq t \leq \lfloor n/2 \rfloor$. Fix some t in this range – by Lemma 6, the probability that $S_i > 0$ for all $0 < i < t$ is $O(t^{-1/2})$ uniformly in t. By definition, $\mathbf{P}\{X_t = \pm n\} = n^{-4}$, so by the strong Markov property,

$$\mathbf{P}\{S_i > 0 \; \forall \; 0 < i < t, X_t = \pm n\} = O\left(\frac{1}{n^4\sqrt{t}}\right),$$

still uniformly in t. Furthermore, by Theorem 2 we have that

$$\sup_r \mathbf{P}\left\{S_n - S_t = r\right\} = O\left(\frac{1}{\sqrt{n-t}}\right) = O\left(\frac{1}{\sqrt{n}}\right). \tag{23}$$

As $S_n - S_t$ and $n - S_t$ are independent, it follows from (23) that $\mathbf{P}\left\{S_n - S_t = n - S_t\right\} = O(n^{-1/2})$, and by another application of the strong Markov property we have that

$$\mathbf{P}\left\{B_t\right\} = \mathbf{P}\left\{S_i > 0\ \forall 0 < i < t, X_t = \pm n, S_n - S_t = n - S_t\right\} = O\left(\frac{1}{n^{9/2}\sqrt{t}}\right).$$

Thus,

$$\mathbf{P}\left\{\bigcup_{t=1}^{\lfloor n/2\rfloor} B_t\right\} \leq \sum_{t=1}^{\lfloor n/2\rfloor} \mathbf{P}\left\{B_t\right\} \leq \sum_{t=1}^{\lfloor n/2\rfloor} O\left(\frac{1}{n^{9/2}\sqrt{t}}\right) = O\left(\frac{1}{n^4}\right). \tag{24}$$

We next bound the probabilities of the events D_t, $\lfloor n/2\rfloor < t \leq n$. By an argument just as above, we have that

$$\mathbf{P}\left\{S_i > 0\ \forall 0 < i < \lfloor n/2\rfloor, X_t = \pm n\right\} = O\left(\frac{1}{n^{9/2}}\right).$$

Also just as above, since $S_n - S_{\lfloor n2\rfloor} - X_t$ and $n - S_{\lfloor n2\rfloor} - X_t$ are independent,

$$\mathbf{P}\left\{S_n - S_{\lfloor n2\rfloor} - X_t = n - S_{\lfloor n2\rfloor} - X_t\right\} = O\left(\frac{1}{\sqrt{n}}\right).$$

By the independence of disjoint sections of the random walk we therefore have that $\mathbf{P}\left\{D_t\right\} = O(n^{-5})$, and so

$$\mathbf{P}\left\{\bigcup_{t=\lfloor n/2\rfloor+1}^{n} D_t\right\} = O\left(\frac{1}{n^4}\right). \tag{25}$$

As $E \cap \{N_k \geq 1\}$ is contained in $\bigcup_{t=1}^{\lfloor n/2\rfloor} B_t \cup \bigcup_{t=\lfloor n/2\rfloor+1}^{n} D_t$, it follows from (24) and (25) that $\mathbf{P}\left\{E, N_k \geq 1\right\} = O(n^{-4})$ as claimed.

We now turn our attention to proving that $\mathbf{P}\left\{E, N_k = 0\right\} = O(n^{-6})$. We will in fact show that

$$\mathbf{P}\left\{N_k = 0, |S_n| \geq 8k\sqrt{n\log n}\right\} = O\left(\frac{1}{n^6}\right), \tag{26}$$

which implies the desired bound. We recall that N_i is the number of times t with $1 \leq t \leq n$ that $X_t = f(i)$. For $0 \leq i < k$, let $S_{n,i} = \sum_{\{1\leq j\leq n: |X_j| = f(i)\}} X_j$. If $N_k = 0$ then either $S_n = \sum_{i=1}^{k-1} S_{n,i}$ or $N_j > 0$ for some $j > k$; thus,

$$\{N_k = 0, |S_n| > 8k\sqrt{n\log n}\} \subseteq \left\{\sum_{j=k+1}^{\infty} N_j > 0\right\} \cup \bigcup_{i=1}^{k-1}\{|S_{n,i}| > 8\sqrt{n\log n}\}. \tag{27}$$

For any $1 \le i < k$, $S_{n,i}$ is the sum of N_i i.i.d. random variables taking values $\pm g(i)$, each with probability $1/2$, and N_i is distributed as $\mathrm{Bin}(n, g(i)^{-4})$. By Lemma 7, therefore,

$$\mathbf{P}\left\{|S_{n,i}| > 8\sqrt{n\log n}\right\} \le 2\exp\left\{\frac{-64n\log n}{\frac{8n}{g(i)^2} + \frac{32\sqrt{n\log n}g(i)}{3}}\right\} + 2\exp\left\{\frac{-n}{3g(i)^4}\right\}.$$

Since $g(i) \le \log n$, presuming n is large enough that $32\sqrt{n\log n}g(i) < n$ we thus have

$$\mathbf{P}\left\{|S_{n,i}| > 8\sqrt{n\log n}\right\} \le 2\exp\left\{\frac{-64\log n}{9}\right\} + 2\exp\left\{\frac{-n}{3\log^4 n}\right\} = O\left(\frac{1}{n^7}\right). \quad (28)$$

Furthermore, it follows directly from the definition of X that $\mathbf{P}\left\{\sum_{j=k+1}^{\infty} N_j > 0\right\} = o(2^{-n})$. Applying this fact, (28), and (27), it follows immediately that

$$\mathbf{P}\left\{N_k = 0, |S_n| > 8k\sqrt{n\log n}\right\} = o(2^{-n}) + O\left(\frac{k}{n^7}\right) = O\left(\frac{1}{n^6}\right)$$

as claimed.

3.3 Optimality for other random walks

Let $g : \mathbb{N} \to \mathbb{N}$ be an rapidly increasing integer-valued function with $g(0) = 1$; in particular, we choose g such that $g \ge f$ where f is the tower function seen in the previous section.

$$X = \begin{cases} \pm 1, & \text{each with probability } \frac{1}{2}\left(1 - \sum_{i=0}^{\infty} \frac{1}{g(k)^{3/2}}\right) \\ \pm g(k), & \text{each with probability } \frac{1}{2g(k)^{3/2}}, \text{ for } k = 1, 2, \ldots \end{cases}$$

and let S be a random walk with steps distributed as X. Clearly $\mathbf{E}X = 0$. For integers $i > 0$, let Pos_i be the event that $S_j > 0$ for all $0 < j \le i$; we also let Pos_0 be some event of probability 1 as it will simplify later equations. We will show that when $n = g(k)^2$ for positive integers k,

$$\mathbf{P}\left\{S_n = \sqrt{n}\right\} = \Omega\left(\frac{1}{\sqrt{\log n}n^{5/8}}\right), \quad (29)$$

and that

$$\mathbf{P}\left\{S_n = \sqrt{n}, Pos_n\right\} = O\left(\frac{\log^{13/2} n}{n^{5/4}}\right), \quad (30)$$

from which it follows by Bayes' formula that for such values of n $\mathbf{P}\left\{Pos_n | S_n = \sqrt{n}\right\}$ is $O(\log^7 n/n^{5/8})$, not $\Theta(1/\sqrt{n})$ as the ballot theorem would suggest. We now prove (29) and (30). In what follows we presume, to avoid cumbersome floors and ceilings, that $g(k) = \sqrt{n}$ has been chosen so that $\sqrt{\log n}$ and $n^{1/8}$ are both integers.

For $j = 1, 2, \ldots$ and $i = 0, 1, \ldots$, we let $\mathcal{X}_{j,i}$ be the random set $\{X_m : 1 \le m \le j, |X_m| = g(i)\}$, let $N_{j,i} = |\mathcal{X}_{j,i}|$, and let $S_{j,i} = \sum_{X_m \in \mathcal{X}_{j,i}} X_m$. For all $j = 1, 2, \ldots$, the sets $\mathcal{X}_{j,0}, \mathcal{X}_{j,1}, \ldots$ partition $\{X_1, \ldots, X_j\}$. For an integer $k \ge 0$, the *k-truncated* random walk $S^{(k)}$ is given by

$$S_j^{(k)} = \sum_{i=1}^{k} S_{j,i} = \sum_{i=1}^{j} X_i \mathbf{1}_{[|X_i| \le g(k)]},$$

for $j = 1, 2, \ldots$. We remark that for any n and any set $\mathcal{X} \subseteq \{X_1, \ldots, X_n\}$, conditional upon the event that $\bigcup_{i=1}^{k} \mathcal{X}_{n,i} = \mathcal{X}$, $S_n^{(k)}$ is simply a sum of $|\mathcal{X}|$ i.i.d. *bounded* random variables with variance at most $g(k)^2$. In particular, this implies that *after* such conditioning, $S_n^{(k)}$ obeys a local central limit theorem around 0. The key consequence of this fact (for our purposes), is that we may choose g to grow fast enough that there exists a small constant $\epsilon > 0$ such that:

$$\begin{aligned} &\forall\, k \ge 0,\ \forall\, n' \ge n \ge g(k+1),\ \forall\, \mathcal{X} \subset \{X_1, \ldots, X_{n'}\} \text{ s.t. } |\mathcal{X}| = n, \\ &\mathbf{P}\left\{ S_{n'}^{(k)} = 0 \,\middle|\, \{X_i : 1 \le i \le n', |X_i| \le g(k)\} = \mathcal{X} \right\} \ge \frac{\epsilon}{g(k)\sqrt{n}}; \end{aligned} \tag{31}$$

such a constant is guaranteed to exist by Theorem 5 and our above observation about the conditional distribution of $S_n^{(k)}$.

Fix some integer $k \ge 1$ and let $n = g(k)^2$. We remark that $\mathbf{E} N_{n,k} = n\mathbf{P}\{X = g(k)\} = n^{1/4}/2$. For $S_n = \sqrt{n}$ to occur, it suffices that $S_{n,k} = g(k) = \sqrt{n}$ and that $S_n - S_{n,k} = 0$. For any subset $\mathcal{K}$ of $\mathcal{S} = \{X_1, \ldots, X_n\}$, $S_{n,k}$ and $S_n - S_{n,k}$ are conditionally independent given that $\mathcal{X}_k = \mathcal{K}$. Letting $\mathcal{Z}$ be the set of subsets of $\mathcal{S}$ of odd parity and of size at most $2n^{1/4}$, we then have

$$\begin{aligned} \mathbf{P}\left\{S_n = \sqrt{n}\right\} &\ge \mathbf{P}\{S_{n,k} = g(k), S_n - S_{n,k} = 0\} \\ &\ge \mathbf{P}\left\{S_{n,k} = g(k), S_n - S_{n,k} = 0, N_{n,i} \le 2n^{1/4}, N_{n,i} \text{ odd}\right\} \\ &\ge \left(\inf_{\mathcal{K}\in\mathcal{Z}} \mathbf{P}\{S_{n,k} = g(k), S_n - S_{n,k} = 0 | \mathcal{X}_k = \mathcal{K}\}\right) \cdot \mathbf{P}\left\{N_{n,i} \le 2n^{1/4}, N_{n,i} \text{ odd}\right\} \\ &= \left(\inf_{\mathcal{K}\in\mathcal{Z}} \mathbf{P}\{S_{n,k} = g(k) | \mathcal{X}_k = \mathcal{K}\} \mathbf{P}\{S_n - S_{n,k} = 0 | \mathcal{X}_k = \mathcal{K}\}\right) \cdot \left(\frac{1}{2} - o(1)\right), \end{aligned} \tag{32}$$

by the aforementioned independence and a Chernoff bound.

To bound this last formula from below, fix an arbitrary element $\mathcal{K}$ of $\mathcal{Z}$. Note that $S_n - S_{n,k} = S_n^{(k-1)}$ unless there is $i > k$ such that $\mathcal{X}_i \ne \emptyset$. Since $\mathbf{P}\{|X| > g(k)\} = O(g(k+1)^{3/2}) = O(2^{-3n/4})$, it is easily seen that $\mathbf{P}\{\bigcup_{i=k}^{\infty} \mathcal{X}_i = \mathcal{K} | \mathcal{X}_k = \mathcal{K}\} = 1 - o(2^{-n/2})$. Thus, by Bayes' formula,

$$\begin{aligned} \mathbf{P}\{S_n - S_{n,k} = 0 | \mathcal{X}_k = \mathcal{K}\} &= \mathbf{P}\left\{ S_n - S_{n,k} = 0 \,\middle|\, \bigcup_{i=k}^{\infty} \mathcal{X}_k = \mathcal{X}_k = \mathcal{K} \right\} (1 - o(2^{-n/2})) \\ &= \mathbf{P}\left\{ S_n^{(k-1)} = 0 \,\middle|\, \bigcup_{i=k}^{\infty} \mathcal{X}_k = \mathcal{X}_k = \mathcal{K} \right\} (1 - o(2^{-n/2})). \end{aligned} \tag{33}$$

Letting $\mathcal{X} = \{X_1, \ldots, X_n\} - \mathcal{K}$, the previous equation implies that

$$\mathbf{P}\left\{S_n - S_{n,k} = 0 | \mathcal{X}_k = \mathcal{K}\right\} = \Omega\left(\mathbf{P}\left\{S_n^{(k-1)} = 0 \;|\; \{X_i : 1 \leq i \leq n, |X_i| \leq g(k-1)\} = \mathcal{X}\right\}\right). \tag{34}$$

As $|\mathcal{K}| \leq 2n^{1/4}$, $|\mathcal{X}| = n - |\mathcal{K}| \geq n - 2n^{1/4} \geq n^{1/2} = g(k)$, so applying (31) in (34) yields that

$$\mathbf{P}\left\{S_n - S_{n,k} = 0 | \mathcal{X}_k = \mathcal{K}\right\} = \Omega\left(\frac{1}{g(k-1)\sqrt{n - |\mathcal{K}|}}\right) = \Omega\left(\frac{1}{g(k-1)\sqrt{n}}\right). \tag{35}$$

Next, for any set $\mathcal{K} \in \mathcal{Z}$, given that $\mathcal{X}_k = \mathcal{K}$, it follows directly from a binomial approximation that $\mathbf{P}\left\{S_{n,k} = g(k) || \mathcal{X}_k| = \mathcal{K}\right\} = \Omega(|\mathcal{K}|^{-1/2}) = \Omega(n^{-1/8})$. Plugging this bound and (35) into (32) yields that

$$\mathbf{P}\left\{S_n = \sqrt{n}\right\} = \Omega\left(\frac{1}{g(k-1)n^{5/8}}\right) = \Omega\left(\frac{1}{\sqrt{\log n}n^{5/8}}\right), \tag{36}$$

establishing (29).

We next turn to our upper bound on $\mathbf{P}\left\{S_n = \sqrt{n}, Pos_n\right\}$. We shall define several ways in which the walk S can "behave unexpectedly". We first show that the walk is unlikely to behave unexpectedly; it will be fairly easy to show that given that none of the unexpected events occur, the probability that $\{S_n = \sqrt{n}\}$ and Pos_n both occur is $O(\log^6 n / n^{5/4})$. Combining this bound with our bounds on the probability of unexpected events will yield (30).

We first describe and bound the probabilities of the so-called "unexpected events". Let B be the event that there is i with $1 \leq i \leq n$ for which $|X_i| > g(k)$. By a union bound,

$$\begin{aligned} \mathbf{P}\left\{B\right\} &\leq n\mathbf{P}\left\{|X_1| > g(k)\right\} = n\left(\sum_{i=k+1}^{\infty} \mathbf{P}\left\{X_1 = g(i)\right\}\right) = n\left(\sum_{i=k+1}^{\infty} \frac{1}{g(i)^{3/2}}\right) \\ &= nO\left(\frac{1}{2^{3n/4}}\right) = o\left(\frac{1}{2^{n/2}}\right) \end{aligned} \tag{37}$$

Next, let T be the first time for which $X_T = g(k) = \sqrt{n}$. Letting $t^* = 5n^{3/4}\log n$, we have

$$\mathbf{P}\left\{T > t^*\right\} \leq \mathbf{P}\left\{\bigcap_{t=1}^{t^*} \{|X_t| \neq \sqrt{n}\}\right\} = \left(1 - \frac{1}{n^{3/4}}\right)^{5n^{3/4}\log n} = O\left(\frac{1}{n^5}\right) \tag{38}$$

Now, by another union bound,

$$\mathbf{P}\left\{S_{T-1} > 8k\sqrt{t^* \log n}, T \leq t^*, B\right\} \leq t^* \sup_{1 \leq t \leq t^*} \mathbf{P}\left\{|S_{t-1}| > 8k\sqrt{t^* \log n}, T = t, B\right\} \tag{39}$$

If $\{T = t\}$ and B occur, then $S_{t-1} = \sum_{i=1}^{k-1} S_{t-1,i}$. Thus, by an argument just as we used to prove (26), we can see that

$$\mathbf{P}\left\{|S_{t-1}| > 8k\sqrt{t^* \log n}, T = t, B\right\} = O\left(\frac{1}{n^6}\right), \tag{40}$$

which, combined with (39), yields that

$$\mathbf{P}\left\{|S_{T-1}| > 8k\sqrt{t^*\log n}, T \le t^*, B\right\} = O\left(\frac{t^*}{n^6}\right) = O\left(\frac{1}{n^5}\right). \tag{41}$$

Finally, for $0 \le t < n$, let $\mathcal{Z}_t$ be the set of subsets of $\{t+1,\ldots,n\}$ of size between $n^{1/4}/3$ and $3n^{1/4}/2$. Let $\mathcal{Z} = \mathcal{Z}_0$, and let $\{T_1,\ldots,T_R\}$, which we denote $\mathcal{T}$, be the set of indices i with $T < i \le n$ for which $|X_i| = \sqrt{n}$, ordered so that $T < T_1 < \ldots < T_R \le n$. R is distributed as $\mathrm{Bin}(n-T, n^{-3/4})$, so by Bayes' formula, (18), and (19),

$$\begin{aligned}
\mathbf{P}\{\mathcal{T} \notin \mathcal{Z}, T \le t^*\} &= \sum_{t=1}^{t^*} \mathbf{P}\{\mathcal{T} \notin \mathcal{Z} | T = t\}\,\mathbf{P}\{T = t\} \\
&\le \sup_{t \le t^*} \mathbf{P}\{\mathcal{T} \notin \mathcal{Z} | T = t\} = \sup_{t \le t^*} \mathbf{P}\{\mathcal{T} \notin \mathcal{Z}_t | T = t\} \\
&= \sup_{t \le t^*} \mathbf{P}\left\{\mathrm{Bin}\left(n-t, \frac{1}{n^{3/4}}\right) < \frac{n^{1/4}}{3} \text{ or } \mathrm{Bin}\left(n-t, \frac{1}{n^{3/4}}\right) > \frac{3n^{1/4}}{2}\right\} \\
&= O\left(\frac{1}{n^6}\right).
\end{aligned} \tag{42}$$

This completes our bounds on the "unexpected events". We next use these inequalities to bound $\mathbf{P}\{S_n = \sqrt{n}, Pos_n\}$. Roughly speaking, in order that $\{S_n = \sqrt{n}\}$ and Pos_n occur, it is necessary that

(a) S stays positive until time T,

(b) The random walk S' given by $S'_i = \sum_{j=1}^{i} X_{T_j}/\sqrt{n}$ does not go "too negative" and additionally $|S'_R|$ is not "too large", and

(c) $S_n - S_T - \sqrt{n}S'_R = \sqrt{n} - S_t - \sqrt{n}S'_R$.

Though the event in (c) is precisely the event that $S_n = \sqrt{n}$, we write it in this form in order to point out that once we have conditioned on *fixed values* for T and $\mathcal{T}$, $S_n - S_T - \sqrt{n}S'_R$ is independent of $\sqrt{n} - S_T - \sqrt{n}S'_R$. We now turn to the details of defining and bounding the events in (a)-(c).

First, recall that $t^* = 5n^{3/4}\log n$. For any $t \le t^*$, by Lemma 6,

$$\begin{aligned}
\mathbf{P}\{Pos_T, T = t\} &\le \mathbf{P}\{Pos_{t-1}, T = t\} \\
&\le \mathbf{P}\left\{Pos_{t-1}, |X_t| = \sqrt{n}\right\} \\
&= \mathbf{P}\{Pos_{t-1}\}\,\mathbf{P}\left\{|X_t| = \sqrt{n}\right\} \\
&= O\left(\frac{1}{\sqrt{t}}\right) \cdot O\left(\frac{1}{n^{3/4}}\right) = O\left(\frac{1}{\sqrt{t}n^{3/4}}\right)
\end{aligned} \tag{43}$$

For *any* events E, F, and G, $\mathbf{P}\{E\} \le \mathbf{P}\{E, F\} + \mathbf{P}\left\{\bar{F}\right\}$, and $\mathbf{P}\{E, F\} \le \mathbf{P}\{E, F, G\} + \mathbf{P}\left\{F, \bar{G}\right\}$. We now apply these bounds together with the bounds (37), (38), and (41), to see that for any event E,

$$\begin{aligned}
\mathbf{P}\{E, Pos_T\} &\le \mathbf{P}\{E, Pos_T, B\} + o(2^{-n/2}) \\
&\le \mathbf{P}\{E, Pos_T, B, T \le t^*\} + O(n^{-5}) \\
&\le \mathbf{P}\left\{E, Pos_T, B, T \le t^*, |S_{T-1}| \le 8k\sqrt{t^* \log n}\right\} \\
&\quad + \mathbf{P}\left\{B, T \le t^*, |S_{T-1}| > 8k\sqrt{t^* \log n}\right\} + O(n^{-5}) \\
&= \mathbf{P}\left\{E, Pos_T, B, T \le t^*, |S_{T-1}| \le 8k\sqrt{t^* \log n}\right\} + O(n^{-5}) \qquad (44)
\end{aligned}$$

Continuing in this fashion using (42), (44), and the fact that $8k\sqrt{t^* \log n} \le 20n^{3/8}\log^{3/2} n$, and letting $j^* = 20n^{3/8}\log^{3/2} n$, we have

$$\begin{aligned}
\mathbf{P}\{E, Pos_T\} &\le \mathbf{P}\{E, Pos_T, B, T \le t^*, |S_{T-1}| \le j^*, \mathcal{T} \in \mathcal{Z}\} \\
&\quad + \mathbf{P}\{\mathcal{T} \notin \mathcal{Z}, T \le t^*\} + O(n^{-5}) \\
&= \mathbf{P}\{E, Pos_T, B, T \le t^*, |S_{T-1}| \le j^*, \mathcal{T} \in \mathcal{Z}\} + O(n^{-5}). \\
&= \sum_{t=1}^{t^*} \mathbf{P}\{E, Pos_T, B, T = t^*, |S_{T-1}| \le j^*, \mathcal{T} \in \mathcal{Z}\} + O(n^{-5}). \qquad (45)
\end{aligned}$$

By applying Bayes' formula and (43), this yields

$$\begin{aligned}
\mathbf{P}\{E, Pos_T\} &\le \sum_{t=1}^{t^*} \mathbf{P}\{Pos_t, T = t\}\, \mathbf{P}\{E | Pos_t, T = t, B, |S_{t-1}| \le j^*, \mathcal{T} \in \mathcal{Z}\} + O(n^{-5}) \\
&= \sum_{t=1}^{t^*} O\left(\frac{1}{\sqrt{t}n^{3/4}}\right) \mathbf{P}\{E | Pos_t, T = t, B, |S_{t-1}| \le j^*, \mathcal{T} \in \mathcal{Z}_t\} + O(n^{-5}) (46)
\end{aligned}$$

Next, for any $1 \le t \le t^*$ we have

$$\mathbf{P}\{E | Pos_t, T = t, B, |S_{t-1}| \le j^*, \mathcal{T} \in \mathcal{Z}_t\} \le \sup_{|s| \le j^*, \mathcal{I} \in \mathcal{Z}_t} \mathbf{P}\{E | Pos_t, T = t, B, S_{t-1} = s, \mathcal{T} = \mathcal{I}\},$$

which together with (46) gives

$$\begin{aligned}
\mathbf{P}\{E, Pos_T\} &\le \sum_{t=1}^{t^*} O\left(\frac{1}{\sqrt{t}n^{3/4}}\right) \sup_{|s| \le j^*, \mathcal{I} \in \mathcal{Z}_t} \mathbf{P}\{E | Pos_t, T = t, B, S_{t-1} = s, \mathcal{T} = \mathcal{I}\} + O(n^{-5}) \\
&= O\left(\frac{\sqrt{t^*}}{n^{3/4}}\right) \sup_{1 \le t \le t^*} \sup_{|s| \le j^*, \mathcal{I} \in \mathcal{Z}_t} \mathbf{P}\{E | Pos_t, T = t, B, S_{t-1} = s, \mathcal{T} = \mathcal{I}\} \\
&\quad + O(n^{-5}) \\
&= O\left(\frac{\sqrt{\log n}}{n^{3/8}}\right) \sup_{1 \le t \le t^*} \sup_{|s| \le j^*, \mathcal{I} \in \mathcal{Z}_t} \mathbf{P}\{E | Pos_t, T = t, B, S_{t-1} = s, \mathcal{T} = \mathcal{I}\} + O(n^{-5}) \\
&\qquad (47)
\end{aligned}$$

We will apply equation (47) with E the event $\{S_n = n\} \cap Pos_n$. We first note that for a given t with $1 \le t \le t^*$, if $|X_t| = \sqrt{n}$ and $|S_{t-1}| = s \le j^* < \sqrt{n}$, then for Pos_t to occur necessarily $X_t = \sqrt{n}$, so $S_t = \sqrt{n} + s$.

Fix any integer t with $1 \le t \le t^*$, any integer s for which $|s| \le j^*$, and any $\mathcal{I} \in \mathcal{Z}_t$. We hereafter denote by *Good* the intersection of events

$$Pos_t \cap \{T = t\} \cap B \cap \{S_{t-1} = s\} \cap \{\mathcal{T} = \mathcal{I}\},$$

and by $\mathbf{P}^c\{\cdot\}$ the conditional probability measure

$$\mathbf{P}\{\cdot|Good\}.$$

Given that $\{T_1, \ldots, T_R\} = \mathcal{I}$, R is deterministic – say $R = r$ – and $n^{1/4}/3 \le r \le 3n^{1/4}/2$. We recall that S' was the random walk with $S'_i = \sum_{j=1}^{i} X_{T_j}/\sqrt{n}$. As previously discussed, given that *Good* occurs, $S_t = \sqrt{n} + s$. In order that $\{S_n = \sqrt{n}\}$ and Pos_n occur, then, it is necessary that either

1. for some integer m with $|m| \le 10\log^2 n$, $S'_r = m$, $S'_j \ge -10\log^2 n$ for all $1 \le j \le r$, and $S_n - S_t - \sqrt{n}S'_r = -s - m\sqrt{n}$ (we denote these events B_m for $|m| \le 10\log^2 n$), or
2. there is j with $t < j \le n$ for which $|\sum_{i=1}^{k-1} S_{j,i}| \ge 10\sqrt{n}\log^2 n$ (we denote this event C).

To see this, observe that if none of the events B_m occurs and C does not occur, then either:

- $|S'_r| > 10\log^2 n$, in which case

  $$S_n \ge S'_r\sqrt{n} - |\sum_{i=1}^{k-1} S_{n,i}| \ge S'_r - 10\sqrt{n}\log^2 n > \sqrt{n},$$

  so $S_n \ne \sqrt{n}$, or
- there is j with $1 \le j \le r$ for which $S'_j < -10\log^2 n$, in which case

  $$S_{T_j} \le S'_j\sqrt{n} + |\sum_{i=1}^{k-1} S_{T_j,i}| \le S'_j\sqrt{n} + 10\sqrt{n}\log^2 n < 0,$$

  so Pos_n does not occur, or
- $S'_r = m$ for some m with $|m| \le 10\log^2 n$, but $S_n - S_t - \sqrt{n} + s \ne -s - m\sqrt{n}$, so $S_n \ne \sqrt{n}$.

Thus, by a union bound,

$$\mathbf{P}^c\left\{S_n=\sqrt{n},Pos_n\right\}\le\mathbf{P}^c\left\{\left(\bigcup_{|m|\le 10\log^2 n}B_m\right)\cup C\right\}\le\mathbf{P}^c\left\{C\right\}+\sum_{m=-10\log^2 n}^{10\log^2 n}\mathbf{P}^c\left\{B_m\right\}. \quad (48)$$

To bound the probabilities $\mathbf{P}^c\{B_m\}$, we first note that, denoting by A_m the event that $S'_r=m$ and $S'_j\ge -10\log^2 n$ for all $1\le j\le r$,

$$B_m=A_m\cap\{S_n-S_t-\sqrt{n}S'_r=\sqrt{n}-s-m\sqrt{n}\}.$$

Furthermore, A_m and $\{S_n-S_t-\sqrt{n}S'_r=\sqrt{n}-s-m\sqrt{n}\}$ are independent as they are determined by disjoint sections of the random walk, so for all m with $|m|\le 10\log^2 n$,

$$\begin{aligned}\mathbf{P}^c\left\{B_m\right\}&=\mathbf{P}^c\left\{A_m,S_n-S_t-\sqrt{n}S'_r=\sqrt{n}-s-m\sqrt{n}\right\}\\&=\mathbf{P}^c\left\{A_m\right\}\mathbf{P}^c\left\{S_n-S_t-\sqrt{n}S'_r=\sqrt{n}-s-m\sqrt{n}\right\}\end{aligned} \quad (49)$$

Now, given that *Good* occurs, S' is nothing but a symmetric simple random walk of length r; thus, by Bertrand's ballot theorem,

$$\mathbf{P}^c\left\{A_m\right\}=O\left(\frac{(m+10\log^2 n+1)(10\log^2 n+1)}{r^{3/2}}\right)=O\left(\frac{\log^4 n}{n^{3/8}}\right). \quad (50)$$

Also, given that *Good* occurs, $S_n-S_t-\sqrt{n}S'_r$ is a sum of $n-t-r=\Omega(n)$ i.i.d. integer-valued random variables that are never zero; Thus, by Theorem 2,

$$\mathbf{P}^c\left\{S_n-S_t-\sqrt{n}S'_r=\sqrt{n}-s-m\sqrt{n}\right\}=O\left(\frac{1}{\sqrt{n}}\right), \quad (51)$$

and combining (49), (50), and (51) yields that

$$\mathbf{P}^c\left\{B_m\right\}=O\left(\frac{\log^4 n}{n^{7/8}}\right). \quad (52)$$

For any j with $t\le j\le n$, an argument just as that leading to (26) shows that

$$\mathbf{P}^c\left\{|\sum_{i=1}^{k-1}S_{j,i}|\ge 10\sqrt{n}\log^2 n\right\}=O\left(\frac{1}{n^6}\right),$$

so

$$\mathbf{P}^c\left\{C\right\}\le\sum_{j=t}^{n}\mathbf{P}^c\left\{|\sum_{i=1}^{k-1}S_{j,i}|\ge 10\sqrt{n}\log^2 n\right\}=O\left(\frac{1}{n^5}\right), \quad (53)$$

and (48), (52), and (53) together yield

$$\mathbf{P}^c\left\{S_n=\sqrt{n},Pos_n\right\}=O\left(\frac{1}{n^5}\right)+\sum_{m=-10\log^2 n}^{10\log^2 n}O\left(\frac{\log^4 n}{n^{7/8}}\right)=O\left(\frac{\log^6 n}{n^{7/8}}\right). \quad (54)$$

Since t, s, and $\mathcal{I} \in \mathcal{Z}_t$ were arbitrary, (47) and (54) combine to give

$$\mathbf{P}\left\{S_n = \sqrt{n}, Pos_n, Pos_T\right\} = O\left(\frac{\sqrt{\log n}}{n^{3/8}}\right) \cdot O\left(\frac{\log^6 n}{n^{7/8}}\right) = O\left(\frac{\log^{13/2} n}{n^{5/4}}\right).$$

Finally, since if Pos_n occurs then either Pos_T occurs or $T > n$, by (38) we have

$$\begin{aligned}\mathbf{P}\left\{S_n = \sqrt{n}, Pos_n\right\} &\geq \mathbf{P}\left\{S_n = \sqrt{n}, Pos_n, Pos_T\right\} - \mathbf{P}\left\{T > n\right\} \\ &= O\left(\frac{\log^{13/2} n}{n^{5/4}}\right) - O\left(\frac{1}{n^5}\right) = O\left(\frac{\log^{13/2} n}{n^{5/4}}\right),\end{aligned}$$

as asserted in (30).

4 Conclusion

The results of this paper raise several questions. While our examples show that Theorem 1 is essentially best possible, is it perhaps possible that a ballot theorem holds for real-valued Markov chains with finite variance? Also, one observation about the example of Section 3.3 is that in that example, the step size X is not in the domain of attraction of *any* distribution. This leaves open the possibility that a ballot-style theorem may hold if the X has mean zero and is in the domain of attraction of some stable law. Such behavior seems unlikely but is not ruled out by our examples.

Finally, it would be very interesting to derive conditions on more general multisets $\mathcal{S}$ of n numbers summing to some value k which guarantee that, in a uniformly random permutation of $\mathcal{S}$, all partial sums are positive with probability of order k/n. Indeed, perhaps such work could end up not only generalizing the work of this paper, but perhaps unifying it with the existing discrete-time ballot theorems based on the “increasing stochastic process” perspective.

References

Erik Sparre Andersen. Fluctuations of sums of random variables. *Mathematica Scandinavica*, 1:263–285, 1953.

Erik Sparre Andersen. Fluctuations of sums of random variables ii. *Mathematica Scandinavica*, 2:195–223, 1954.

J. Bertrand. Solution d’un probleme. *Comptes Rendus de l’Academie des Sciences*, 105:369, 1887.

H. Chernoff. A measure of the asymptotic efficiency for tests of a hypothesis based on the sum of observables. *Ann. Math. Statist.*, 2:493–509, 1952.

William Feller. *An Introduction to Probability Theory and Its Applications, Volume 1*, volume 1. John Wiley & Sons, Inc, third edition, 1968a.

William Feller. *An Introduction to Probability Theory and Its Applications, Volume 2*. John Wiley & Sons, Inc, third edition, 1968b.

S. Janson, T. Łuczak, and A. Ruciński. *Random Graphs*. Wiley, New York, 2000.

Olav Kallenberg. *Foundations of Modern Probability*. Probability and Its Applications. Springer Verlag, second edition, 2003.

Olav Kallenberg. Ballot theorems and sojourn laws for stationary processes. *The Annals of Probability*, 27(4):2011–2019, 1999.

Harry Kesten. Sums of independent random variables–without moment conditions. *Annals of Mathematical Statistics*, 43(3):701–732, June 1972.

Robin Pemantle and Yuval Peres. Critical random walk in random environment on trees. *The Annals of Probability*, 23(1):105–140, 1995.

Frank Spitzer. A combinatorial lemma and its applications to probability theory. *Transactions of the American Mathematical Society*, 82(2):323–339, July 1956.

Charles J. Stone. On local and ratio limit theorems. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 217–224, 1965.

Lajos Takács. Ballot problems. *Zeitschrift für Warscheinlichkeitstheorie und verwandte Gebeite*, 1:154–158, 1962a.

Lajos Takács. A generalization of the ballot problem and its application in the theory of queues. *Journal of the American Statistical Association*, 57(298):327–337, 1962b.

Lajos Takács. The time dependence of a single-server queue with poisson input and general service times. *The Annals of Mathematical Statistics*, 33(4):1340–1348, December 1962c.

Lajos Takács. The distribution of majority times in a ballot. *Zeitschrift für Warscheinlichkeitstheorie und verwandte Gebeite*, 2(2):118–121, January 1963.

Lajos Takács. Combinatorial methods in the theory of dams. *Journal of Applied Probability*, 1(1):69–76, 1964a.

Lajos Takács. Fluctuations in the ratio of scores in counting a ballot. *Journal of Applied Probability*, 1(2):393–396, 1964b.

Lajos Takács. *Combinatorial Methods in the Theory of Stochastic Processes*. John Wiley & Sons, Inc, New York, NY, first edition, 1967.