RANDOM WALK 2−ARMED BANDITS

Tommaso Cesari[Υ]
EECS
University of Ottawa
Ottawa, Canada, K1N 6N5
tcesari@uottawa.ca

Roberto Colomboni[Υ]
DEIB
Politecnico di Milano
Department of Computer Science
Università degli Studi di Milano
Milano, Italy, 20133
roberto.colomboni@polimi.it

Christian Paravalos[Υ]
EECS
University of Ottawa
Ottawa, Canada, K1N 6N5
cpara006@uottawa.ca

Shivani Singal[Υ]
Department of Math & Stats
University of Ottawa
Ottawa, Canada, K1N 6N5
ssing294@uottawa.ca

November 28, 2025

ABSTRACT

In this preliminary (and unpolished) work, we study a non-stationary two-armed bandit in which one arm has an unbounded, drifting reward and the other arm is a safe baseline. The "walk" arm generates rewards according to a simple symmetric random walk, so that both its magnitude and total variation grow without bound. This regime falls outside the scope of standard non-stationary models, which typically assume bounded rewards and impose regularity conditions such as a total-variation budget, a finite number of change points, or a slow-drift constraint. In our setting, the reward of the "walk" arm can change by $\Omega(1)$ every round and classical guarantees do not apply. We introduce the Shivgorithm, a policy that alternates between querying the walk arm and playing the safe arm. We measure performance against a clairvoyant oracle that plays the best arm at each time and whose cumulative reward grows on the order of $T^{3/2}$. Our analysis shows that Shivgorithm achieves an $O(T^{5/6})$ regret rate. To the best of our knowledge, we achieve the first provably sublinear regret guarantees for the unbounded random-walk rewards problem.[1]

1 Setting

We consider a two-armed bandit over rounds $t = 1, 2, \ldots, T$, where $T \in \mathbb{N}$ is the time horizon.

Let $(\xi_t)_{t \geq 1}$ be an i.i.d. sequence of Rademacher random variables with $\mathbb{P}(\xi_t = 1) = \mathbb{P}(\xi_t = -1) = 1/2$. Define the partial sums $X_0 \coloneqq 0, X_t \coloneqq \sum_{s=1}^{t} \xi_s, t \geq 1$. Thus, $(X_t)_{t \geq 0}$ is a symmetric random walk on $\mathbb{Z}$. At each round t, the learner chooses an action A_t to select the "walk" arm or the "safe" arm, which we encode with 1 and 0 respectively. So, $A_t \in \mathcal{A} \coloneqq \{0, 1\}$. The realized reward at time t is

$$Y_t = \begin{cases} X_t, & \text{if } A_t = 1, \\ 0, & \text{if } A_t = 0. \end{cases}$$

The walk arm therefore has a time-varying mean equal to the random walk X_t, which grows unboundedly and has unbounded total variation, while the safe arm is constantly 0.

[Υ] Equal contribution.

[1]Note that this preprint has not been reviewed for publication yet and may contain some inaccuracies.

We can encode the actions by indicator variables: $I_t^{\text{walk}} := \mathbb{I}\{A_t = 1\}$, $I_t^{\text{safe}} := \mathbb{I}\{A_t = 0\} = 1 - I_t^{\text{walk}}$.

Within this setting, the online protocol is as follows.

Online Protocol

For each time step $t = 1, 2, \ldots, T$:
  1: The learner chooses an action $A_t \in \{0, 1\}$.
  2: Then learner earns the reward $Y_t = X_t I_t^{\text{walk}} + I_t^{\text{safe}} \cdot 0$.

Formally, we now describe the information structure of the interaction and the class of policies the learner may use.

Information Structure

The learner does not have direct access to $(X_t)_{t\geq 0}$; it only observes the action-reward pair (A_t, Y_t) from the round at time t. So, if the learner plays the safe arm at time t, it does not observe X_t, only a zero reward. For each $t \in [T]$, let

$$H_t := (A_1, Y_1, A_2, Y_2, \ldots, A_t, Y_t)$$

denote the interaction history up to time t.

A policy is a sequence of decision rules

$$\pi = (\pi_t)_{t=1}^T, \qquad \pi_t \colon \mathcal{H}_{t-1} \to \mathcal{A},$$

mapping histories to actions. $\mathcal{H}_{t-1}$ denotes the set of all possible histories of length $t - 1$. At time t, given history H_{t-1}, the policy selects $A_t := \pi_t(H_{t-1})$. Intuitively, A_t may depend on everything the learner has seen so far, but not on future rewards or future increments of the random walk.

All expectations $\mathbb{E}[\cdot]$ are taken with respect to the distribution of $(\xi_t)_{t\geq 1}$ and the randomness of the policy (if any).

Oracle and regret

As a benchmark, we consider a clairvoyant oracle that, at each round t, observes X_t and selects the arm with the larger instantaneous reward. Since the safe arm always yields 0, the oracle's per-round reward is $\max\{X_t, 0\}$, obtained by choosing the walk arm when $X_t > 0$ and the safe arm otherwise. The oracle's cumulative reward over horizon T is therefore $\sum_{t=1}^T \max\{X_t, 0\}$, which grows on the order of $T^{3/2}$ in expectation for a simple random walk. (See Appendix - Proposition 5.1)

Given a policy π, if the Y_t are generated according to π, its cumulative reward is

$$\sum_{t=1}^T Y_t = \sum_{t=1}^T (I_t^{\text{walk}} X_t + I_t^{\text{safe}} \cdot 0) = \sum_{t=1}^T I_t^{\text{walk}} X_t.$$

We define regret of π with respect to the oracle as

$$R_T := \sum_{t=1}^T \left(\max\{X_t, 0\} - I_t^{\text{walk}} X_t \right).$$

2   Challenges

This setting lies outside classical non-stationary bandit models in two respects:

1. **Unbounded growth of rewards.** The walk arm has reward X_t, which typically satisfies $|X_t| = \Theta(\sqrt{t})$ (Appendix - Proposition 5.2), so rewards grow unboundedly.
2. **Unbounded variation.** The sequence $(X_t)_{t\geq 1}$ has almost surely infinite total variation; there is no finite variation budget, no finite set of change points, and no slow-drift constraint.

As a consequence, existing non-stationary bandit algorithms do not directly apply in this regime. Moreover, the usual notion of bandit regret of comparing against the best fixed arm is not informative here since in expectation, both the walk and safe arms have mean zero. Instead, we compare against a much stronger, path-dependent benchmark, a clairvoyant oracle that observes X_t at every round and always plays the arm with larger instantaneous reward. The oracle's cumulative reward grows on the order of $T^{3/2}$ in expectation.

With this benchmark, the problem is intrinsically hard. It can be shown that there exists a universal constant $c > 0$ such that for any policy π, $\mathbb{E}[R_T] \geq c\sqrt{T}$, reflecting the fact that whenever $X_{t-1} = 0$, both actions incur expected instantaneous regret $1/2$, and the expected number of visits to 0 is $\Theta(\sqrt{T})$.

The next section introduces the Shivgorithm algorithm, a particular policy for choosing A_t, and analyzes its regret R_T relative to the oracle.

3 Algorithm

In this section we specify the policy we study, which we call Shivgorithm. At each round, Shivgorithm decides whether to probe the walk arm or to play the safe arm. The key idea is to space probes based on the last observed value of the random walk: after seeing a negative value $x \leq 0$, the algorithm waits for a number of rounds proportional to $x^2 / \ln(1/\varepsilon)$, chosen so that the probability of a zero-crossing during that waiting period is at most ε.

A guiding design principle is that Shivgorithm should be extremely simple and computationally lightweight. In particular, we restrict attention to policies that use only $O(1)$ memory and $O(1)$ arithmetic per round, and whose decisions can be expressed in closed form from the last observed value. Shivgorithm satisfies this: it maintains a single integer state variable and uses only basic integer operations. The goal of this work is not to characterize all optimal policies in this model, but to analyze a natural, easily implementable rule that still achieves sublinear regret in a highly non-stationary environment.

From now on we use the terms "probe" and "wait" instead of "play walk" and "play safe" respectively. We set $I_t^{\text{probe}} := I_t^{\text{walk}}$ (indicator that Shivgorithm probes at time t), and $I_t^{\text{wait}} := I_t^{\text{safe}}$ (indicator that Shivgorithm waits at time t), so that $I_t^{\text{probe}} + I_t^{\text{wait}} = 1$ for all t.

Shivgorithm takes as input the horizon $T \in \mathbb{N}$ and a parameter $\varepsilon \in (0, 1)$, which we interpret as a target upper bound on the probability that the random walk crosses zero during any single waiting period. The algorithm maintains a single state variable `next_pull`, denoting the next time index at which the walk arm will be probed. Initially, `next_pull` is set to 1.

Algorithm Shivgorithm

1: **Input:** Time horizon $T \in \mathbb{N}$; target flip-probability $\varepsilon \in (0, 1)$
2: **Initialize:** `next_pull` $\leftarrow 1$
3: **for** $t = 1, 2, \ldots, T$ **do**
4:    **if** $t \geq$ `next_pull` **then**
5:       **Pull** WALK; **obtain reward** X_t
6:       $\hat{X} \leftarrow X_t$
7:       **if** $\hat{X} > 0$ **then**
8:          $w \leftarrow 1$
9:       **else**
10:          $w \leftarrow \left\lfloor \dfrac{\hat{X}^2}{2 \ln(2/\varepsilon)} \right\rfloor$
      end if
11:       `next_pull` $\leftarrow t + w$
12:    **else**
13:       **Pull** SAFE; **obtain** 0
   end if
 end for

Thus, Shivgorithm probes the walk arm exactly at those times when the current time reaches the scheduled `next_pull`. After a positive probe, it continues probing every step; after a negative probe of depth $|X_t|$, it waits a number of rounds

quadratic in $|X_t|$, scaled by $\ln(1/\varepsilon)$, before probing again. The entire behavior is determined by the last observed value of the walk and the current time.

We now analyze the theoretical guarantees of Shivgorithm.

Theorem 3.1. *Consider the two-arm model of Section 1. Fix a time horizon $T \in \mathbb{N}$ and set $\varepsilon_T := \min\{1/2, T^{-5/3}\}$. If we run Shivgorithm with parameters T and ε_T, then*

$$\mathbb{E}[R_T] = \widetilde{O}\left(T^{5/6}\right).$$

We now turn to the proof of Theorem 3.1. Our starting point is a structural decomposition of the regret R_T into a probe component and a wait component, expressed in terms of the random probe times and the associated waiting intervals. We will show that the bound consists of two terms: the first arising from the cost of probing at negative depths; the second, from missed positive excursions during waiting periods.

3.1 Regret Decomposition

We can define the cumulative count of probes as $N_t := \sum_{s=1}^{t} I_s^{\text{probe}}$ (nondecreasing, integer-valued, $N_0 := 0$). Let $M := N_T$ be the total number of probes up to time T. The probe times are then $\tau_m := \inf\{t \geq 1 : N_t \geq m\}, m = 1, 2, \ldots, M$ so that $\tau_1 < \tau_2 < \cdots < \tau_M \leq T$ are the times at which we probe the walk arm. For every $t \geq 1$, we have the identity:

$$I_t^{\text{probe}} = N_t - N_{t-1} = \sum_{m=1}^{M} \mathbb{I}\{t = \tau_m\}. \tag{i}$$

We are particularly interested in probes at nonpositive values of the walk. The cumulative number of negative probes is given by $N_t^- := \sum_{s=1}^{t} \mathbb{I}\{I_s^{\text{probe}} = 1, X_s \leq 0\}$, (nondecreasing, $N_0^- := 0$). Let $\kappa_k := \inf\{t \geq 1 : N_t^- \geq k\}, k = 1, 2, \ldots$ be the time of the k-th negative probe. Let $K := N_T^-$ be the total number of negative probes up to time T. Then $\kappa_1 < \cdots < \kappa_K \leq T$ are exactly the probe times with $X_{\kappa_k} \leq 0$.

For each $k = 1, 2, \ldots, K$, the waiting period chosen by the algorithm after the k-th negative probe is

$$W_k := \left\lfloor \frac{X_{\kappa_k}^2}{2\ln(2/\varepsilon)} \right\rfloor \in \mathbb{N} \cup \{0\}.$$

The corresponding waiting interval is $\mathcal{I}_k := \{\kappa_k + 1, \ldots, \min(\kappa_k + W_k, T)\}$. By construction of the algorithm, these intervals are disjoint, and every time at which Shivgorithm waits belongs to exactly one such interval. Consequently, $I_t^{\text{wait}} = \sum_{k=1}^{K} \mathbb{I}\{t \in \mathcal{I}_k\}, t = 1, \ldots, T$.

Now, at time t, the oracle's reward is $\max\{X_t, 0\}$, while Shivgorithm's reward is $I_t^{\text{probe}} X_t + I_t^{\text{wait}} \cdot 0$. Thus, the total regret is given by

$$R_T := \sum_{t=1}^{T} \left(\max\{X_t, 0\} - (I_t^{\text{probe}} X_t + I_t^{\text{wait}} \cdot 0) \right).$$

Using $(x)_+ := \max\{x, 0\}$ and $(x)_- = \max\{-x, 0\}$, we can write the regret as

$$R_T = \sum_{t=1}^{T} I_t^{\text{probe}}(X_t)_- + \sum_{t=1}^{T} I_t^{\text{wait}}(X_t)_+ := R_{\text{probe}} + R_{\text{wait}},$$

where $R_{\text{probe}} := \sum_{t=1}^{T} I_t^{\text{probe}}(X_t)_-$ is the regret incurred when Shivgorithm probes at negative values, and $R_{\text{wait}} := \sum_{t=1}^{T} I_t^{\text{wait}}(X_t)_+$ is the regret incurred when Shivgorithm waits while the walk is positive.

The lemma below allows us to decompose the regret further in terms of the negative probes κ_k and the waiting interval $\mathcal{I}_k$. This decomposition will be the starting point for the upper-bound analysis.

Lemma 3.2. *The regret of Shivgorithm satisfies $R_T = R_{\text{probe}} + R_{\text{wait}}$ where*

$$R_{\text{probe}} = \sum_{k=1}^{K} |X_{\kappa_k}| \quad and \quad R_{\text{wait}} = \sum_{k=1}^{K} \sum_{t=\kappa_k+1}^{\min(\kappa_k+W_k, T)} (X_t)_+.$$

4

Proof.

$$R_T = \sum_{t=1}^{T} I_t^{\text{probe}} (X_t)_- \; + \; \sum_{t=1}^{T} I_t^{\text{wait}} (X_t)_+$$

$$= \sum_{t=1}^{T} (N_t - N_{t-1}) (X_t)_- \; + \; \sum_{t=1}^{T} I_t^{\text{wait}} (X_t)_+ \qquad \text{(using identity (i))}$$

$$= \sum_{t=1}^{T} \left(\sum_{m=1}^{M} \mathbb{I}\{t = \tau_m\} \right) (X_t)_- \; + \; \sum_{t=1}^{T} I_t^{\text{wait}} (X_t)_+ \qquad \text{(expand } N_t - N_{t-1})$$

$$= \sum_{m=1}^{M} (X_{\tau_m})_- \; + \; \sum_{t=1}^{T} I_t^{\text{wait}} (X_t)_+$$

$$= \sum_{k=1}^{K} |X_{\kappa_k}| \; + \; \sum_{t=1}^{T} \left(\sum_{k=1}^{K} \mathbb{I}\{t \in \mathcal{I}_k\} \right) (X_t)_+ \qquad \text{(using } (x)_- = |x|\mathbb{I}\{x \leq 0\}, \text{ and the definition of } I_t^{\text{wait}})$$

$$= \sum_{k=1}^{K} |X_{\kappa_k}| \; + \; \sum_{k=1}^{K} \sum_{t=\kappa_k+1}^{\min(\kappa_k+W_k, T)} (X_t)_+$$

$$=: R_{\text{probe}} \; + \; R_{\text{wait}}.$$

$\square$

3.2 Upper Bound

In this section we prove the regret bound in Theorem 3.1. using the structural decomposition from Lemma 3.3.

Bounding the wait regret R_{wait}

Proposition 3.3. *For any time horizon $T \in \mathbb{N}$ and $\varepsilon \in (0, 1/2]$, the wait component of the regret satisfies*

$$\mathbb{E}[R_{wait}] \leq C_0 \varepsilon T^{3/2},$$

for some absolute constant $C_0 > 0$ (independent of T and ε).

Regret during waiting periods can only occur if, while Shivgorithm is playing the safe arm, the random walk is actually positive. During the k-th waiting interval, this requires the walk to hit 0 and then make a positive excursion before the next probe. By construction of the waiting time W_k, the probability that a hit occurs in that interval is at most ε. Conditioned on such a hit, the expected regret suffered over the remaining time grows at most of order $W_k^{3/2}$. Summing this over all waiting intervals and using the fact that the total length of all waiting intervals is at most T gives the bound. We provide a formal proof below.

Proof. Define $R_{\text{wait}} = \sum_{k=1}^{K} \sum_{t \in \mathcal{I}_k} (X_t)_+$. For convenience, set $\kappa_0 := 0$ and $W_0 := 0$. For each $k \geq 1$, define the event that the walk hits 0 at least once during the k-th waiting interval: $A_k := \bigcup_{t=\kappa_k+1}^{\kappa_k+W_k} \{X_t = 0\}$.

For each k, recall that the k-th waiting interval is $\mathcal{I}_k := \{\kappa_k + 1, \ldots, \min(\kappa_k + W_k, T)\} \subseteq \{1, \ldots, T\}$. The intervals $\mathcal{I}_1, \ldots, \mathcal{I}_K$ are disjoint subsets of $\{1, \ldots, T\}$, so $\sum_{k=1}^{K} |\mathcal{I}_k| \leq T$.

Then, taking expectations, we have

$$\mathbb{E}[R_{\text{wait}}] = \mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=\kappa_k+1}^{\kappa_k+|\mathcal{I}_k|}(X_t)_+\right]$$

$$= \sum_{k=1}^{\infty}\mathbb{E}\left[\mathbb{I}\{k \leq K\}\sum_{t=\kappa_k+1}^{\kappa_k+|\mathcal{I}_k|}(X_t)_+\right]$$

$$\overset{(\star)}{=} \sum_{k=1}^{\infty}\mathbb{E}\left[\mathbb{P}(A_k)\cdot\mathbb{E}\left[\mathbb{I}\{k \leq K\}\sum_{t=\kappa_k+1}^{\kappa_k+|\mathcal{I}_k|}(X_t)_+ \,\Big|\, A_k\right] + \underbrace{\mathbb{P}(A_k^c)\cdot\mathbb{E}\left[\mathbb{I}\{k \leq K\}\sum_{t=\kappa_k+1}^{\kappa_k+|\mathcal{I}_k|}(X_t)_+ \,\Big|\, A_k^c\right]}_{=0}\right]$$

$$= \sum_{k=1}^{\infty}\mathbb{E}\left[\mathbb{P}(A_k)\cdot\mathbb{E}\left[\mathbb{I}\{\kappa_k \leq T\}\cdot\mathbb{E}\left[\sum_{t=\kappa_k+1}^{\kappa_k+|\mathcal{I}_k|}(X_t)_+ \,\Big|\, \kappa_k, |\mathcal{I}_k|, A_k\right] \,\Big|\, A_k\right]\right]$$

$$\leq \sum_{k=1}^{\infty}\mathbb{E}\left[\varepsilon\cdot\mathbb{E}\left[\mathbb{I}\{\kappa_k \leq T\}\, C_0|\mathcal{I}_k|^{3/2} \,\Big|\, A_k\right]\right] \qquad \text{(by Appendix Lemma 5.3 \& Lemma 5.4)}$$

$$\leq C_0\,\varepsilon\sum_{k=1}^{\infty}\mathbb{E}\left[\mathbb{I}\{\kappa_k \leq T\}\,|\mathcal{I}_k|^{3/2}\right]$$

$$= C_0\,\varepsilon\,\mathbb{E}\left[\sum_{k=1}^{K}|\mathcal{I}_k|^{3/2}\right]$$

$$\overset{(\star\star)}{\leq} C_0\,\varepsilon\,\mathbb{E}\left[\left(\sum_{k=1}^{K}|\mathcal{I}_k|\right)^{3/2}\right]$$

$$\leq C_0\,\varepsilon\,T^{3/2}.$$

Note that $(\star)$ arises since on the event A_k^c, the walk never hits 0 in the interval $\kappa_k + 1, \ldots, \kappa_k + W_k$, so it stays nonpositive throughout that interval, and hence $(X_t)_+ = 0$ for all those t. And $(\star\star)$ follows from the ℓ_p-norm inequality: $\left(\sum_{k=1}^{K}|\mathcal{I}_k|^{3/2}\right)^{2/3} = \|(|\mathcal{I}_1|, \ldots, |\mathcal{I}_K|)\|_{3/2} \leq \|(|\mathcal{I}_1|, \ldots, |\mathcal{I}_K|)\|_1 = \sum_{k=1}^{K}|\mathcal{I}_k|$. And so we have, $\sum_{k=1}^{K}|\mathcal{I}_k|^{3/2} \leq \left(\sum_{k=1}^{K}|\mathcal{I}_k|\right)^{3/2} \leq T^{3/2}$.

$\square$

Bounding the probe regret R_{probe}

Proposition 3.4. *Let $L := \ln(2/\varepsilon)$. For any time horizon $T \in \mathbb{N}$ and $\varepsilon \in (0, 1/2]$, the probe component of the regret satisfies*

$$\mathbb{E}[R_{probe}] \leq C_1 L^{2/3}T^{5/6} + C_2\varepsilon\sqrt{L}T^{5/2},$$

for some absolute constants $C_1, C_2 > 0$.

Proof. Recall the waiting time $W_k = \left\lfloor \frac{X_{\kappa_k}^2}{2L} \right\rfloor \in \mathbb{N}_0$ and the probe regret $R_{\text{probe}} = \sum_{k=1}^{K}|X_{\kappa_k}|$.

From the definition of W_k, we have

$$X_{\kappa_k}^2 \leq 2L(W_k + 1) \implies |X_{\kappa_k}| \leq \sqrt{2L(W_k + 1)} \leq 2\sqrt{L}(1 + \sqrt{W_k}) \implies R_{\text{probe}} \leq 2\sqrt{L}K + 2\sqrt{L}\sum_{k=1}^{K}\sqrt{W_k}.$$

Note: The contribution of the linear term K is of the same order as the final bound (See Appendix - Lemma 5.12). The main contribution comes from the $\sum\sqrt{W_k}$ term. For simplicity, we write

$$R_{\text{probe}} \leq c\sqrt{L}\sum_{k=1}^{K}\sqrt{W_k},$$

for some absolute constant $c > 0$. So, probe regret is controlled by the square roots of the waiting times.

We now group negative probes into "negative regions." Define the regions $[s_i, f_i]$ recursively as follows. Set $f_0 := 0$. For $i \geq 1$, we set s_i to be the first negative probe strictly after the previous region,

$$s_i := \min\{ t \in \{\kappa_1, \ldots, \kappa_K\} : f_{i-1} < t, \ X_t < 0 \},$$

and we set f_i to be the first subsequent probe at which the walk is nonnegative,

$$f_i := \min\{ t \in \{\kappa_1 + W_1 + 1, \ldots, \kappa_K + W_K + 1\} : s_i < t \leq T, \ X_t \geq 0 \}.$$

Let $I := \max\{i : s_i \leq T\}$ be the (random) number of negative regions up to time T. For each region i, define m_i as the number of negative probes in that region,

$$m_i := \sum_{k:\ \kappa_k \in [s_i, f_i]} 1,$$

and E_i as the sum of the waiting times attached to those probes,

$$E_i := \sum_{k:\ \kappa_k \in [s_i, f_i]} W_k.$$

By construction, every negative probe belongs to exactly one region. Hence,

$$R_{\text{probe}} \leq c\sqrt{L} \sum_{k=1}^{K} \sqrt{W_k}$$

$$= c\sqrt{L} \sum_{i=1}^{I} \sum_{k:\kappa_k \in [s_i, f_i]} \sqrt{W_k}$$

$$\leq c\sqrt{L} \sum_{i=1}^{I} \sqrt{\left(\sum_{k:\ \kappa_k \in [s_i, f_i]} 1 \right) \left(\sum_{k:\ \kappa_k \in [s_i, f_i]} W_k \right)} \qquad \text{(applying Cauchy–Schwarz inside each region)}$$

$$= c\sqrt{L} \sum_{i=1}^{I} \sqrt{m_i E_i}$$

Taking expectations:

$$\mathbb{E}[R_{\text{probe}}] \leq c\sqrt{L}\, \mathbb{E}\Big[\sum_{i=1}^{I} \sqrt{m_i E_i}\Big]$$

$$= c\sqrt{L} \sum_{i=1}^{\infty} \mathbb{E}\Big[\mathbb{I}\{i \leq I\}\, \sqrt{m_i E_i}\Big]$$

$$= c\sqrt{L} \sum_{i=1}^{\infty} \mathbb{E}\Big[\mathbb{I}\{s_i \leq T\}\, \sqrt{m_i E_i}\Big]$$

$$= c\sqrt{L} \sum_{i=1}^{\infty} \mathbb{E}\Big[\mathbb{I}\{s_i \leq T\}\, \mathbb{E}\Big[\sqrt{m_i E_i}\,\Big|\, s_i\Big]\Big]$$

$$\leq c\sqrt{L} \sum_{i=1}^{\infty} \mathbb{E}\Big[\mathbb{I}\{s_i \leq T\}\, \underbrace{\mathbb{E}\Big[\sqrt{m_1 E_1}\,\Big|\, s_1\Big]}_{\leq\ C\ L^{1/6}\ T^{1/3}\ +\ \varepsilon\, T^2}\Big] \qquad \text{(Appendix - Lemma 5.6)}$$

$$\leq c\sqrt{L}\Big(C\ L^{1/6}\ T^{1/3}\ +\ \varepsilon\, T^2\Big) \cdot \underbrace{\mathbb{E}[I]}_{C_2\sqrt{T}} \qquad \text{(Appendix - Lemma 5.7)}.$$

$$= CL^{2/3}T^{5/6} + C'\varepsilon\sqrt{L}T^{5/2}$$

for some absolute constants $C, C' > 0$.

$\square$

By Lemma 3.2 we have the decomposition $R_T = R_{\text{probe}} + R_{\text{wait}}$. Combining the bounds on R_{wait} and R_{probe} from the previous two propositions, we obtain that for any $T \in \mathbb{N}$ and any $\varepsilon \in (0, 1/2]$, with $L := \ln(2/\varepsilon)$, there exist absolute constants $C_0, C_1, C_2 > 0$ such that

$$\mathbb{E}[R_T] \leq C_0\,\varepsilon\,T^{3/2} + C_1\,L^{2/3}T^{5/6} + C_2\,\varepsilon\,\sqrt{L}\,T^{5/2}.$$

To turn this into a rate purely in terms of T, we now tune ε as a function of the horizon. Set $\varepsilon_T := \min\{1/2,\ T^{-5/3}\}$ and $L_T := \ln(2/\varepsilon_T)$. Then for all $T \geq 2$, we have $\mathbb{E}[R_T] \leq C_1'(\ln T)^{2/3}T^{5/6} + C_2'(\ln T)^{1/2}T^{5/6} + C_0\,T^{-1/6}$, for suitable absolute constants $C_1', C_2' > 0$. Absorbing the lower-order terms into the leading $(\ln T)^{2/3}T^{5/6}$ term, we conclude that there exists a constant $C > 0$ such that

$$\mathbb{E}[R_T] \leq C\,(\ln T)^{2/3}T^{5/6} \qquad \text{for all } T \geq 2.$$

Equivalently, $\mathbb{E}[R_T] = \widetilde{O}(T^{5/6})$, which completes the proof of Theorem 3.1.

4 Conclusion

The analysis conducted shows that, despite the unbounded drift and variation of the walk arm and the strength of the clairvoyant benchmark, a remarkably simple probing rule is sufficient to guarantee sublinear regret. Shivgorithm uses only the most recent observation and a single integer state variable to decide when to probe or wait, yet it achieves expected regret of order $T^{5/6}$ (up to logarithmic factors) against an oracle whose cumulative reward grows on the order of $T^{3/2}$. This provides, to the best of our knowledge, the first provable guarantees in this particular bandit setting. More broadly, our results suggest that carefully spacing information-gathering actions based on the current "depth" of the process can be an effective strategy in highly non-stationary environments.

5 Appendix

5.1 Missing Details in Setting

Proposition 5.1 (Oracle reward grows as $T^{3/2}$)**.** *Let $(\xi_t)_{t\geq 1}$ be an i.i.d. sequence of Rademacher random variables with $\mathbb{P}(\xi_t = 1) = \mathbb{P}(\xi_t = -1) = \frac{1}{2}$, and define the simple symmetric random walk $X_0 := 0$, $X_t := \sum_{s=1}^{t} \xi_s\ (t \geq 1)$. Then, for all integers $T \geq 1$,*

$$\mathbb{E}\left[\sum_{t=1}^{T}(X_t)_+\right] = \Theta\left(T^{3/2}\right).$$

Proof. We begin by providing an upper bound.

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}(X_t)_+\right] &= \sum_{t=1}^{T}\mathbb{E}\left[\max\{0, X_t\}\right]\\
&\overset{(\star)}{=} \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\left[|X_t|\right]\\
&\leq \frac{1}{2}\sum_{t=1}^{T}\sqrt{\mathbb{E}\left[X_t^2\right]}\\
&= \frac{1}{2}\sum_{t=1}^{T}\sqrt{\text{Var}\left[X_t\right]}\\
&= \frac{1}{2}\sum_{t=1}^{T}\sqrt{t}\\
&\leq \frac{1}{2}T^{3/2},
\end{aligned}$$

where $(\star)$ follows by the symmetry of the random walk.

Proof of $(\star)$:

Let $X_t^+ = \max\{0, X_t\}$ and $X_t^- = \max\{0, -X_t\}$. Then $|X_t| = X_t^+ + X_t^-$. Since the random walk is symmetric, $X_t \overset{d}{=} -X_t$. Applying this to the positive component, $\mathbb{E}[X_t^+] = \mathbb{E}[(-X_t)^+] = \mathbb{E}[X_t^-]$.

Therefore, $\mathbb{E}|X_t| = \mathbb{E}[X_t^+] + \mathbb{E}[X_t^-] = 2\,\mathbb{E}[X_t^+] \implies \mathbb{E}[X_t^+] = \frac{1}{2}\,\mathbb{E}|X_t|$.

Now we prove the lower bound.

$$\mathbb{E}\left[\sum_{t=1}^{T}(X_t)_+\right] = \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}[|X_t|] \overset{(\star\star)}{\geq} \frac{1}{24\sqrt{2}}\sum_{t=1}^{T}\sqrt{t} \geq \frac{1}{24\sqrt{2}}\int_0^T \sqrt{x}\,dx = \frac{1}{24\sqrt{2}}\cdot\frac{2}{3}T^{3/2}.$$

We obtain $(\star\star)$ since for all $t \geq 1$, we have

$$\mathbb{E}[|X_t|] \geq \sqrt{\frac{t}{2}}\mathbb{P}\left(|X_t| \geq \sqrt{t/2}\right) \overset{(\star\star\star)}{\geq} \sqrt{\frac{t}{2}}\cdot\frac{1}{12} = \frac{1}{12\sqrt{2}}\sqrt{t},$$

where $(\star\star\star)$ is attained using the Paley-Zygmund inquality. Note that $\mathbb{E}[X_t^2] = t$ and $\mathbb{E}[X_t^4] = 3t^2 - 2t \leq 3t^2$. So,

$$\mathbb{P}\left(|X_t| \geq \sqrt{t/2}\right) = \mathbb{P}\left(X_t^2 \geq \frac{t}{2}\right) \geq (1 - 1/2)^2\frac{(\mathbb{E}[X_t^2])^2}{\mathbb{E}[X_t^4]} = (1 - 1/2)^2\frac{t^2}{3t^2} = \frac{1}{12}.$$

$\square$

5.2   Missing Details in Challenges

Proposition 5.2. *Let $(\xi_s)_{s\geq 1}$ be i.i.d. Rademacher random variables with $\mathbb{P}(\xi_s = 1) = \mathbb{P}(\xi_s = -1) = 1/2$, and define the simple symmetric random walk $X_t := \sum_{s=1}^{t}\xi_s$, $t \geq 1$. Then there exist constants $0 < c_1 \leq c_2 < \infty$ such that*

$$c_1\sqrt{t} \leq \mathbb{E}|X_t| \leq c_2\sqrt{t} \qquad \text{for all } t \geq 1.$$

In particular, $\mathbb{E}|X_t| = \Theta(\sqrt{t})$ as $t \to \infty$.

Proof. We begin by proving the upper bound. By Cauchy–Schwarz, we have

$$\mathbb{E}|X_t| \leq \sqrt{\mathbb{E}[X_t^2]} = \sqrt{\mathrm{Var}(X_t)} = \sqrt{\sum_{s=1}^{t}\mathrm{Var}(\xi_s)} = \sqrt{t\cdot 1} = \sqrt{t}.$$

We now prove the lower bound.

Fix $t \geq 1$. For any $a > 0$, we have $\mathbb{E}|X_t| \geq a\,\mathbb{P}(|X_t| \geq a)$, since $|X_t| \geq a\,\mathbb{I}\{|X_t| \geq a\}$. Choose $a_t := \sqrt{t/2}$. Then

$$\mathbb{E}|X_t| \geq \sqrt{\frac{t}{2}}\,\mathbb{P}\left(\frac{|X_t|}{\sqrt{t}} \geq \frac{1}{\sqrt{2}}\right).$$

Define the normalized walk $Z_t := X_t/\sqrt{t}$. By the central limit theorem, $Z_t \xrightarrow{t\to\infty} Z$, $Z \sim \mathcal{N}(0, 1)$. Hence

$$\mathbb{P}\left(|Z_t| \geq \frac{1}{\sqrt{2}}\right) \xrightarrow{t\to\infty} \mathbb{P}\left(|Z| \geq \frac{1}{\sqrt{2}}\right) =: p_0,$$

where $p_0 \in (0, 1)$ is a fixed constant. Therefore, there exists $t_0 \in \mathbb{N}$ such that for all $t \geq t_0$, $\mathbb{P}(|Z_t| \geq 1/\sqrt{2}) \geq \frac{p_0}{2}$. For all $t \geq t_0$, we then have

$$\mathbb{E}|X_t| \geq \sqrt{\frac{t}{2}}\,\mathbb{P}\left(|Z_t| \geq \frac{1}{\sqrt{2}}\right) \geq \sqrt{\frac{t}{2}}\cdot\frac{p_0}{2} = \left(\frac{p_0}{2\sqrt{2}}\right)\sqrt{t}.$$

For the finitely many $1 \leq t < t_0$, define $c_1' := \min_{1\leq t<t_0}\mathbb{E}|X_t|/\sqrt{t}$ and set $c_1 := \min\left\{c_1', p_0/(2\sqrt{2})\right\}$. Therefore, for all $t \geq 1$, $\mathbb{E}|X_t| \geq c_1\sqrt{t}$.

$\square$

5.3 Missing Details in Bounding the wait regret R_{wait}

Lemma 5.3 (Probability of a hit in a waiting interval). *For each $k \geq 1$, let $W_k = \left\lfloor \frac{X_{\kappa_k}^2}{2 \ln(2/\varepsilon)} \right\rfloor$, where κ_k is the k-th negative probe time. Define the event that the walk hits 0 during the k-th waiting interval (trucated at horizon T) by*

$$A_k := \bigcup_{t=\kappa_k+1}^{\kappa_k+|\mathcal{I}_k|} \{X_t = 0\}.$$

Then, for all $k \geq 1$, we have $\mathbb{P}(A_k) \leq \varepsilon$.

Proof. Fix k and condition on $X_{\kappa_k} = x \leq 0$. Define the increment walk $S_t := X_{\kappa_k+t} - X_{\kappa_k} = \sum_{i=\kappa_k+1}^{\kappa_k+t} \xi_i, t \geq 0$. By the strong Markov property, at stopping time κ_k (See Appendix - Lemma 5.5), $(S_t)_{t \geq 0}$ is a simple symmetric random walk started at 0.

By the definition of A_k, we have

$$
\begin{aligned}
\mathbb{P}[A_k \mid X_{\kappa_k} = x] &= \mathbb{P}[\exists t \in \{\kappa_k + 1, \ldots, \kappa_k + |\mathcal{I}_k|\} : X_t = 0 \mid X_{\kappa_k} = x] \\
&= \mathbb{P}[\exists t \in \{\kappa_k + 1, \ldots, \kappa_k + |\mathcal{I}_k|\} : X_t \geq 0 \mid X_{\kappa_k} = x] \\
&= \mathbb{P}\left[\sup_{\kappa_k+1 \leq t \leq \kappa_k+|\mathcal{I}_k|} X_t \geq 0 \,\middle|\, X_{\kappa_k} = x\right] \\
&= \mathbb{P}\left[\max_{\kappa_k+1 \leq t \leq \kappa_k+|\mathcal{I}_k|} X_t \geq 0 \,\middle|\, X_{\kappa_k} = x\right] \\
&= \mathbb{P}\left[\max_{1 \leq t \leq |\mathcal{I}_k|} (x + S_t) \geq 0\right] \\
&= \mathbb{P}\left[\max_{1 \leq t \leq |\mathcal{I}_k|} S_t \geq -x\right] \\
&\leq 2\,\mathbb{P}\left[S_{|\mathcal{I}_k|} \geq -x\right] \qquad \text{(Reflection principle)} \\
&\leq 2\exp\left(-\frac{2x^2}{4|\mathcal{I}_k|}\right), \qquad \text{(Hoeffding's Inequality)} \\
&\leq 2\exp\left(-\frac{x^2}{2W_k}\right) \qquad \text{(since } V_k \leq W_k\text{)}
\end{aligned}
$$

By the choice of the waiting time, $W_k = \left\lfloor \frac{x^2}{2\ln(2/\varepsilon)} \right\rfloor$, this bound is at most ε, i.e., $\mathbb{P}[A_k \mid X_{\kappa_k} = x] \leq \varepsilon$ for all $x \leq 0$.

Then, averaging over X_{κ_k} gives

$$\mathbb{P}(A_k) = \sum_x \mathbb{P}[A_k \mid X_{\kappa_k} = x]\,\mathbb{P}[X_{\kappa_k} = x] \leq \varepsilon \sum_x \mathbb{P}[X_{\kappa_k} = x] = \varepsilon.$$

$\square$

Lemma 5.4 (Expected regret conditioned on a hit). *For each $k \geq 1$, let A_k be the event that the walk hits 0 at least once during the k-th waiting interval $\mathcal{I}_k$, and let W_k be the waiting time. Then*

$$\mathbb{E}\left[\sum_{t=\kappa_k+1}^{\kappa_k+|\mathcal{I}_k|} (X_t)_+ \,\middle|\, \kappa_k,\, |\mathcal{I}_k|,\, A_k\right] \leq C_0\,|\mathcal{I}_k|^{3/2}$$

for some absolute constant $C_0 > 0$ (independent of k, T, and ε).

Proof. On the event A_k, the walk hits 0 at some random time $\sigma_k \in \mathcal{I}_k$. Let σ_k be the first such hitting time in the interval, and define the residual length $D_k := (\kappa_k + |\mathcal{I}_k|) - \sigma_k$. Once the walk hits 0 at σ_k, the future increments are

independent of the past and form a fresh symmetric random walk. Thus, starting from σ_k, the sequence $(X_{\sigma_k+t})_{t\geq 0}$ has the same distribution as $(S_t)_{t\geq 0}$ started at 0.

Conditioning on (κ_k, W_k, A_k), the expected regret suffered from step κ_k up to $\kappa_k + |\mathcal{I}_k|$ is

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=\kappa_k+1}^{\kappa_k+|\mathcal{I}_k|}(X_t)_+ \,\middle|\, \kappa_k, |\mathcal{I}_k|, A_k\right] &= \mathbb{E}\left[\sum_{t=\sigma_k}^{\kappa_k+|\mathcal{I}_k|}(X_t)_+ \,\middle|\, \kappa_k, |\mathcal{I}_k|, A_k\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{D_k}(S_t)_+ \,\middle|\, \kappa_k, |\mathcal{I}_k|, A_k, \sigma_k\right] \\
&\overset{(\star)}{\leq} C_0 D_k^{3/2} \\
&\leq C_0 |\mathcal{I}_k|^{3/2} \qquad (\text{since } D_k \leq |\mathcal{I}_k|.)
\end{aligned}
$$

Note that we have $(\star)$ since for a simple symmetric random walk, it is known that $\mathbb{E}[(S_t)_+] = \sqrt{t/(2\pi)}$. Therefore,

$$
\sum_{t=1}^{D_k}\mathbb{E}[(S_t)_+] \leq \sum_{t=1}^{D_k}\sqrt{\frac{t}{2\pi}} \leq C_0 D_k^{3/2}
$$

for some absolute constant $C_0 > 0$, using the bound $\sum_{t=1}^n \sqrt{t} \leq C\, n^{3/2}$.

$\square$

Lemma 5.5 (Negative probes are stopping times)**.** *Let $(\xi_t)_{t\geq 1}$ be i.i.d. Rademacher random variables. Let $X_0 := 0, X_t := \sum_{s=1}^t \xi_s, t \geq 1$, so that $(X_t)_{t\geq 0}$ is a simple symmetric random walk on $\mathbb{Z}$. Let $\mathcal{F}_t := \sigma(X_0, \xi_1, \ldots, \xi_t) = \sigma(X_0, X_1, \ldots, X_t)$ be the natural filtration.*

Assume that the probe indicators I_t^{probe} are $\mathcal{F}_{t-1}$-measurable for all $t \geq 1$. Define the cumulative number of negative probes up to time t by

$$
N_t^- := \sum_{s=1}^t \mathbf{1}\{I_s^{\mathrm{probe}} = 1,\ X_s \leq 0\}, \quad t \geq 1,
$$

and for each integer $k \geq 1$ define $\kappa_k := \inf\{t \geq 1 : N_t^- \geq k\}$. Then, for every fixed $k \in \mathbb{N}$, the random time κ_k is a stopping time with respect to the filtration $(\mathcal{F}_t)_{t\geq 0}$.

Moreover, if we fix a horizon $T \in \mathbb{N}$ and define $K := N_T^-$ (the total number of negative probes up to time T), then on the event $\{k \leq K\}$ we have $\kappa_k \leq T$.

Proof. (We first look at the measurability of the negative-probe counts.) For each time $s \geq 1$, we have that X_s is $\mathcal{F}_s$-measurable by construction. And by assumption, I_s^{probe} is $\mathcal{F}_{s-1}$-measurable, hence also $\mathcal{F}_s$-measurable. Therefore the indicator $\mathbb{I}\{I_s^{\mathrm{probe}} = 1,\ X_s \leq 0\}$ is $\mathcal{F}_s$-measurable. It follows that for each $t \geq 1$, the sum N_t^- is $\mathcal{F}_t$-measurable. By definition, $(N_t^-)_{t\geq 1}$ is non-decreasing in t.

(Now, we look at the stopping-time property.) Fix $k \in \mathbb{N}$. By definition, $\kappa_k := \inf\{t \geq 1 : N_t^- \geq k\}$. For any integer $t \geq 1$, the event $\{\kappa_k \leq t\}$ can be written as $\{\kappa_k \leq t\} = \{\exists\, s \leq t : N_s^- \geq k\}$. Since $(N_s^-)_{s\geq 1}$ is nondecreasing in s, this is equivalent to $\{\kappa_k \leq t\} = \{N_t^- \geq k\}$. And, as N_t^- is $\mathcal{F}_t$-measurable, we have $\{N_t^- \geq k\} \in \mathcal{F}_t$. Therefore $\{\kappa_k \leq t\} \in \mathcal{F}_t$ for all $t \geq 1$, which shows that κ_k is a stopping time with respect to $(\mathcal{F}_t)$.

(Finally we look at the boundedness up to the horizon.) Fix $T \in \mathbb{N}$ and set $K := N_T^-$, the total number of negative probes up to time T. On the event $\{k \leq K\}$, there are at least k negative probes among times $1, \ldots, T$. Equivalently, there exists some time $s \leq T$ such that $N_s^- \geq k$. By the definition of κ_k as the first time when N_t^- reaches level k, this implies $\kappa_k \leq s \leq T$ on $\{k \leq K\}$. In particular, on the event $\{k \leq K\}$, the stopping time κ_k is bounded above by the deterministic constant T.

$\square$

5.4 Missing Details in Bounding the probe regret R_{probe}

Lemma 5.6. *Let $L = \ln(2/\varepsilon)$. Consider the first negative region $[s_1, f_1]$, with m_1 the number of negative probes in the region and E_1 the total waiting time in the region. Then there exists an absolute constant $C > 0$ such that, for every realization of s_1 with $s_1 \leq T$,*

$$\mathbb{E}\big[\sqrt{m_1 E_1} \,\big|\, s_1\big] \;\leq\; C\, L^{1/6}\, T^{1/3} \;+\; \varepsilon\, T^2.$$

Proof. Let $(X'_t)_{t \geq 0}$ be an independent copy of the random walk, started at $X'_0 = 0$, and run the same algorithm on this copy from time 0 up to horizon T. Let $[s'_1, f'_1]$ be the first negative region of this canonical system and let m'_1 be the number of negative probes and E'_1 the total waiting time.

By the strong Markov property at stopping time s_1, the process $(X_{s_1+t} - X_{s_1})_{t \geq 0}$ is a fresh simple symmetric random walk, independent of the past, and has the same distribution as $(X_t)_{t \geq 0}$. So, the process "viewed from s_1" has the same distribution as the canonical process "viewed from time 0." In particular, for every fixed value of s_1, $(m_1, E_1)|(s_1) \overset{d}{=} (m'_1, E'_1)$. So, $\mathbb{E}[\sqrt{m_1 E_1}|s_1] = \mathbb{E}[\sqrt{m'_1 E'_1}]$. So, it suffices to bound $\mathbb{E}[\sqrt{m'_1 E'_1}]$.

Let the first return time to 0 after entering the region, measured from its start be

$$\eta' \;=\; \min\{\,t:\; t > s'_1,\; X_t = 0\,\} \;-\; s'_1,$$

and its truncation

$$\eta'_T \;=\; \min\{\eta',\, T\}.$$

Let the total length of the first negative region (waiting plus probe steps) be

$$T'_1 \;=\; E'_1 + m'_1.$$

Define the "bad" event that the negative region continues after the first hit of 0:

$$H'_1 \;=\; \{\, T'_1 \neq \eta'_T \,\}.$$

Let m''_1 be the number of negative probes that occur before time $s'_1 + \eta'_T$:

$$m''_1 \;=\; \sum_{k:\; \kappa'_k \in [s'_1,\, s'_1 + \eta'_T]} 1.$$

On the event $H^{c\prime}_1$, the region ends exactly when the walk first returns to 0, (or at T if that happens earlier), so $T'_1 = \eta'_T$ and all negative probes in the region occur before that time. Hence on $H^{c\prime}_1$ we have $m'_1 = m''_1$ and $E'_1 \leq T'_1 = \eta'_T$.

So, we have the following:

$$
\begin{aligned}
\mathbb{E}\Big[\sqrt{m'_1 E'_1}\Big] &\leq \mathbb{E}\Big[\sqrt{m'_1 T'_1}\Big] \\[4pt]
&= (1 - \mathbb{P}[H'_1])\, \mathbb{E}\Big[\sqrt{m'_1 T'_1}\,\Big|\, H^{c\prime}_1\Big] \;+\; \mathbb{P}[H'_1]\, \mathbb{E}\Big[\sqrt{m'_1 T'_1}\,\Big|\, H'_1\Big] \\[4pt]
&\leq (1 - \mathbb{P}[H'_1])\, \mathbb{E}\Big[\sqrt{m'_1 T'_1}\,\Big|\, H^{c\prime}_1\Big] \;+\; \varepsilon\, T\, \mathbb{E}\Big[\sqrt{m'_1 T'_1}\,\Big|\, H'_1\Big] \quad \text{(See Appendix - Lemma 5.8)} \\[4pt]
&\leq (1 - \mathbb{P}[H'_1])\, \mathbb{E}\Big[\sqrt{m'_1 T'_1}\,\Big|\, H^{c\prime}_1\Big] \;+\; \varepsilon\, T^2 \\[4pt]
&= (1 - \mathbb{P}[H'_1])\, \mathbb{E}\Big[\sqrt{m''_1 \eta'_T}\,\Big|\, H^{c\prime}_1\Big] \;+\; \varepsilon\, T^2 \\[4pt]
&\leq \mathbb{E}\Big[\sqrt{m''_1 \eta'_T}\Big] \;+\; \varepsilon\, T^2 \\[4pt]
&\leq \sqrt{\underbrace{\mathbb{E}[\eta'_T]}_{\leq C_1 \sqrt{T}}} \sqrt{\underbrace{\mathbb{E}[m''_1]}_{\leq C_2 (L\sqrt{T})^{1/3}}} \;+\; \varepsilon\, T^2 \quad \text{(See Appendix - Lemma 5.10 \& Lemma 5.9)} \\[4pt]
&\leq C\, L^{1/6}\, T^{1/3} \;+\; \varepsilon\, T^2,
\end{aligned}
$$

for some universal constant $C > 0$.

Therefore, $\mathbb{E}[\sqrt{m_1 E_1}|s_1] = \mathbb{E}[\sqrt{m'_1 E'_1}] \leq C\, L^{1/6}\, T^{1/3} \;+\; \varepsilon\, T^2$.

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \square$

Lemma 5.7. *For some absolute constant $C_2 > 0$, the expected number of negative regions satisfies*

$$\mathbb{E}[I] \le C_2 \sqrt{T}.$$

Proof. We bound $\mathbb{E}[I]$, the number of regions. Each negative region starts immediately after a visit of the walk to 0. So the number of regions I is at most the number of visits to 0 up to time T:

$$I \le \sum_{t=0}^{T} \mathbb{I}\{X_t = 0\} \implies \mathbb{E}[I] \le \sum_{t=0}^{T} \mathbb{P}\{X_t = 0\}.$$

For a simple symmetric walk, $\mathbb{P}(X_{2n} = 0) \le 1/\sqrt{\pi n}$ for all $n \ge 1$, and $\mathbb{P}(X_{2n+1} = 0) = 0$ for all $n \ge 0$. Therefore,

$$\mathbb{E}[I] \le \sum_{t=0}^{T} \mathbb{P}(X_t = 0) = \sum_{n=0}^{\lfloor T/2 \rfloor} \mathbb{P}(X_{2n} = 0) \le C_2 \sqrt{T},$$

for some absolute constant C_2.

$\square$

Lemma 5.8. *Let H_1 be the event that the first negative region continues after the first return to 0, and let A_k be the event that the walk hits 0 during the k-th waiting interval. Then $\mathbb{P}(H_1) \le \varepsilon T$.*

Proof.

$$\mathbb{P}[H_1] \le \mathbb{P}\left[\bigcup_{k:\kappa_k \in [s_1, f_1]} \{A_k\} \right] \le \mathbb{P}\left[\bigcup_{k=1}^{T} \{A_k\} \right] \le \sum_{k=1}^{T} \mathbb{P}\left[A_k\right] \le \varepsilon T,$$

where we have $\mathbb{P}(A_k) \le \varepsilon$ from the earlier R_{wait} analysis.

$\square$

Lemma 5.9. *Let $L = \ln(2/\varepsilon)$ and m_1 be the number of negative probes in the first negative region. Then there exists an absolute constant $C > 0$ such that*

$$\mathbb{E}[m_1] \le C \left(L\sqrt{T}\right)^{1/3}.$$

Proof. Fix a region $[s_i, f_i]$. Let $\eta^{(i)} := \min\{t > 0 : X_{s_i + t} = 0\}$ and $\eta_T^{(i)} := \min\{\eta^{(i)}, T\}$.

Define

$$m_i' := \sum_{k:\, \kappa_k \in [s_i,\, s_i + \eta_T^{(i)}]} 1,$$

the number of negative probes in this region before the random walk first hits 0 (or time T, whichever occurs first). Clearly $m_i \le m_i'$, so it suffices to bound $\mathbb{E}[m_i']$.

We introduce a depth cutoff $D \in \mathbb{N}$ and split "shallow" negative probes at depths $-1, \ldots, -D$ and "deep" negative probes below $-D$. Formally,

$$m_{i, \ge -D}' := \left| \{\kappa_k \in [s_i, s_i + \eta_T^{(i)}] :\ X_{\kappa_k} \in \{-1, \ldots, -D\}\} \right|, \quad m_{i, < -D}' := m_i' - m_{i, \ge -D}'.$$

For each depth $d \in \{1, \ldots, D\}$, let $N_{-d}^{(i)}$ be the number of visits to level $-d$ during the time interval $[s_i,\ s_i + \eta_T^{(i)}]$. Every shallow negative probe occurs at some level $-d$ with $1 \le d \le D$, so

$$m_{i, \ge -D}' \le \sum_{d=1}^{D} N_{-d}^{(i)} \implies \mathbb{E}[m_{i, \ge -D}'] \le \sum_{d=1}^{D} \mathbb{E}\left[N_{-d}^{(i)}\right] \overset{\star}{\le} 2D,$$

where $(\star)$ follows from applying Appendix - Lemma 5.11 to the excursion starting at s_i for every $d \ge 1$.

Now, we enumerate the "deep" negative probes in this region as

$$\kappa_1^{(<-D)} < \cdots < \kappa_{m_{i, <-D}'}^{(<-D)},$$

13

where $X_{\kappa_r(<-D)} \leq -(D+1)$ for each r. For each such probe, the waiting time satisfies

$$|X_{\kappa_r^{(<-D)}}| \geq D+1 \implies W_{\kappa_r^{(<-D)}} \geq \frac{D^2}{2L}$$

$$\implies \kappa_{r+1}^{(<-D)} \geq \kappa_r^{(<-D)} + \frac{D^2}{2L}$$

$$\implies \kappa_{m'_{i,<-D}}^{(<-D)} - \kappa_1^{(<-D)} \geq (m'_{i,<-D} - 1)\frac{D^2}{2L}$$

$$\implies (m'_{i,<-D} - 1)\frac{D^2}{2L} \leq \eta_T^{(i)} \quad \text{(since all these deep probes occurs before time } s_i + \eta_T^{(i)})$$

$$\implies m'_{i,<-D} \leq 1 + \frac{2L}{D^2}\eta_T^{(i)}$$

$$\implies \mathbb{E}\left[m'_{i,<-D}\right] \leq 1 + \frac{2L}{D^2}\mathbb{E}\left[\eta_T^{(i)}\right]$$

$$\implies \mathbb{E}\left[m'_{i,<-D}\right] \leq 1 + \frac{2C_1 L}{D^2}\sqrt{T} \quad \text{(by lemma)}$$

Then,

$$\mathbb{E}[m_i] \leq \mathbb{E}[m'_i] = \mathbb{E}\left[m'_{i,\geq -D}\right] + \mathbb{E}\left[m'_{i,<-D}\right] \leq 2D + 1 + \frac{2C_1 L}{D^2}\sqrt{T}.$$

Optimizing over D with $D := \left(4\,C_1 L\sqrt{T}\right)^{1/3}$ yields $\mathbb{E}[m_i] \leq C_2\,(L\sqrt{T})^{1/3}$ for all regions i. This bound is uniform over regions i.

$\square$

Lemma 5.10 (Truncated return time to 0)**.** *Let η be the first return time to 0 for SRW started at -1, and $\eta_T := \min\{\eta, T\}$. Then there exists an absolute constant C_1 such that $\mathbb{E}[\eta_T] \leq C_1\sqrt{T}$ for all T.*

Proof. By reflection/ballot, for $n \geq 1$, $\mathbb{P}(\eta > 2n) = \binom{2n}{n}/2^{2n} \leq (\pi n)^{-1/2}$. Hence

$$\mathbb{E}[\eta_T] = \sum_{t=0}^{T-1} \mathbb{P}(\eta > t) \leq 1 + 2\sum_{n=1}^{\lfloor T/2 \rfloor} (\pi n)^{-1/2} \leq C_1\sqrt{T}.$$

$\square$

Lemma 5.11 (Green's function on the half-line)**.** *Let $(Y_t)_{t\geq 0}$ be a simple symmetric random walk on $\{0, 1, 2, \dots\}$, started at $Y_0 = 1$, with 0 absorbing. For any $d \geq 1$, let $N_d := |\{t \geq 0 : Y_t = d \text{ before the first visit to } 0\}|$ be the number of visits to d before absorption. Then $\mathbb{E}_1[N_d] = 2$. Equivalently, for the reflected walk $X_t := -Y_t$ on $\{0, -1, -2, \dots\}$ started at -1 and absorbed at 0, the expected number of visits to $-d$ before hitting 0 is also 2.*

Proof. For each starting state $i \geq 0$, define $a_i := \mathbb{E}_i[N_d]$. By definition $a_0 = 0$. For every $i \geq 1$, a first-step decomposition gives

$$a_i = \mathbf{1}\{i = d\} + \frac{1}{2}a_{i-1} + \frac{1}{2}a_{i+1},$$

or equivalently $a_{i+1} - 2a_i + a_{i-1} = -2\,\mathbb{I}\{i = d\}$, $i \geq 1$.

First consider the region $1 \leq i \leq d-1$. There the right-hand side of $-2\,\mathbb{I}\{i = d\} = 0$, so

$$a_{i+1} - 2a_i + a_{i-1} = 0, \qquad 1 \leq i \leq d-1.$$

The general solution of this homogeneous second-order difference equation is affine:

$$a_i = A + Bi, \qquad 0 \leq i \leq d.$$

Using $a_0 = 0$ gives $A = 0$, so $a_i = Bi$ for $0 \leq i \leq d$.

14

Next consider $i \geq d$. From any starting point $i \geq d$, the walk must hit d before it can hit 0. After the first visit to d, the future evolution is that of the walk started from d. Hence the expected number of visits to d is the same for all $i \geq d$. So, $a_i = a_d$ for all $i \geq d$. In particular, $a_{d+1} = a_d$.

Then at $i = d$, we have:
$$a_d = 1 + \frac{1}{2}a_{d-1} + \frac{1}{2}a_{d+1}.$$

Substituting $a_{d-1} = B(d-1)$ and $a_d = a_{d+1} = Bd$ yields
$$Bd = 1 + \frac{1}{2}B(d-1) + \frac{1}{2}Bd \quad \Longrightarrow \quad 2Bd = 2 + B(2d-1) \quad \Longrightarrow \quad B = 2.$$

Thus $a_i = 2i$ for $0 \leq i \leq d$, and $a_i = a_d = 2d$ for all $i \geq d$. In particular,
$$\mathbb{E}_1[N_d] = a_1 = 2.$$

For the reflected walk $X_t := -Y_t$ on $\{0, -1, -2, \dots\}$ started at -1 and absorbed at 0, we have that $X_t = -d$ if and only if $Y_t = d$, so the expected number of visits to $-d$ before absorption is also equal to 2.

$\square$

Lemma 5.12. *Let $L := \ln(2/\varepsilon)$ and let K denote the total number of negative probes up to time T under Shivgorithm, $K = N_T^-$. Then, for any time horizon $T \in \mathbb{N}$ and any $\varepsilon \in (0, 1/2]$, there exists an absolute constant $C' > 0$ such that*
$$\mathbb{E}[K] \ \leq \ C' \, L^{1/3} T^{5/6}.$$

Proof. Applying earlier techniques, we have
$$
\begin{aligned}
\mathbb{E}[K] &= \mathbb{E}\left[\sum_{i=1}^{I} m_i\right] \\
&= \sum_{i=1}^{\infty} \mathbb{E}\left[m_i \mathbb{I}\{i \leq I\}\right] \\
&= \sum_{i=1}^{\infty} \mathbb{E}\left[m_i \mathbb{I}\{s_i \leq T\}\right] \\
&= \sum_{i=1}^{\infty} \mathbb{E}\left[\mathbb{I}\{s_i \leq T\} \, \mathbb{E}[m_i \mid s_i]\right] \\
&\leq \sum_{i=1}^{\infty} \mathbb{E}\left[\mathbb{I}\{s_i \leq T\} \, C(L\sqrt{T})^{1/3}\right] \\
&= C(L\sqrt{T})^{1/3} \sum_{i=1}^{\infty} \mathbb{P}(s_i \leq T) \\
&= C(L\sqrt{T})^{1/3} \, \mathbb{E}\left[\sum_{i=1}^{\infty} \mathbf{1}\{s_i \leq T\}\right] \\
&= C(L\sqrt{T})^{1/3} \, \mathbb{E}[I] \\
&\leq C'(L\sqrt{T})^{1/3} \sqrt{T} \\
&= C' \, L^{1/3} T^{5/6}.
\end{aligned}
$$

$\square$