Perturbed-History Exploration in Stochastic Multi-Armed Bandits

Branislav Kveton$^1$, Csaba Szepesvári$^2$, Mohammad Ghavamzadeh$^2$, Craig Boutilier$^1$

$^1$Google Research
$^2$DeepMind
$^3$Facebook AI Research

Abstract

We propose an online algorithm for cumulative regret minimization in a stochastic multi-armed bandit. The algorithm adds $O(t)$ i.i.d. pseudo-rewards to its history in round $t$ and then pulls the arm with the highest average reward in its perturbed history. Therefore, we call it perturbed-history exploration (PHE). The pseudo-rewards are designed to offset the underestimated mean rewards of arms in round $t$ with a high probability. We analyze PHE in a $K$-armed bandit and derive both $O(K\Delta^{-1}\log n)$ and $O(\sqrt{Kn}\log n)$ bounds on its $n$-round regret, where $\Delta$ denotes the minimum gap between the mean rewards of the optimal and suboptimal arms. The key to our analysis is a novel argument that shows that randomized Bernoulli rewards lead to optimism. We compare PHE empirically to several baselines and show that it is competitive with the best of them.

1 Introduction

A multi-armed bandit [Lai and Robbins, 1985; Auer et al., 2002; Lattimore and Szepesvari, 2019] is an online learning problem where actions of the learning agent are represented by arms. After the arm is pulled, the agent receives its stochastic reward. The objective of the agent is to maximize its expected cumulative reward. The agent does not know the mean rewards of the arms in advance and faces the so-called exploration-exploitation dilemma: explore, and learn more about the arm; or exploit, and pull the arm with the highest average reward thus far. The arm may be a treatment in a clinical trial and its reward is the outcome of that treatment on some patient population.

Thompson sampling (TS) [Thompson, 1933; Russo et al., 2018] and optimism in the face of uncertainty (OFUL) [Auer et al., 2002; Dani et al., 2008; Abbasi-Yadkori et al., 2011] are the most celebrated and studied exploration strategies in stochastic multi-armed bandits. These strategies are near optimal in multi-armed [Garivier and Cappe, 2011; Agrawal and Goyal, 2013a] and linear [Abbasi-Yadkori et al., 2011; Agrawal and Goyal, 2013b] bandits. However, they typically do not generalize easily to complex problems. For instance, in generalized linear bandits [Filippi et al., 2010], we only know how to construct approximate confidence sets and posterior distributions, which affects the statistical efficiency of bandit algorithms [Filippi et al., 2010; Zhang et al., 2016; Abeille and Lazaric, 2017; Jun et al., 2017; Li et al., 2017].

Another example is online learning to rank [Radlinski et al., 2008], where we only have statistically efficient algorithms for simple user interaction models, such as the cascade model [Kveton et al., 2015; Katariya et al., 2016]. If the user interaction model was a graphical model with latent variables [Chapelle and Zhang, 2009], we would not know how to design a bandit algorithm with regret guarantees. In general, it is hard to design efficient approximations to confidence sets and posterior distributions [Gopalan et al., 2014; Kawale et al., 2015; Lu and Van Roy, 2017; Riquelme et al., 2018; Lipton et al., 2018; Liu et al., 2018].

In this work, we propose a novel exploration strategy that is conceptually straightforward and has the potential to easily generalize to complex problems. In round $t$, we add $O(t)$ i.i.d. pseudo-rewards to the history of the learning agent, which treats them as if they were generated by actual arm pulls. The agent pulls the arm with the highest average reward in this perturbed history and then updates its history with the observed reward. The pseudo-rewards are drawn from the same family of distributions as the actual rewards, but generate maximum variance randomized data.

Our algorithm, perturbed-history exploration (PHE), is inherently optimistic. To see this, note that the lack of “optimism” regarding arm $i$ in round $t$, that the estimated mean reward of arm $i$ is below the mean, is due to a specific history of the past $t – 1$ rewards, some of which are low simply because of randomness. Note that these rewards are independent noisy realizations of the mean arm reward. Therefore, if we add $O(t)$ i.i.d. pseudo-rewards to the history, they can offset the lack of optimism due to a specific reward realization. This is the key idea in our design. An appealing aspect of the design is its conceptual simplicity, because maximum variance rewards can be easily added to complex problems.

We make the following contributions in this paper. First, we propose PHE, a multi-armed bandit algorithm where the mean rewards of arms are estimated using a mixture of actual rewards and i.i.d. pseudo-rewards. Second, we analyze PHE in a $K$-armed bandit with $[0, 1]$ rewards, and prove both $O(K\Delta^{-1}\log n)$ and $O(\sqrt{Kn}\log n)$ bounds on its $n$-round regret, where $\Delta$ is the minimum gap between the mean re-
wards of the optimal and suboptimal arms. The key to our analysis is a novel argument that shows that randomized Bernoulli rewards lead to optimism. Finally, we empirically compare PHE to several baselines and show that it is competitive with Thompson sampling.

2 Setting

We use the following notation. The set \( \{1, \ldots, n\} \) is denoted by \([n]\). We define \( \text{Ber}(x; p) = p^x(1-p)^{1-x} \) and let \( \text{Ber}(p) \) be the corresponding Bernoulli distribution. We also define \( B(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x} \) and let \( B(n, p) \) be the corresponding binomial distribution. For any event \( E \), \( \mathbb{1}\{E\} = 1 \) if and only if event \( E \) occurs, and is zero otherwise.

We study the problem of cumulative regret minimization in a stochastic multi-armed bandit. Formally, a stochastic multi-armed bandit \([Lai and Robbins, 1985; Auer et al., 2002; Lattimore and Szepesvari, 2019]\) is an online learning problem where the learning agent sequentially pulls \( K \) arms in \( n \) rounds. In round \( t \in [n] \), the agent pulls arm \( i_t \in [K] \) and receives its reward. The reward of arm \( i \in [K] \) in round \( t \), \( Y_{i,t} \), is drawn i.i.d. from a distribution of arm \( i \), \( P_i \), with mean \( \mu_i \) and support \([0,1]\). The goal of the agent is to maximize its expected cumulative reward in \( n \) rounds. The agent does not know the mean rewards of the arms in advance and learns them by pulling the arms.

Without loss of generality, we assume that the first arm is \emph{optimal}, that is \( \mu_1 > \max_{i>1} \mu_i \). Let \( \Delta_i = \mu_1 - \mu_i \) denote the \emph{gap} of arm \( i \). Maximization of the expected cumulative reward in \( n \) rounds is equivalent to minimizing the expected \( n \)-round regret, which we define as

\[
R(n) = \sum_{i=2}^K \Delta_i \mathbb{E} \left[ \sum_{t=1}^n \mathbb{1}\{I_t = i\} \right].
\]

3 Perturbed-History Exploration

Our new algorithm, \emph{perturbed-history exploration} (PHE), is presented in Algorithm 1. PHE pulls the arm with the highest average reward in its perturbed history, which is estimated as follows. Let \( T_{i,t} = \sum_{s=1}^t \mathbb{1}\{I_s = i\} \) denote the number of pulls of arm \( i \) in the first \( t \) rounds and \( s = T_{i,t-1} \). Then the estimated reward of arm \( i \) in round \( t \), \( \hat{\mu}_{i,t} \), is the average of its past \( s \) rewards and \( s \) i.i.d. pseudo-rewards \( \{Z_{i,t}\}_{t=1}^s \), for some tunable integer \( s > 0 \). In line 9, \( \hat{\mu}_{i,t} \) is computed from the sum of the rewards of arm \( i \) after \( s \) pulls, \( V_{i,s} \), and the sum of its pseudo-rewards, \( U_{i,s} \). After the arm is pulled, the cumulative reward of that arm is updated with its reward in round \( t \) (line 17). All arms are initially pulled once (line 11).

\[
\hat{\mu}_i = \frac{V_{i,s} + U_{i,s}}{(a+1)s}.
\]

This estimator has two key properties that allow us to bound the regret of PHE in Section 4. First, it \emph{concentrates} at the scaled and shifted mean reward of arm \( i \). More specifically, let \( U_{i,s} = \mathbb{E} [U_{i,s}] \) and \( V_{i,s} = \mathbb{E} [V_{i,s}] \). Then we have

\[
\mathbb{E} \left[ \hat{\mu} \right] = \frac{\hat{V}_{i,s} + \hat{U}_{i,s}}{a(a+1)s} = \mu_i + \frac{a}{a+1},
\]

\[
\text{var} [\hat{\mu}] \leq \frac{\sigma_{\text{max}}^2}{(a+1)s},
\]

where \( \sigma_{\text{max}}^2 \) is the maximum variance of any random variable on \([0,1]\). By Popoviciu’s inequality on variances \([Popoviciu, 1935]\), we have \( \sigma_{\text{max}}^2 = 1/4 \), which is precisely the variance of \( Z \sim \text{Ber}(1/2) \).

Second, \( \hat{\mu} \) is sufficiently \emph{optimistic} in the following sense. Let \( E = \{ V_{i,s}/s - V_{i,s}/s = \varepsilon \} \) be the event that the estimated mean reward of arm \( i \) is below the mean by \( \varepsilon > 0 \). We informally below that any \( \varepsilon > 0 \) suffices for sublinear regret. We prove in Section 4 that any \( \varepsilon > 2 \) guarantees it.

Now we examine how exploration emerges within our algorithm. Fix arm \( i \) and the number of its pulls \( s \). Let \( V_{i,s} \) be the cumulative reward of arm \( i \) after \( s \) pulls. Let \( \{Z_{i,s}\}_{s=1}^s \) \( \sim \text{Ber}(1/2) \) be as i.i.d. pseudo-rewards and \( U_{i,s} = \sum_{t=1}^s Z_{i,t} \) be their sum. The mean reward of arm \( i \) in round \( t \) (line 9) is estimated as

\[
\hat{\mu}_i = \frac{V_{i,s} + U_{i,s}}{(a+1)s},
\]

Algorithm 1 Perturbed-history exploration in a multi-armed bandit with \([0,1]\) rewards.

1: \textbf{Inputs:} Perturbation scale \( a \)

2: \textbf{for} \( i = 1, \ldots, K \) \textbf{do} \hfill \textit{\textbf{Initialization}}

3: \hspace{1em} \( T_{i,0} \leftarrow 0, \ V_{i,0} \leftarrow 0 \)

4: \textbf{for} \( t = 1, \ldots, n \) \textbf{do}

5: \hspace{1em} \textbf{for} \( i = 1, \ldots, K \) \textbf{do} \hfill \textit{\textbf{Estimate mean arm rewards}}

6: \hspace{2em} \textbf{if} \( T_{i,t-1} > 0 \) \textbf{then}

7: \hspace{3em} \( s \leftarrow T_{i,t-1} \)

8: \hspace{4em} \( U_{i,s} \leftarrow \sum_{t=1}^s Z_{i,t}, \text{where} \ (Z_{i,t})_{t=1}^s \sim \text{Ber}(1/2) \)

9: \hspace{4em} \( \hat{\mu}_{i,t} \leftarrow \frac{V_{i,s} + U_{i,s}}{(a+1)s} \)

10: \hspace{1em} \textbf{else}

11: \hspace{2em} \( \hat{\mu}_{i,t} \leftarrow +\infty \)

12: \hspace{1em} \( I_t \leftarrow \text{arg max}_{i \in [K]} \hat{\mu}_{i,t} \) \hfill \textit{\textbf{Pulled arm}}

13: \hspace{1em} \( \text{Pull arm} \ I_t \) and get reward \( Y_{i,t} \)

14: \hspace{1em} \textbf{for} \( i = 1, \ldots, K \) \textbf{do} \hfill \textit{\textbf{Update statistics}}

15: \hspace{2em} \textbf{if} \( i = I_t \) \textbf{then}

16: \hspace{3em} \( T_{i,t} \leftarrow T_{i,t-1} + 1 \)

17: \hspace{3em} \( V_{i,T_{i,t}} \leftarrow V_{i,T_{i,t}-1} + Y_{i,t} \)

18: \hspace{2em} \textbf{else}

19: \hspace{3em} \( T_{i,t} \leftarrow T_{i,t-1} \)


say that \( \hat{\mu} \) is optimistic if
\[
\mathbb{P} \left( \frac{V_{i,s} + U_{i,s}}{(a + 1)s} \geq \frac{V_{i,s} + \bar{U}_{i,s}}{(a + 1)s} \ \bigg| E \right) > \mathbb{P} (E) \tag{5}
\]
for any \( \varepsilon > 0 \) such that \( \mathbb{P} (E) > 0 \). That is, the probability of overestimating the mean reward of the arm conditioned on any harmful deviation \( \varepsilon > 0 \) is at least high as the probability of that deviation. Under this condition, PHE explores enough and can escape potentially harmful deviations.

Now we argue informally that (5) holds for \( a > 1 \) in PHE. Fix any \( \varepsilon > 0 \). First, note that
\[
\mathbb{P} (E) \leq \mathbb{P} \left( \frac{V_{i,s}}{s} - \frac{V_{i,s}}{s} \geq \varepsilon \right)
\]
and
\[
\mathbb{P} \left( \frac{V_{i,s} + U_{i,s}}{(a + 1)s} \geq \frac{V_{i,s} + \bar{U}_{i,s}}{(a + 1)s} \right.
\]
\[
= \mathbb{P} \left( \frac{V_{i,s} + U_{i,s}}{(a + 1)s} \geq \frac{V_{i,s} + \bar{U}_{i,s}}{(a + 1)s} \right)
\]
\[
= \mathbb{P} \left( \frac{U_{i,s}}{s} - \frac{\bar{U}_{i,s}}{s} \geq \varepsilon \right). \tag{6}
\]
The last equality holds because \( U_{i,s} - \bar{U}_{i,s} \) is independent of past rewards. Based on the above two inequalities, it follows that (5) holds when
\[
\mathbb{P} \left( \frac{U_{i,s}}{s} - \frac{\bar{U}_{i,s}}{s} \geq \varepsilon \right) > \mathbb{P} \left( \frac{V_{i,s}}{s} - \frac{V_{i,s}}{s} \geq \varepsilon \right). \tag{6}
\]
Now suppose that both \( V_{i,s}/s \) and \( U_{i,s}/s \) were normally distributed. Then (6) would hold if the variance of \( V_{i,s}/s \) was lower than that of \( U_{i,s}/s \). This is indeed true, since
\[
\text{var} \left[ \frac{V_{i,s}}{s} \right] \leq \sigma_{\text{max}}^2/s, \quad \text{var} \left[ \frac{U_{i,s}}{s} \right] = a\sigma_{\text{max}}^2/s;
\]
and \( a > 1 \) from our assumption. This concludes our informal argument. We experiment with this suggested value of \( a \) in Section 5.

4 Analysis

PHE is an instance of general randomized exploration in Section 3 of Kveton et al. [2019b]. So, the regret of PHE can be bounded using their Theorem 1, which we restate below.

**Theorem 1.** For any \((\tau_i)_{i=2}^K \in \mathbb{R}^{K-1}\), the expected \( n \)-round regret of Algorithm 1 in Kveton et al. [2019b] can be bounded from above as
\[
R(n) \leq \sum_{i=2}^K \Delta_i (a_i + b_i),
\]
where
\[
a_i = \sum_{s=0}^{n-1} \mathbb{E} \left[ \min \left\{ 1/Q_{1,s}(\tau_i) - 1, n \right\} \right],
\]
\[
b_j = \sum_{s=0}^{n-1} \mathbb{P} (Q_{1,s}(\tau_i) > 1/n) + 1.
\]
For any arm \( i \) and the number of its pulls \( s \in [n] \cup \{0\} \),
\[
Q_{i,s}(\tau) = \mathbb{P} (\hat{\mu} \geq \tau | \hat{\mu} \sim p(\mathcal{H}_{i,s}, \mathcal{H}_{i,s})) \tag{7}
\]
is the tail probability that the estimated mean reward of arm \( i \), \( \hat{\mu} \), is at least \( \tau \) conditioned on the history of the arm after \( s \) pulls, \( \mathcal{H}_{i,s} \); where \( p \) is the sampling distribution of \( \hat{\mu} \) and \( \tau \) is a tunable parameter. In PHE, the history \( \mathcal{H}_{i,s} \) is \( V_{i,s} \) and \( \hat{\mu} \) is defined in (2). Following Kveton et al. [2019b], we choose \( \tau_i \) to be the average of the scaled and shifted mean rewards of arms 1 and \( i \),
\[
\tau_i = \frac{\mu_i + a/2}{a + 1} + \frac{\Delta_i}{2(a + 1)}, \tag{8}
\]
which are defined in (3). This setting leads to the following gap-dependent regret bound.

**Theorem 2.** For any \( a > 2 \), the expected \( n \)-round regret of PHE is bounded as
\[
R(n) \leq \sum_{i=2}^K \Delta_i \left( \frac{16ac}{\Delta_i^2} \log n + 3 + \frac{8a}{\Delta_i^2} \log n + 2 \right), \tag{9}
\]
where
\[
c = \frac{e^2 \sqrt{2a}}{\sqrt{\pi}} \exp \left[ \frac{16}{a - 2} \right] \left( 1 + \sqrt{\frac{\pi a}{8(a - 2)}} \right). \tag{10}
\]

**Proof.** The proof has two parts. In Section 4.2, we prove an upper bound on \( b_i \) in Theorem 1. In Section 4.3, we prove an upper bound on \( a_i \) in Theorem 1. Finally, we add these upper bounds for all arms \( i > 0 \).

A standard reduction yields a gap-free regret bound.

**Theorem 3.** For any \( a > 2 \), the expected \( n \)-round regret of PHE is bounded as
\[
R(n) \leq 4 \sqrt{2a(2c + 1)K n \log n + 5K},
\]
where \( c \) is defined in Theorem 2.

**Proof.** Let \( \mathcal{K} = \{ i \in [K] : \Delta_i \geq \varepsilon \} \) be the set of arms whose gaps are at least \( \varepsilon > 0 \). Then by the same argument as in the proof of Theorem 2 and from the definition of \( \mathcal{K} \), we have
\[
R(n) \leq \sum_{i \in \mathcal{K}} \frac{8a(2c + 1) \log n + \varepsilon n + 5 |\mathcal{K}|}{\varepsilon} \leq \frac{8a(2c + 1)K n \log n + \varepsilon n + 5K}{\varepsilon}.
\]
Now we choose \( \varepsilon = \sqrt{\frac{8a(2c + 1)K n \log n}{n}} \), which completes the proof.

4.1 Discussion

We derive two regret bounds. The gap-dependent bound in Theorem 2 is \( O(\sqrt{K \Delta^{-1} \log n}) \), where \( \Delta = \min_{i > 1} \Delta_i \) is the minimum gap, \( K \) is the number of arms, and \( n \) is the number of rounds. This scaling is considered near-optimal in stochastic multi-armed bandits. The gap-free bound in Theorem 3 is \( O(\sqrt{K n \log n}) \). This scaling is again near-optimal, up to the factor of \( \sqrt{\log n} \), in stochastic multi-armed bandits.

A potentially large factor in our bounds is \( \exp[16/(a - 2)] \) in (10). It arises in the lower bound on the probability of a binomial tail (Appendix A) and is likely to be loose. Nevertheless, it is constant in \( K, \Delta, \) and \( n \); and decreases significantly even for small \( a \). For instance, it is only \( e^4 \) at \( a = 6 \).
4.2 Upper Bound on $b_i$ in Theorem 1

Fix arm $i > 1$. Based on our choices of $\mathcal{H}_{1,s}$, $\tilde{\mu}$, and $\tau_i$, we have for $s > 0$ that
\[
Q_{1,s}(\tau_i) = \mathbb{P}\left( \frac{V_{1,s} + U_{1,s}}{(a+1)s} \geq \frac{\mu_i + a/2 + \Delta_i/2}{a+1} \mid V_{1,s} \right).
\]
We set $Q_{1,0}(\tau_i) = 1$, because of the optimistic initialization in line 11 of PHE. We abbreviate $Q_{1,s}(\tau_i)$ as $Q_{1,s}$.

Fix the number of pulls $s$ and let $m = 8a\Delta_i^2 \log n$. If $s \leq m$, we bound $\mathbb{P}(Q_{1,s} > 1/n)$ trivially by 1. If $s > m$, we split our proof based on the event that $V_{1,s}$ is not much larger than its expectation,
\[
E = \{ V_{1,s} - \mu_i s \leq \Delta_i s/4 \}.
\]
On event $E$,
\[
Q_{1,s} = \mathbb{P}\left( V_{1,s} + U_{1,s} - \mu_i s - \frac{as}{2} \geq \frac{\Delta_i s^2}{2} \mid V_{1,s} \right)
\leq \mathbb{P}\left( U_{1,s} - \frac{as}{2} \geq \frac{\Delta_i s^2}{4} \mid V_{1,s} \right)
\leq \exp\left[ -\frac{\Delta_i^2 s^2}{8a} \right] \leq n^{-1},
\]
where the first inequality is by the definition of event $E$, the second is by Hoeffding’s inequality, and the last is from $s > m$. On the other hand, event $\tilde{E}$ is unlikely because
\[
\mathbb{P}(\tilde{E}) \leq \exp\left[ -\frac{\Delta_i^2 s^2}{8} \right] \leq \exp\left[ -\frac{\Delta_i^2 s^2}{8a} \right] \leq n^{-1},
\]
where the first inequality is from Hoeffding’s inequality, the second is from $a > 1$, and the last is from $s > m$. Now we apply the last two inequalities and get
\[
P\left( Q_{1,s} > 1/n \right) = \mathbb{E}\left[ \mathbb{P}\left( Q_{1,s} > 1/n \mid V_{1,s} \right) \mathbb{I}\{E\} \right] + \mathbb{E}\left[ \mathbb{P}\left( Q_{1,s} > 1/n \mid V_{1,s} \right) \mathbb{I}\{\tilde{E}\} \right]
\leq 0 + \mathbb{P}\{\tilde{E}\} \leq n^{-1}.
\]
Finally, we chain our upper bounds for all $s \in [n]$ and get
\[
b_i \leq 1 + \sum_{s = 0}^{[m]} 1 + \sum_{s = [m]+1}^{n} n^{-1} \leq \frac{8a}{\Delta_i^2} \log n + 3.
\]
This completes our proof.

4.3 Upper Bound on $a_i$ in Theorem 1

Fix arm $i > 1$. Based on our choices of $\mathcal{H}_{1,s}$, $\tilde{\mu}$, and $\tau_i$, we have for $s > 0$ that
\[
Q_{1,s}(\tau_i) = \mathbb{P}\left( \frac{V_{1,s} + U_{1,s}}{(a+1)s} \geq \frac{\mu_i + a/2 + \Delta_i/2}{a+1} \mid V_{1,s} \right).
\]
We set $Q_{1,0}(\tau_i) = 1$, because of the optimistic initialization in line 11 of PHE. We abbreviate $Q_{1,s}(\tau_i)$ as $Q_{1,s}$, and define
\[
F_s = 1/Q_{1,s} - 1.
\]
Fix the number of pulls $s$ and let $m = 16a\Delta_i^2 \log n$. If $s = 0$, $Q_{1,s} = 1$ and we obtain $\mathbb{E}\left[ \min\{F_s, n\} \right] = 0$. Now consider the case of $s > 0$. If $s \leq m$, we apply the upper bound in Theorem 4 in Appendix A and get
\[
\begin{align*}
\mathbb{E}\left[ \min\{F_s, n\} \right] &\leq \mathbb{E}\left[ 1/Q_{1,s} \right] \\
&\leq \mathbb{E}\left[ 1/P\left( V_{1,s} + U_{1,s} \geq \mu_i s + as/2 \mid V_{1,s} \right) \right] \leq c,
\end{align*}
\]
where $c$ is defined in (10). Note that $2a$ in Theorem 4 plays the role of $a$ in this claim.

If $s > m$, we split our argument based on the event that $V_{1,s}$ is not much smaller than its expectation,
\[
E = \{ \mu_i s - V_{1,s} \leq \Delta_i s/4 \}.
\]
On event $E$,
\[
Q_{1,s} = \mathbb{P}\left( \mu_i s + \frac{as}{2} - V_{1,s} - U_{1,s} \leq \frac{\Delta_i s^2}{2} \mid V_{1,s} \right)
\geq \mathbb{P}\left( \frac{as}{2} - U_{1,s} \leq \frac{\Delta_i s^2}{4} \mid V_{1,s} \right)
= 1 - \mathbb{P}\left( \frac{as}{2} - U_{1,s} > \frac{\Delta_i s^2}{4} \mid V_{1,s} \right)
\geq 1 - \exp\left[ -\frac{\Delta_i^2 s^2}{8a} \right] \geq \frac{n^2 - 1}{n^2},
\]
where the first inequality is by the definition of event $E$, the second is by Hoeffding’s inequality, and the last is from $s > m$. This lower bound yields
\[
F_s = \frac{1}{Q_{1,s}} - 1 \leq \frac{n^2}{n^2 - 1} - 1 = \frac{1}{n^2 - 1} \leq n^{-1}
\]
for $n \geq 2$. On the other hand, event $\tilde{E}$ is unlikely because
\[
\mathbb{P}\{\tilde{E}\} \leq \exp\left[ -\frac{\Delta_i^2 s^2}{8} \right] \leq \exp\left[ -\frac{\Delta_i^2 s^2}{8a} \right] \leq n^{-2},
\]
where the first inequality is from Hoeffding’s inequality, the second is from $a > 1$, and the last is from $s > m$. Now we apply the last two inequalities and get
\[
\begin{align*}
\mathbb{E}\left[ \min\{F_s, n\} \right] &= \mathbb{E}\left[ \mathbb{E}\left[ \min\{F_s, n\} \mid V_{1,s} \right] \mathbb{I}\{E\} \right] + \mathbb{E}\left[ \mathbb{E}\left[ \min\{F_s, n\} \mid V_{1,s} \right] \mathbb{I}\{\tilde{E}\} \right] \\
&\leq n^{-1} \mathbb{P}\{E\} + n \mathbb{P}\{\tilde{E}\} \leq 2n^{-1}.
\end{align*}
\]
Finally, we chain our upper bounds for all $s \in [n]$ and get
\[
a_i \leq \sum_{s = 1}^{[m]} c + \sum_{s = [m]+1}^{n} 2n^{-1} \leq \frac{16ac}{\Delta_i^2} \log n + 2.
\]
This completes our proof.

5 Experiments

We compare PHE to five baselines: UCB1 [Auer et al., 2002], KL-UCB [Garivier and Cappe, 2011], Bernoulli TS [Agrawal and Goyal, 2013a], Giro [Kveton et al., 2019b], and FPL [Hannan, 1957; Kalai and Vempala, 2005]. The prior in TS is Beta(1, 1). We choose our baselines for the following reasons. KL-UCB and TS are statistically near-optimal in Bernoulli bandits. We implement them with $[0, 1]$ rewards as described in Agrawal and Goyal [2013a]. Specifically, for
any observed reward \( Y_{i,t} \in [0, 1] \), we draw \( \hat{Y}_{i,t} \sim \text{Ber}(Y_{i,t}) \) and then use it instead of \( Y_{i,t} \). 
Giro is chosen because it explores by adding pseudo-rewards to its history, similarly to PHE (Section 6). We implement it with \( a = 1 \), as analyzed in Kveton et al. [2019b]. FPL is chosen because it perturbs the estimates of mean rewards similarly to PHE (Section 6). We implement it with geometric resampling and exponential noise [Neu and Bartok, 2013].

We experiment with three values of the perturbation scale \( a \) in PHE: 2.1, 1.1, and 0.5. The first value is greater than 2 and it is justified by Theorem 2. The second value is greater than 1 and it is informally justified in Section 3. The last value is used to show that the regret of PHE can grow linearly when PHE is not parameterized properly.

To run PHE with a non-integer perturbation scale \( a \), we replace \( a \) in PHE with \( \lceil a \rceil \). The analysis of PHE in Section 4 can be extended to this setting. We also experimented with \( a = 1 \) and \( a = 2 \). We do not report these results because they are similar to those at \( a = 1.1 \) and \( a = 2.1 \).

5.1 Comparison to Baselines

In the first experiment, we evaluate PHE on two classes of the bandit problems in Kveton et al. [2019b]. The first class is a Bernoulli bandit where \( P_i = \text{Ber}(\mu_i) \). The second class is a beta bandit where \( P_i = \text{Beta}(\mu_i, v(1 - \mu_i)) \) and \( v = 4 \). We experiment with 100 randomly chosen problems in each class. Each problem has \( K = 10 \) arms and their means are chosen randomly from interval \([0.25, 0.75]\). The horizon is \( n = 10000 \) rounds.

Our results are reported in Figures 1a and 1b. We observe that PHE with \( a > 1 \) outperforms four of our baselines: UCB1, KL-UCB, Giro, and FPL. This is unexpected, since the design of PHE is conceptually simple; and neither requires nor uses confidence intervals or posteriors. PHE becomes competitive with TS at \( a = 1.1 \). We also note that the regret of PHE is linear in the Bernoulli bandit at \( a = 0.5 \). This shows that our informal and formal recommendations for setting the perturbation scale \( a \) are reasonably tight.

5.2 Computational Cost

In the second experiment, we compare the run times of three randomized algorithms: TS, which samples from a beta posterior; Giro, which bootstraps from a history with pseudo-rewards; and PHE, which samples pseudo-rewards from a binomial distribution. The number of arms is between 5 and 20, and the horizon is up to \( n = 10000 \) rounds.

Our results are reported in Figure 1c. In all settings, the run time of PHE is comparable to that of TS. The run time of Giro is an order of magnitude higher. The reason is that the computational cost of bootstrapping grows linearly with the number of past observations.

6 Related Work

Our algorithm design bears a similarity to three existing designs, which we discuss in detail below.

Giro is a bandit algorithm where the mean reward of the arm is estimated by its average reward in a bootstrap sample of its history with pseudo-rewards [Kveton et al., 2019b]. The algorithm has a provably sublinear regret in a Bernoulli bandit. PHE improves over Giro in three respects. First, its design is simpler, because PHE merely adds random pseudo-rewards and does not bootstrap. Second, PHE has a sublinear regret in any \( K \)-armed bandit with \([0, 1]\) rewards. Third, PHE is computationally efficient beyond a Bernoulli bandit. We discuss this in Section 3.

Our work is also closely related to posterior sampling. In particular, let \( \mu \sim \mathcal{N}(\mu_0, \sigma^2) \) and \( (Y_{i,t})_{t=1}^s \sim \mathcal{N}(\mu, \sigma^2) \) be s i.i.d. noisy observations of \( \mu \). Then the posterior distribution of \( \mu \) conditioned on \( (Y_{i,t})_{t=1}^s \) is

\[
\mathcal{N}\left( \mu_0 + \frac{\sum_{t=1}^s Y_{i,t}}{s+1}, \frac{\sigma^2}{s+1} \right).
\]  

A sample from this distribution can be also drawn as follows. First, draw \( s + 1 \) i.i.d. samples \( (Z_t)_{t=0}^s \sim \mathcal{N}(0, \sigma^2) \). Then

\[
\mu_0 + \frac{\sum_{t=1}^s Y_{i,t}}{s+1} + \frac{\sum_{t=0}^s Z_t}{s+1}
\]

is a sample from (11). Unfortunately, the above equivalence holds only for normal random variables. Therefore, it cannot justify PHE as a form of Thompson sampling. Nevertheless, the scale of the perturbation is similar to (2), which suggests that PHE is sound.

Follow the perturbed leader (FPL) [Hannan, 1957; Kalai and Vempala, 2005] is an algorithm design where the learning agent pulls the arm with the lowest perturbed cumulative
cost. In our notation, \( I_t = \arg \min_{i \in [K]} \tilde{V}_{i,t-1} + \tilde{U}_{i,t} \), where \( \tilde{V}_{i,t-1} \) is the cumulative cost of arm \( i \) in the first \( t - 1 \) rounds and \( \tilde{U}_{i,t} \) is the perturbation of arm \( i \) in round \( t \). PHE differs from FPL in three respects. First, \( \tilde{U}_{i,t} = O(\sqrt{n}) \) in FPL. In PHE, the noise in round \( t \) is \( U_{i,T_{i,t-1}} = O(T_{i,t-1}) \), and thus it adapts to the number of pulls. Second, FPL has been traditionally studied in the non-stochastic full-information setting. PHE is designed for the stochastic bandit setting. Neu and Bartok [2013] extended FPL to the bandit setting using geometric resampling and we compare to their algorithm in Section 5. Finally, all existing FPL analyses derive gap-free regret bounds. We derive a gap-dependent regret bound.

7 Conclusions

We propose a new online algorithm, PHE, for cumulative regret minimization in stochastic multi-armed bandits. The key idea in PHE is to add \( O(t) \) i.i.d. pseudo-rewards to the history in round \( t \) and then pull the arm with the highest average reward in this perturbed history. The pseudo-rewards are drawn from the maximum variance distribution. We derive \( O(K\Delta^{-1} \log n) \) and \( O(\sqrt{k \log n}) \) bounds on the \( n \)-round regret of PHE, where \( K \) is the number of arms and \( \Delta \) is the minimum gap between the mean rewards of the optimal and suboptimal arms. This result is unexpected, given how simple the design of PHE is. We empirically compare PHE to several baselines and show that it is competitive with Bernoulli Thompson sampling.

PHE can be applied to reward distributions with a bounded support. Specifically, if \( Y_{i,t} \in [m,M] \), \( Y_{i,t} \) in line 17 of PHE should be replaced with \( (Y_{i,t} - m)/(M - m) \).

PHE can be applied to structured problems, such as generalized linear bandits [Filippi et al., 2010], as follows. Let \( x_i \) be the feature vector of arm \( i \). Then \((x_{i,t}, Y_{i,t})\) for \( t \leq T_{i,t-1} \) is the history in round \( t \) and a natural choice for the pseudo-history is \((x_{i,t}, Z_{i,t})\) for \( t \), where \( Z_{i,t} \sim \text{Ber}(1/2) \) are i.i.d. random variables. In round \( t \), a reward generalization model is trained on both histories and the learning agent pulls the arm with the highest estimated mean reward in that model. We leave the analysis of this algorithm for future work. It is analyzed in a linear bandit in Kveton et al. [2019a].

We believe that PHE can be extended to other perturbation schemes. For instance, since \( \text{var}[V_{i,t}] \leq s/4 \), it is plausible that any \( s \) i.i.d. pseudo-rewards with comparable variances, such as \((Z_{i,t})_{t=1}^{\infty} \sim \mathcal{N}(0,1/4)\), would lead to optimism. We leave the analyses of such designs for future work.

A Technical Lemmas

Fix arm \( i \) and the number of its pulls \( n \). Let \( X \) be the cumulative reward of arm \( i \) after \( n \) pulls and \( Y = \sum_{t=1}^{2an} Z_t \) be the sum of \( 2an \) i.i.d. pseudo-rewards \((Z_{i,t})_{t=1}^{2an} \sim \text{Ber}(1/2)\). Note that both \( X \) and \( Y \) are random variables. Let \( \bar{X} = \mathbb{E}[X] \) and \( \bar{Y} = \mathbb{E}[Y] \). Our main theorem is stated and proved below.

Theorem 4. For any \( a > 1 \),
\[
\mathbb{E}\left[\frac{1}{\mathbb{P}}(X + Y \geq \bar{X} + \bar{Y} \mid X)\right] \leq 2e^2\sqrt{\frac{a}{\pi}} \exp\left[\frac{8}{a - 1}\right] \left(1 + \sqrt{\frac{\pi a}{2(a - 1)}}\right).
\]

Proof. Let \( W = \mathbb{E}\left[\frac{1}{\mathbb{P}}(X + Y \geq \bar{X} + \bar{Y} \mid X)\right] \). Note that \( W \) can be rewritten as
\[
f(X) = \left[\sum_{y = \lfloor \bar{X} - X + an \rfloor}^{m} B(y; m, 1/2)\right]^{-1},
\]
and \( m = 2an \). This follows from the definition of \( Y \) and that \( \bar{Y} = an \).

Note that \( f(X) \) decreases in \( X \), as required by Lemma 1, because the probability of observing at least \( \lfloor \bar{X} - X + an \rfloor \) ones increases with \( X \). So we can apply Lemma 1 and get
\[
W \leq \sum_{i=0}^{i_0-1} \exp[-2i^2]\left[\sum_{y = \lfloor an + (i+1)\sqrt{n} \rfloor}^{m} B(y; m, 1/2)\right]^{-1} + \exp[-2i_0^2]\left[\sum_{y = \lfloor an + \bar{X} \rfloor}^{m} B(y; m, 1/2)\right]^{-1},
\]
where \( i_0 \) is the smallest integer such that \((i_0 + 1)\sqrt{n} \geq \bar{X} \), as defined in Lemma 1.

Now we bound the sums in the reciprocals from below using Lemma 2. For \( \delta = (i + 1)\sqrt{n} \),
\[
\sum_{y = \lfloor an + (i+1)\sqrt{n} \rfloor}^{m} B(y; m, 1/2) \geq \frac{\sqrt{\pi}}{e^2\sqrt{\pi}} \exp\left[-\frac{2(i + 2)^2}{a}\right].
\]

For \( \delta = \bar{X} \),
\[
\sum_{y = \lfloor an + \bar{X} \rfloor}^{m} B(y; m, 1/2) \geq \frac{\sqrt{\pi}}{e^2\sqrt{\pi}} \exp\left[-\frac{2(\bar{X} + \sqrt{n})^2}{an}\right] \geq \frac{\sqrt{\pi}}{e^2\sqrt{\pi}} \exp\left[-\frac{2(i_0 + 2)^2}{a}\right],
\]
where the last inequality is from the definition of \( i_0 \). Then we chain the above three inequalities and get
\[
W \leq \frac{e^2\sqrt{\pi}}{\sqrt{\pi}} \sum_{i=0}^{i_0} \exp\left[-2ai^2 - \frac{2(i + 2)^2}{a}\right].
\]

Now note that
\[
2ai^2 - \frac{2(i + 2)^2}{a} = 2(a - 1)i^2 - 8i - 8 = 2(a - 1) \left(\frac{i^2 - \frac{4i}{a - 1} + \frac{4}{(a - 1)^2}}{a - 1}\right) - 8 \leq 2(a - 1) \left(\frac{i^2 - \frac{2}{a - 1}}{a - 1}\right)^2 - \frac{8a}{a - 1}.
\]
It follows that

\[
W \leq \frac{e^2 \sqrt{\pi}}{\sqrt{\pi}} \sum_{i=0}^{i_0} \exp \left[ -\frac{2(a-1)}{a} \left( i - \frac{2}{a-1} \right)^2 + \frac{8}{a-1} \right]
\]

\[
\leq \frac{2e^2 \sqrt{\pi}}{\sqrt{\pi}} \exp \left[ -\frac{8}{a-1} \sum_{i=0}^{\infty} \exp \left[ -\frac{2(a-1)}{a} \right] \right]
\]

\[
\leq \frac{2e^2 \sqrt{\pi}}{\sqrt{\pi}} \exp \left[ \frac{8}{a-1} \int_{0}^{\infty} \exp \left[ -\frac{2(a-1)}{a} \right] du \right]
\]

\[
= \frac{2e^2 \sqrt{\pi}}{\sqrt{\pi}} \exp \left[ \frac{8}{a-1} \left( 1 + \sqrt{\frac{\pi a}{8(a-1)}} \right) \right].
\]

This concludes our proof.

**Lemma 1.** Let \( f(X) \) be a non-negative decreasing function of random variable \( X \) in Theorem 4 and \( i_0 \) be the smallest integer such that \( (i_0+1)\sqrt{n} \geq X \). Then

\[
\mathbb{E}[f(X)] \leq \sum_{i=0}^{i_0-1} \exp\left[-2i^2\right] f(X - (i + 1)\sqrt{n}) + \exp\left[-2i_0^2\right] f(0).
\]

**Proof.** Let

\[
\mathcal{P}_i = \begin{cases}
\{\{X - \sqrt{n}, 0\}, n\}, & i = 0; \\
\{\{X - (i+1)\sqrt{n}, 0\}, X - i\sqrt{n}\}, & i > 0;
\end{cases}
\]

for \( i \in [i_0] \cup \{0\} \). Then \( \mathcal{P}_i \) is a partition of \([0, n]\). Based on this observation,

\[
\mathbb{E}[f(X)] = \sum_{i=0}^{i_0-1} \mathbb{E}[1\{X \in \mathcal{P}_i\} f(X)]
\]

\[
\leq \sum_{i=0}^{i_0-1} f(X - (i + 1)\sqrt{n}) \mathbb{P}(X \in \mathcal{P}_i) + f(0) \mathbb{P}(X \in \mathcal{P}_{i_0}),
\]

where the inequality holds because \( f(x) \) is a decreasing function of \( x \). Now fix \( i > 0 \). Then from the definition of \( \mathcal{P}_i \) and Hoeffding’s inequality, we have

\[
\mathbb{P}(X \in \mathcal{P}_i) \leq \mathbb{P}(X < X - i\sqrt{n}) \leq \exp[-2i^2].
\]

Trivially, \( \mathbb{P}(X \in \mathcal{P}_0) \leq 1 = \exp[-2 \cdot 0^2] \). Finally, we chain all inequalities and get our claim. ■

**Lemma 2.** Let \( m = 2an \). Then for any \( \delta \in [0, an] \),

\[
\sum_{y=\lfloor an + \delta \rfloor}^{\lfloor an + \delta \rfloor} B(y; m, 1/2) \geq \frac{\sqrt{\pi}}{e^2} \exp \left[ \frac{-2\left(\delta + \sqrt{n}\right)^2}{an} \right].
\]

**Proof.** By Lemma 4 in Appendix of Kveton et al. [2019b],

\[
B(y; m, 1/2) \geq \frac{\sqrt{2\pi}}{e^2} \sqrt{\frac{m}{y(m-y)}} \exp \left[ \frac{-2(y - an)^2}{an} \right].
\]

Also note that

\[
\frac{y(m-y)}{m} \leq \frac{1}{m} \frac{m^2}{4} = \frac{an}{2}
\]

for any \( y \in [0, m] \). Now we combine the above two inequalities and get

\[
B(y; m, 1/2) \geq \frac{\sqrt{\pi}}{e^2} \exp \left[ \frac{-2(y - an)^2}{an} \right].
\]

Finally, we note the following. First, the above lower bound decreases in \( y \) for \( y \geq an + \delta \), since \( \delta \geq 0 \). Second, by the pigeonhole principle, there are at least \( \lfloor \sqrt{n} \rfloor \) integers between \( an + \delta \) and \( an + \delta + \sqrt{n} \), starting with \( \lfloor an + \delta \rfloor \). This leads to the following lower bound

\[
\sum_{y=\lfloor an + \delta \rfloor}^{\lfloor an + \delta \rfloor} B(y; m, 1/2) \geq \sqrt{n} \frac{\sqrt{\pi}}{e^2} \exp \left[ \frac{-2\left(\delta + \sqrt{n}\right)^2}{an} \right]
\]

\[
\geq \frac{\sqrt{\pi}}{e^2} \exp \left[ \frac{-2\left(\delta + \sqrt{n}\right)^2}{an} \right].
\]

The last inequality is from \( \sqrt{n} / \sqrt{n} \geq 1/2 \), which holds for \( n \geq 1 \). This concludes our proof. ■

**References**


