arXiv:1811.05154v1 [cs.LG] 13 Nov 2018

Garbage In, Reward Out: Bootstrapping Exploration in Multi-Armed Bandits

Branislav Kveton
Google Research

Csaba Szepesvári
DeepMind and University of Alberta

Zheng Wen
Adobe Research

Mohammad Ghavamzadeh
Facebook AI Research

Tor Lattimore
DeepMind

Abstract

We propose a multi-armed bandit algorithm that explores based on randomizing its history. The key idea is to estimate the value of the arm from the bootstrap sample of its history, where we add pseudo observations after each pull of the arm. The pseudo observations seem to be harmful. But on the contrary, they guarantee that the bootstrap sample is optimistic with a high probability. Because of this, we call our algorithm `Giro`, which is an abbreviation for *garbage in, reward out*. We analyze `Giro` in a K-armed Bernoulli bandit and prove a $O(K\Delta^{-1} \log n)$ bound on its n-round regret, where Δ denotes the difference in the expected rewards of the optimal and best suboptimal arms. The main advantage of our exploration strategy is that it can be applied to any reward function generalization, such as neural networks. We evaluate `Giro` and its contextual variant on multiple synthetic and real-world problems, and observe that `Giro` is comparable to or better than state-of-the-art algorithms.

1 Introduction

Multi-armed bandits [20, 8, 6] are a classical framework for sequential decision-making under uncertainty, where the actions of the *learning agent* are represented by *arms*. The arms can be treatments in a clinical trial or advertisements on a website. Each arm has some *expected reward*. The agent does not know the expected rewards of arms in advance and has to learn them by pulling the arms. The goal of the agent is to maximize its expected cumulative reward. This results in the so-called *exploration-exploitation trade-off*: *explore*, gain more information about an arm; and *exploit*, pull the arm with the highest estimated reward thus far.

Contextual bandits [32, 27, 18, 11, 23, 4] are a generalization of multi-armed bandits where the learning agent has additional information in each round, in the form of context. The context can encode the medical data of a patient in a clinical trial or the demographic information of a targeted user on a website. In this case, the expected reward is an unknown function of the arm and context. This function is typically parametric and its parameters are learned from observations. In *linear bandits* [29, 11, 1], the function is linear, that is the expected reward is the dot product of a known context vector and an unknown parameter vector.

Arguably, three most popular exploration strategies in multi-armed and contextual bandits are the *ϵ-greedy policy (EG)* [21], *optimism in the face of uncertainty (OFU)* [5, 1], and *Thompson sampling (TS)* [4]. The ϵ-greedy policy is simple to implement and widely used in practice. However, it is statistically suboptimal. From a practical point of view, its performance depends heavily on choosing the right exploration parameter and the strategy for annealing it.

OFU methods act based on high-probability confidence sets, which are statistically and computationally efficient in the bandit [6] and linear bandit [1] settings. However, when the reward function is a non-linear function of context, we only know how to construct approximate confidence sets [15, 33, 24, 17]. These sets tend to be conservative [15] and statistically suboptimal.

TS constructs a posterior distribution over model parameters from its prior and observed rewards. Then it acts optimistically with respect to posterior samples. TS is computationally efficient when the posterior has a closed form, such in bandits with Bernoulli and Gaussian reward distributions. If the posterior does not have a closed from, it has

to be approximated. When the reward distribution is bounded on $[0, 1]$, computationally-efficient approximations exist [4]. In general, however, the posterior has to approximated by computationally-expensive techniques [16, 19, 28].

To address these issues, bootstrapping [14] has been proposed in both multi-armed and contextual bandits [7, 13, 26, 30, 25, 31]. Bootstrapping has two advantages over existing exploration strategies. First, unlike OFU and TS, it is simple to implement in any problem, because it does not require problem-specific confidence sets and posteriors. Second, unlike the ϵ-greedy policy, it is data-driven and not very sensitive to tuning. Despite its advantages and good empirical performance, bootstrapping is poorly understood theoretically in the bandit setting. The strongest theoretical result is that of Osband and Van Roy [26], who show that a form of bootstrapping in a Bernoulli bandit is equivalent to Thompson sampling. This result was recently rediscovered by Vaswani et al. [31] and generalized to other reward distributions, such as categorical and Gaussian.

This paper makes the following contributions. A natural exploration policy in a multi-armed bandit is to resample the history of the arm and then estimate the value of the arm by the empirical mean in its resampled history. First, we show that this policy can have linear regret. Second, we propose a computationally-efficient algorithm in a Bernoulli bandit, `Giro`, that addresses this issue by adding positive and negative pseudo observations to the history of the arm. The number of pseudo observations grows with the history and transforms the problem such that a bootstrap sample of the history is optimistic with a high probability. Third, we analyze `Giro` in a K-armed Bernoulli bandit and derive a $O(K\Delta^{-1} \log n)$ bound on its n-round regret, where Δ is the difference in the expected rewards of the optimal and best suboptimal arms. Fourth, we propose a contextual variant of `Giro`. Finally, we empirically evaluate `Giro` and its contextual variant on several synthetic and real-world problems. We observe that `Giro` is comparable to or better than state-of-the-art algorithms.

We adopt the following notation. The set $\{1, \dots, n\}$ is denoted by $[n]$. We define $\mathrm{Ber}(x; p) = p^x(1-p)^{1-x}$ and let $\mathrm{Ber}(p)$ be the corresponding Bernoulli distribution. In addition, we 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 $\mathcal{E}$, the indicator $\mathbb{1}\{\mathcal{E}\}$ takes the value of one when $\mathcal{E}$ occurs and the value of zero otherwise.

2 Background

Consider the following n-round online learning problem. In round $t \in [n]$, the learning agent observes context x_t with label Y_t, which is not revealed. The label is generated by some fixed unknown function of x_t. The agent predicts label I_t and receives reward $\mathbb{1}\{I_t = Y_t\}$. That is, the reward is one if the agent predicts correctly and zero otherwise. The goal of the learning agent is to predict labels in n rounds as well as the best model in hindsight.

Our learning problem can be formulated as a contextual bandit with K arms, where the context in round t is a feature vector $x_t \in \mathbb{R}^{d \times 1}$. The reward of arm $i \in [K]$ in round t, $V_{i,t} = \mathbb{1}\{i = Y_t\}$, is conditioned on context x_t. If

$$V_{i,t} \sim \mathrm{Ber}\left(\frac{1}{1 + \exp[-x_t^\intercal \theta_i]}\right)$$

for parameter vector $\theta_i \in \mathbb{R}^{d \times 1}$ of arm i, our problem could be solved as a generalized linear bandit [15]. However, if $V_{i,t}$ was a more complicated function of x_t, such as a neural network, we would not know how to design a sound bandit algorithm. The difficulty is not necessarily in modeling the uncertainty. For that, one can follow the approach in Lemma 2 of Lai and Robbins [20]. The problem is the lack of computationally efficient methods to do so.

We study a natural exploration strategy that can be applied to any estimator of rewards. Let $\mathcal{H}_{i,s}$ be a vector of s past observations of arm i, which represents the *history* of the arm. To behave optimistically, we resample a vector of the same length, $\mathcal{B}_{i,s}$, from entries of $\mathcal{H}_{i,s}$ with replacement and estimate the parameters associated with arm i from $\mathcal{B}_{i,s}$. Then we estimate the value of arm i from these parameters and x_t. The arm with the highest value is pulled.

The pseudocode of this algorithm is in Algorithm 1. We denote the number of observations of arm i in the first t rounds by $T_{i,t}$. The history of the arm is initialized by an empty vector $()$. The entries of the history $\mathcal{H}_{i,s}$ are pairs of context vectors and observed rewards. For vectors u and v, $u \oplus v$ denotes their concatenation. The parameter vector $\theta_{i,t}$ can be estimated by the maximum likelihood estimation, for instance.

Unfortunately, Algorithm 1 can have linear regret, and we show this in a non-contextual Bernoulli bandit in Section 3. We propose a solution to this problem in Section 4.1, again in a non-contextual Bernoulli bandit, and prove

Algorithm 1 Bootstrapping for contextual bandits. We do *not* recommend this algorithm. Our recommended variant of this algorithm is proposed in Section 4.3.

for all $i \in [K]$ **do** ▷ Initialization
  $T_{i,0} \leftarrow 0$, $\mathcal{H}_{i,0} \leftarrow ()$
for $t = 1, \ldots, n$ **do**
  for all $i \in [K]$ **do** ▷ Choose the arm
    if $T_{i,t-1} > 0$ **then**
      $s \leftarrow T_{i,t-1}$
      $\mathcal{B}_{i,s} \leftarrow$ Sample $|\mathcal{H}_{i,s}|$ times from $\mathcal{H}_{i,s}$
      $\theta_{i,t} \leftarrow$ Estimate model parameters θ_i from $\mathcal{B}_{i,s}$
      $\hat{V}_{i,t} \leftarrow$ Estimate value of arm i, $\mathbb{E}\left[V_{i,t} \mid x_t\right]$, under model $\theta_{i,t}$
    else
      $\hat{V}_{i,t} \leftarrow +\infty$
  $I_t \leftarrow \arg\max_{i \in [K]} \hat{V}_{i,t}$ ▷ Pull the arm
  Pull arm I_t and get reward $V_{I_t,t} = \mathbb{1}\{I_t = Y_t\}$
  for all $i \in [K]$ **do** ▷ Update statistics
    if $i = I_t$ **then**
      $T_{i,t} \leftarrow T_{i,t-1} + 1$
      $\mathcal{H}_{i,T_{i,t}} \leftarrow \mathcal{H}_{i,T_{i,t-1}} \oplus ((x_t, V_{i,t}))$
    else
      $T_{i,t} \leftarrow T_{i,t-1}$

that the resulting algorithm has sublinear regret. Based on this solution, we suggest a straightforward modification of Algorithm 1 in Section 4.3 that performs well in practice. We leave the analysis of that algorithm for future work.

3 Lower Bound

We show that Algorithm 1 can have linear regret in a Bernoulli bandit with $K = 2$ arms. Let μ_i denote the expected reward of arm $i \in [K]$, $\mu_1 > \mu_2$, and $\Delta = \mu_1 - \mu_2$. This problem is not contextual, and therefore we set $x_t = \emptyset$ for all $t \in [n]$. The n-round regret is defined as $R(n) = \mathbb{E}\left[\sum_{t=1}^{n} \Delta \mathbb{1}\{I_t = 2\}\right]$.

Algorithm 1 is implemented as follows. Because the problem is not contextual, we directly estimate the value of arm i from $\mathcal{B}_{i,s}$ as

$$\theta_{i,t} = \frac{1}{|\mathcal{H}_{i,s}|} \sum_{(x,v) \in \mathcal{B}_{i,s}} v \,.$$

When $\theta_{1,t} = \theta_{2,t}$, the ties are broken by a fixed rule that is chosen randomly in advance. In particular, Algorithm 1 draws $Z \sim \mathrm{Ber}(0.5)$ before the start of round 1. In any round $t \in [n]$, if $Z = 0$ and $\theta_{1,t} = \theta_{2,t}$, $I_t = 1$; and if $Z = 1$ and $\theta_{1,t} = \theta_{2,t}$, $I_t = 2$. The following lemma shows that the expected n-round regret of Algorithm 1 can be linear.

Lemma 1. *In a Bernoulli bandit with 2 arms and $\mu_1 > \mu_2$, the expected n-round regret of the above implementation of Algorithm 1 is bounded from below as $R(n) \geq 0.5(1 - \mu_1)\Delta(n - 2)$.*

Proof. By the design of the algorithm, arm 1 is never pulled after events $Z = 1$ and $\mathcal{B}_{1,1} = \{(\emptyset, 0)\}$ occur. Since the events are independent,

$$\mathbb{P}\left(Z = 1,\, \mathcal{B}_{1,1} = \{(\emptyset, 0)\}\right) = \mathbb{P}\left(Z = 1\right) \mathbb{P}\left(\mathcal{B}_{1,1} = \{(\emptyset, 0)\}\right) = \frac{1 - \mu_1}{2} \,.$$

Moreover, if $\mathcal{B}_{1,1} = \{(\emptyset, 0)\}$ occurs, it must occur by the end of round 2 because Algorithm 1 pulls all arms once in the first K rounds (line 11). Finally, we combine the above two facts and get

$$\begin{aligned} R(n) &\geq \mathbb{E}\left[\left(\sum_{t=1}^{n} \Delta \mathbb{1}\{I_t = 2\}\right) \mathbb{1}\{Z = 1,\, \mathcal{B}_{1,1} = \{(\emptyset, 0)\}\}\right] \\ &\geq \mathbb{E}\left[\sum_{t=1}^{n} \Delta \mathbb{1}\{I_t = 2\} \,\middle|\, Z = 1,\, \mathcal{B}_{1,1} = \{(\emptyset, 0)\}\right] \mathbb{P}\left(Z = 1,\, \mathcal{B}_{1,1} = \{(\emptyset, 0)\}\right) \\ &\geq \frac{(1-\mu_1)\Delta(n-2)}{2}\,. \end{aligned}$$

This concludes the proof. ■

4 Algorithm

One seemingly naive solution to the issues in Section 3 is to add a sufficient number of positive and negative pseudo observations to $\mathcal{H}_{i,s}$. This increases the variance of the bootstrap distribution, which may lead to optimism and thus exploration. However, the pseudo observations also introduce bias, which needs to be controlled.

4.1 Bernoulli `Giro`

Algorithm 2 shows the pseudocode of the algorithm that implements the above idea in a Bernoulli bandit. We refer to it as `Giro`, which is an abbreviation for *garbage in, reward out*. This is an informal description of our exploration strategy, which adds seemingly useless observations to the history of the pulled arm. We refer to these observations as *pseudo observations*, to distinguish them from the actual observations. `Giro` has one tunable parameter, the number of positive and negative pseudo observations a that are added to the history of the arm after each pull.

`Giro` does not seem sound, because it adds a large number of pseudo observations to the history of the arm. It is sound and performs well for the following reason. For the appropriate choice of a, the variance of the bootstrap distribution is guaranteed to be above that of the history. Therefore, the probability of generating a bootstrap sample with a higher mean than the mean history is higher than the probability of that history, for any history. Then, roughly speaking, history resampling is a good exploration strategy.

4.2 Optimistic Design

In this section, we informally argue for the correctness of `Giro`. Specifically, we analyze $\theta_{i,t}$ in line 9, and show that it concentrates and is sufficiently optimistic. The formal regret analysis is deferred to Section 5.

Before we analyze the distribution of $\theta_{i,t}$, we analyze the history $\mathcal{H}_{i,s}$ and its bootstrap sample $\mathcal{B}_{i,s}$. Fix round t, arm i, and the number of pulls s of arm i. Let $X_{i,s}$ be the number of ones in history $\mathcal{H}_{i,s}$, which includes a positive and negative pseudo observations per each observation of arm i. Because of this, $X_{i,s} - as$ is the number of positive observations of arm i. These observations are drawn independently from $\mathrm{Ber}(\mu_i)$, and hence $X_{i,s} - as \sim B(s, \mu_i)$. From the definition of $X_{i,s}$,

$$\mathbb{E}\left[X_{i,s}\right] = (a + \mu_i)s\,, \quad \mathrm{var}\left[X_{i,s}\right] = \mu_i(1-\mu_i)s\,. \tag{1}$$

Let $\alpha = (2a+1)$ and $Y_{i,s}$ be the number of ones in the bootstrap sample $\mathcal{B}_{i,s}$. Then $Y_{i,s} \sim B\left(\alpha s, X_{i,s}/(\alpha s)\right)$, since it is equivalent to draw αs samples with replacement from $\mathcal{H}_{i,s}$ and αs independent samples from $\mathrm{Ber}(X_{i,s}/(\alpha s))$. From the definition of $Y_{i,s}$,

$$\mathbb{E}\left[Y_{i,s} \mid X_{i,s}\right] = X_{i,s}\,, \quad \mathrm{var}\left[Y_{i,s} \mid X_{i,s}\right] = \frac{X_{i,s}}{\alpha s}\left(1 - \frac{X_{i,s}}{\alpha s}\right)\alpha s \tag{2}$$

for any fixed $X_{i,s}$.

Algorithm 2 `Giro` for Bernoulli bandits.

Inputs: Pseudo observations per unit of history a

for all $i \in [K]$ **do** ▷ Initialization
  $T_{i,0} \leftarrow 0$, $\mathcal{H}_{i,0} \leftarrow ()$
for all $t \in [n]$ **do**
  for all $i \in [K]$ **do** ▷ Choose the arm
    if $T_{i,t-1} > 0$ **then**
      $s \leftarrow T_{i,t-1}$
      $\mathcal{B}_{i,s} \leftarrow$ Sample $|\mathcal{H}_{i,s}|$ times from $\mathcal{H}_{i,s}$
      $\theta_{i,t} \leftarrow \dfrac{1}{|\mathcal{H}_{i,s}|} \displaystyle\sum_{v \in \mathcal{B}_{i,s}} v$
    else
      $\theta_{i,t} \leftarrow +\infty$
  $I_t \leftarrow \arg\max_{i \in [K]} \theta_{i,t}$ ▷ Pull the arm
  Pull arm I_t and get reward $V_{I_t,t}$
  for all $i \in [K]$ **do** ▷ Update statistics
    if $i = I_t$ **then**
      $T_{i,t} \leftarrow T_{i,t-1} + 1$
      $\mathcal{H}_{i,T_{i,t}} \leftarrow \mathcal{H}_{i,T_{i,t-1}} \oplus (V_{i,t})$
      for all $\ell \in [a]$ **do**
        $\mathcal{H}_{i,T_{i,t}} \leftarrow \mathcal{H}_{i,T_{i,t}} \oplus (0, 1)$
    else
      $T_{i,t} \leftarrow T_{i,t-1}$

Now we are ready to analyze $\theta_{i,t}$. First, we argue for the concentration of $\theta_{i,t}$. Since $\theta_{i,t} = Y_{i,s}/(\alpha s)$, we have from the properties of $Y_{i,s}$ in (2) that

$$\mathbb{E}\left[\theta_{i,t} \,|\, X_{i,s}\right] = \frac{X_{i,s}}{\alpha s}\,, \quad \operatorname{var}\left[\theta_{i,t} \,|\, X_{i,s}\right] = O(1/s)$$

for any fixed $X_{i,s}$. Moreover, we have from the properties of $X_{i,s}$ in (1) that

$$\mathbb{E}\left[\frac{X_{i,s}}{\alpha s}\right] = \frac{a + \mu_i}{\alpha}\,, \quad \operatorname{var}\left[\frac{X_{i,s}}{\alpha s}\right] = O(1/s)\,.$$

Based on the above equalities, $\theta_{i,t}$ concentrates at $(a + \mu_i)/\alpha$, the scaled and shifted expected reward of arm i, at the rate of $1/\sqrt{s}$. This transformation does not change the order of arms. However, it changes the gaps, the differences in the expected rewards of the optimal and suboptimal arms.

Second, we argue that $\theta_{i,t}$ is optimistic. To show this, we examine the variances of $X_{i,s}$ and $Y_{i,s}$ conditioned on $X_{i,s}$. Trivially, $\operatorname{var}\left[X_{i,s}\right] \leq s/4$. From the properties of $Y_{i,s}$ in (2) and that $X_{i,s} \in [as, (a+1)s]$,

$$\operatorname{var}\left[Y_{i,s} \,|\, X_{i,s}\right] \geq \alpha s \min_{x \in [as,(a+1)s]} \frac{x}{\alpha s}\left(1 - \frac{x}{\alpha s}\right) = \alpha s \frac{as}{\alpha s}\left(1 - \frac{as}{\alpha s}\right) = \frac{a(a+1)}{2a+1} s\,.$$

It follows that for any number of pseudo observations a that satisfy

$$\frac{a(a+1)}{2a+1} \geq \frac{1}{4}\,,$$

including $a = 1$, $\mathrm{var}\left[Y_{i,s} \mid X_{i,s}\right] \geq \mathrm{var}\left[X_{i,s}\right]$ for all $X_{i,s}$; and in turn $\mathrm{var}\left[\theta_{i,t} \mid X_{i,s}\right] \geq \mathrm{var}\left[X_{i,s}/(\alpha s)\right]$ for all $X_{i,s}$. Since $\mathbb{E}\left[\theta_{i,t} \mid X_{i,s}\right] = X_{i,s}/(\alpha s)$ and $\mathbb{E}\left[X_{i,s}/(\alpha s)\right] = (a + \mu_i)/\alpha$, this indicates that

$$\mathbb{P}\left(X_{i,s} = x\right) \leq \mathbb{P}\left(X_{i,s} \leq x\right) \leq \mathbb{P}\left(\theta_{i,t} \geq \frac{a + \mu_i}{\alpha} \,\middle|\, X_{i,s} = x\right)$$

for any $x \leq (a + \mu_i)s$, under normal-like assumptions on the distributions of $\theta_{i,t}$ and $X_{i,s}/(\alpha s)$. Therefore, for any fixed history $X_{i,s} = x$, the probability of generating a bootstrap sample with a higher mean than the mean history is higher than the probability of that history, which we wanted to show.

4.3 Contextual `Giro`

It is straightforward to propose contextual `Giro`. In particular, it is Algorithm 1 where we add a times

$$\mathcal{H}_{i,T_{i,t}} \leftarrow \mathcal{H}_{i,T_{i,t}} \oplus ((x_t, 0), (x_t, 1))$$

in line 17, to increase the conditional variance of observations given context. The function approximation technique in Algorithm 1 should permit any constant shift of any representable function. In linear models, this can be achieved by adding a constant bias feature to the feature vector.

5 Analysis

This section has two subsections. In Section 5.1, we prove an upper bound on the expected cumulative regret of `Giro` in a Bernoulli bandit. In Section 5.2, we discuss the results of our analysis.

5.1 Regret Bound

We analyze `Giro` in a K-armed Bernoulli bandit. Without loss of generality, we assume that arm 1 is optimal, that is $\mu_1 > \max_{i \neq 1} \mu_i$. The gap of arm i is $\Delta_i = \mu_1 - \mu_i$. The *expected n-round regret* is defined as

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

Our regret bound is stated below.

Theorem 1. *The expected n-round regret of* `Giro` *is*

$$R(n) \leq \sum_{i=2}^{K} \Delta_i \left[\underbrace{\left(\frac{16(2a+1)c}{\Delta_i^2} \log n + 2\right)}_{\text{Upper bound on } a_i \text{ in Theorem 3}} + \underbrace{\left(\frac{8(2a+1)}{\Delta_i^2} \log n + 2\right)}_{\text{Upper bound on } b_i \text{ in Theorem 3}} \right], \tag{3}$$

where

$$b = \frac{2a+1}{a+1}, \quad c = \frac{2e^2\sqrt{2a+1}}{\sqrt{2\pi}} \exp\left[\frac{8b}{2-b}\right] \left(1 + \sqrt{\frac{2\pi}{4-2b}}\right). \tag{4}$$

Proof. Our claim is proved as an instance of Theorem 3 in Appendix B, because `Giro` is an instance of Algorithm 3 there. With the notation in Appendix B, $Q_{i,s}(\tau)$ in (5) is the probability that a sample from the bootstrap distribution of arm i after s pulls exceeds τ, conditioned on the history of the arm.

Let i be the index of any suboptimal arm. Based on Section 4.2, the bootstrap distribution of arm i concentrates at $\mu_i' = (\mu_i + a)/\alpha$, where $\alpha = 2a + 1$. Following the discussion on the choice of τ_i that precedes Theorem 3, we select $\tau_i = (\mu_i + a)/\alpha + \Delta_i/(2\alpha)$, which is halfway between μ_i' and μ_1'. As in Section 4.2, let $X_{i,s}$ be the number of ones

in history $\mathcal{H}_{i,s}$, which includes pseudo observations, and $Y_{i,s}$ be the number of ones in its bootstrap sample $\mathcal{B}_{i,s}$. We abbreviate $Q_{i,s}(\tau_i)$ as $Q_{i,s}$, and define it based on the above as

$$Q_{i,s} = \mathbb{P}\left(\frac{Y_{i,s}}{\alpha s} \geq \frac{\mu_i + a}{\alpha} + \frac{\Delta_i}{2\alpha} \,\middle|\, X_{i,s}\right) \text{ for } s > 0\,,$$
$$Q_{i,0} = 1\,,$$

where $Q_{i,0} = 1$ because of the optimistic initialization in line 11 of `Giro`. Note that $X_{i,s}$, $Y_{i,s}$, and $Q_{i,s}$ are random variables.

Upper Bound on b_i in Theorem 3 (Appendix B)

Our first objective it to bound

$$b_i = \sum_{s=0}^{n-1} \mathbb{P}\left(Q_{i,s} > 1/n\right) + 1\,.$$

Fix the number of pulls s. When the number of pulls is "small", $s \leq \frac{8\alpha}{\Delta_i^2}\log n$, we bound $\mathbb{P}\left(Q_{i,s} > 1/n\right)$ trivially by 1. When the number of pulls is "large", $s > \frac{8\alpha}{\Delta_i^2}\log n$, we divide the proof based on the event that $X_{i,s}$ is not much larger than its expectation. Define

$$\mathcal{E} = \left\{X_{i,s} - (\mu_i + a)s \leq \frac{\Delta_i s}{4}\right\}.$$

On event $\mathcal{E}$,

$$Q_{i,s} = \mathbb{P}\left(Y_{i,s} - (\mu_i + a)s \geq \frac{\Delta_i s}{2} \,\middle|\, X_{i,s}\right) \leq \mathbb{P}\left(Y_{i,s} - X_{i,s} \geq \frac{\Delta_i s}{4} \,\middle|\, X_{i,s}\right) \leq \exp\left[-\frac{\Delta_i^2 s}{8}\right] \leq n^{-1}\,,$$

where the first inequality is from the definition of event $\mathcal{E}$, the second inequality is by Hoeffding's inequality, and the third inequality is by our assumption on s. On the other hand, event $\overline{\mathcal{E}}$ is unlikely because

$$\mathbb{P}\left(\overline{\mathcal{E}}\right) \leq \exp\left[-\frac{\Delta_i^2 s}{8}\right] \leq n^{-1}\,,$$

where the first inequality is by Hoeffding's inequality and the last inequality is by our assumption on s. Now we apply the last two inequalities to

$$\begin{aligned}\mathbb{P}\left(Q_{i,s} > 1/n\right) &= \mathbb{E}\left[\mathbb{P}\left(Q_{i,s} > 1/n \,|\, X_{i,s}\right) \mathbb{1}\{\mathcal{E}\}\right] + \mathbb{E}\left[\mathbb{P}\left(Q_{i,s} > 1/n \,|\, X_{i,s}\right) \mathbb{1}\left\{\overline{\mathcal{E}}\right\}\right] \\ &\leq 0 + \mathbb{E}\left[\mathbb{1}\left\{\overline{\mathcal{E}}\right\}\right] \leq n^{-1}\,.\end{aligned}$$

Finally, we chain our upper bounds for all $s \in [n]$ and get the upper bound on b_i in (3).

Upper Bound on a_i in Theorem 3 (Appendix B)

Our second objective it to bound

$$a_i = \sum_{s=0}^{n-1} \mathbb{E}\left[\min\left\{\frac{1}{Q_{1,s}(\tau_i)} - 1, n\right\}\right].$$

We redefine τ_i as $\tau_i = (\mu_1 + a)/\alpha - \Delta_i/(2\alpha)$ and abbreviate $Q_{1,s}(\tau_i)$ as $Q_{1,s}$. Since i is fixed, this slight abuse of notation should not cause any confusion. For $s > 0$, we have

$$Q_{1,s} = \mathbb{P}\left(\frac{Y_{1,s}}{\alpha s} \geq \frac{\mu_1 + a}{\alpha} - \frac{\Delta_i}{2\alpha} \,\middle|\, X_{1,s}\right).$$

Let $F_s = 1/Q_{1,s} - 1$. Fix the number of pulls s. When $s = 0$, $Q_{1,s} = 1$ and $\mathbb{E}\left[\min\left\{F_s, n\right\}\right] = 0$. When the number of pulls is "small", $0 < s \leq \dfrac{16\alpha}{\Delta_i^2}\log n$, we apply the upper bound from Theorem 2 in Appendix A and get

$$\mathbb{E}\left[\min\left\{F_s, n\right\}\right] \leq \mathbb{E}\left[1/Q_{1,s}\right] \leq \mathbb{E}\left[\frac{1}{\mathbb{P}\left(Y_{1,s} \geq (\mu_1 + a)s \,|\, X_{1,s}\right)}\right] \leq c\,,$$

where c is defined in (4). The last inequality is by Theorem 2 for $p = \mu_1$ and $n = s$.

When the number of pulls is "large", $s > \dfrac{16\alpha}{\Delta_i^2}\log n$, we divide the proof based on the event that $X_{1,s}$ is not much smaller than its expectation. Define

$$\mathcal{E} = \left\{(\mu_1 + a)s - X_{1,s} \leq \frac{\Delta_i s}{4}\right\}.$$

On event $\mathcal{E}$,

$$\begin{aligned} Q_{1,s} &= \mathbb{P}\left((\mu_1 + a)s - Y_{1,s} \leq \frac{\Delta_i s}{2} \,\middle|\, X_{1,s}\right) = 1 - \mathbb{P}\left((\mu_1 + a)s - Y_{1,s} > \frac{\Delta_i s}{2} \,\middle|\, X_{1,s}\right) \\ &\geq 1 - \mathbb{P}\left(X_{1,s} - Y_{1,s} > \frac{\Delta_i s}{4} \,\middle|\, X_{1,s}\right) \geq 1 - \exp\left[-\frac{\Delta_i^2 s}{8\alpha}\right] \geq \frac{n^2 - 1}{n^2}\,, \end{aligned}$$

where the first inequality is from the definition of event $\mathcal{E}$, the second inequality is by Hoeffding's inequality, and the third inequality is by our assumption on s. The above 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 $\overline{\mathcal{E}}$ is unlikely because

$$\mathbb{P}\left(\overline{\mathcal{E}}\right) \leq \exp\left[-\frac{\Delta_i^2 s}{8}\right] \leq n^{-2}\,,$$

where the first inequality is by Hoeffding's inequality and the last inequality is by our assumption on s. Now we apply the last two inequalities to

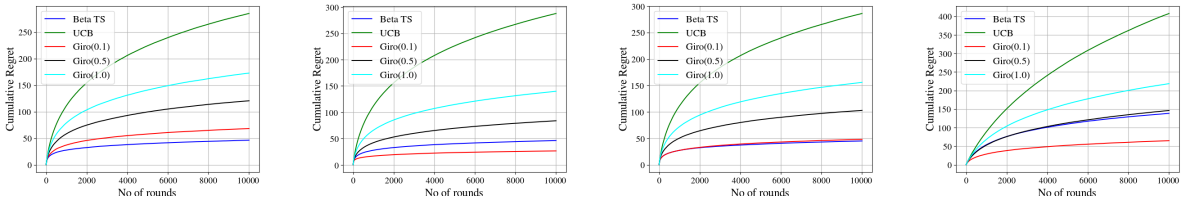
$$\begin{aligned} \mathbb{E}\left[\min\left\{F_s, n\right\}\right] &= \mathbb{E}\left[\mathbb{E}\left[\min\left\{F_s, n\right\} \,|\, X_{1,s}\right] \mathbb{1}\{\mathcal{E}\}\right] + \mathbb{E}\left[\mathbb{E}\left[\min\left\{F_s, n\right\} \,|\, X_{1,s}\right] \mathbb{1}\left\{\overline{\mathcal{E}}\right\}\right] \\ &\leq \mathbb{E}\left[n^{-1}\mathbb{1}\{\mathcal{E}\}\right] + \mathbb{E}\left[n\mathbb{1}\left\{\overline{\mathcal{E}}\right\}\right] \leq 2n^{-1}\,. \end{aligned}$$

Finally, we chain our upper bounds for all $s \in [n]$ and get the upper bound on a_i in (3). This concludes our proof. ■

5.2 Discussion

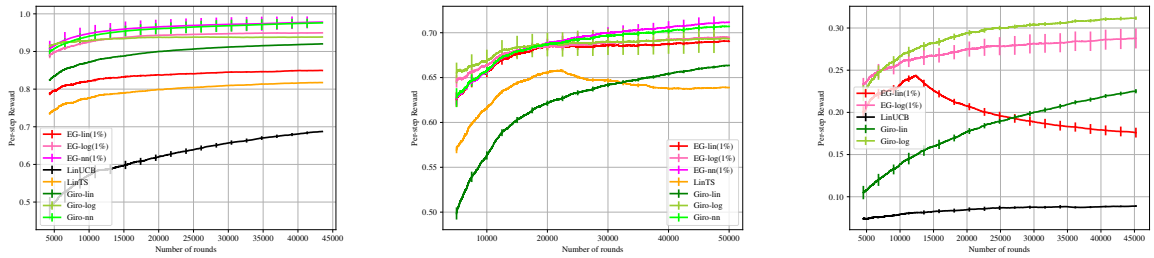
The regret bound of `Giro` is presented in Theorem 1. It is $O(K\Delta^{-1}\log n)$, where K is the number of arms, $\Delta = \min_{i>1}\Delta_i$ is the minimum gap, and n is the number of rounds. Therefore, it scales with all quantities of interest as the bounds of other popular bandit algorithms.

We would like to discuss the role of two constants in Theorem 1, a and c. The bound increases with the number of pseudo observations a. This is expected, since pseudo observations turn the original problem into a problem with



(a) Bernoulli (b) Truncated normal (c) Beta (d) Triangular

Figure 1: The expected n-round regret of `UCB1`, TS, and `Giro` in four synthetic bandit problems in Section 6.1. The regret is reported as a function of round n.



(a) Statlog (b) CovType (c) Adult

Figure 2: The expected per-round reward in n rounds in three contextual bandit problems in Section 6.2. The reward is reported as a function of round n.

$2a+1$ times smaller gaps (Section 4.2). The benefit of this transformation is that exploration becomes easy. Although the gaps are smaller, we observe that `Giro` outperforms `UCB1` in all synthetic experiments in Section 6.1, even when $a = 1$. In the same experiments, we also observe that the regret of `Giro` increases with a.

Our upper bound on the inverse probability of being optimistic c is defined for all $a > 1/\sqrt{2}$. It can be large due to factor $\exp[8b(2-b)^{-1}]$. For instance, when $a = 1$, $b = 3/2$ and the factor is e^{24}. The good news is that the factor drops significantly with increasing a. At $a = 2$, $b = 5/6$ and the factor is about $e^{5.7}$; and at $a = 3$, $b = 7/12$ and the factor is about $e^{3.3}$. We believe that c is large due to the looseness of the upper bound in Appendix A. In numerical experiments, we observed the maximum inverse probability of being optimistic around 3, when $a = 1$.

6 Experiments

Our experiments are organized as follows. In Section 6.1, we compare `Giro` to Thompson sampling and `UCB1` in four synthetic multi-armed bandit problems. In Section 6.2, we evaluate `Giro` in three contextual bandit problems. We experiment with multiple reward function generalizations, including neural networks.

6.1 Multi-Armed Bandit

We experiment with $K = 10$ arms, a horizon of $n = 10$k rounds, and average our results over 1k runs. We consider four reward distributions on $[0, 1]$: Bernoulli, truncated normal, beta, and the triangular distribution. In each run, we draw the expected reward μ_i of each arm i uniformly at random from $[0, 1]$. Then we fit the parameters of the reward distribution such that its mean is μ_i. In the truncated normal distribution, the standard deviation is $\sigma = 10^{-4}$. In the beta distribution, the shape parameters of arm i are set as $\alpha = \mu_i$ and $\beta = 1 - \mu_i$.

We run Giro with different values of a: 1, 0.5, and 0.1. In Thompson sampling, the prior is $\mathrm{Beta}(1, 1)$. We adopt the procedure of Agrawal and Goyal [3] to run TS with non-Bernoulli rewards. In particular, if the reward of arm i in round t is $V_{i,t} \in [0, 1]$, we sample pseudo reward $\hat{V}_{i,t} \sim \mathrm{Ber}(V_{i,t})$ and then update the posterior of arm i with it. The confidence interval in UCB1 is $\sqrt{1.5 \log(t-1)/s}$, where s is the number of pulls of the arm up to round t.

Our results are reported in Figure 1. We observe two main trends. First, Giro consistently outperforms UCB1 for all values of a. Second, for $a = 0.1$, Giro is comparable to or better than Thompson sampling.

6.2 Contextual Bandit

We adopt the one-versus-all multi-class classification setting for the evaluation of contextual bandits [2, 25, 28]. In this setting, arm i corresponds to class $i \in [K]$. In round t, the algorithm observes context vector $x_t \in \mathbb{R}^{d \times 1}$ and then pulls an arm. It receives a reward of one if the pulled arm corresponds to the correct class, and zero otherwise. Each arm maintains an independent set of statistics that map x_t to the observed binary reward. We use three multi-class datasets from Riquelme et al. [28]: Statlog ($d = 9, K = 7$), CovType ($d = 54, K = 7$) and Adult ($d = 94, K = 14$).

The horizon is $n = 50$k rounds and our results are averaged over 5 runs. We compare Giro to LinUCB [1], linear TS (LinTS) [4], and the ϵ-greedy policy (EG) [21]. We also implemented GLM-UCB [24]. GLM-UCB consistently over-explored and had worse performance than EG and LinTS. Therefore, we do not report these results in this paper.

We run EG and Giro with three classes of models: linear regression (suffix *lin* in plots), logistic regression (suffix *log* in plots), and a single hidden-layer fully-connected neural network (suffix *nn* in plots) with 10 hidden neurons. We experimented with different exploration schedules in EG. The best performing schedule across all three datasets was $\epsilon_t = b/t$, where b is set to achieve 1% exploration in n rounds. Note that this gives EG an unfair advantage over other compared methods, since such tuning cannot be done for a new online problem.

In Giro, we set $a = 1$ in all experiments. We solve the maximum likelihood estimation (MLE) problem in each round using stochastic optimization, which is initialized by the solution from the previous round. In linear and logistic regression, we optimize until the error drops below 10^{-3}. In neural networks, we make one pass over the history in each round. To ensure that our results do not depend on the specific choice of optimization, we use publicly available optimization libraries. In linear and logistic regression, we use scikit-learn [9] with stochastic optimization and its default options. In neural networks, we use the Keras library [10] with the ReLU non-linearity for hidden layers and a sigmoid output layer, along with SGD and its default configuration. This is in contrast to McNellis et al. [25] and Tang et al. [30], who approximate bootstrapping by an ensemble of models. Our preliminary experiments suggested that our procedure yields similar solutions to McNellis et al. [25] with lower run time, and better solutions than Tang et al. [30] without any tuning.

Since we compare different bandit algorithms and reward generalization models, we use the expected per-round reward in n rounds, $\mathbb{E}\left[\sum_{t=1}^{n} V_{I_t,t}\right]/n$, as our performance metric. The expected per-round reward of all algorithms in all three datasets is reported in Figure 2. We observe the following trends. First, both linear methods, LinTS and LinUCB, perform the worst.[1] Second, linear Giro is comparable to linear EG in the Statlog and CovType datasets. In the Adult dataset, EG does not explore enough for the relatively larger number of arms. In contrast, Giro explores enough and performs well. Third, non-linear variants of EG and Giro generally outperform their linear counterparts. The most expressive model, neural networks, outperforms logistic regression in both the Statlog and CovType datasets. In the Adult dataset, the neural network, which we do not plot, performs the worst. To investigate this further, we trained a neural network offline for each arm with all available data. Even then, we observed that the neural network performs worse than linear regression. We conclude that the poor performance is due to poor generalization, and not because of the lack of exploration.

7 Related Work

Osband and Van Roy [26] proposed bootstrapping exploration as an alternative to Thompson sampling. Interestingly, they also showed equivalence of *weighted bootstrapping* and Thompson sampling in a Bernoulli bandit. In particular,

[1] To avoid clutter in the plots, we do not report the performance of LinUCB and LinTS in the CovType and Adult datasets, respectively. They are the worst performing methods in these datasets.

let $\mathrm{Beta}(n_1+1, n_0+1)$ be the posterior of an arm with $\mathrm{Beta}(1,1)$ prior, n_1 positive observations, and n_0 negative observations. Then sampling $X \sim \mathrm{Beta}(n_1+1, n_0+1)$ is equivalent to $X = (\sum_{i=1}^{n_1+1} W_i)/(\sum_{i=1}^{n_0+n_1+2} W_i)$, where $W_i \sim \mathrm{Exp}(1)$ is an i.i.d. exponential weight, for any $i \in [n_0+n_1+2]$. Vaswani et al. [31] rediscovered this result and generalized it to other reward distributions, such as categorical and Gaussian. They also proposed contextual bandit algorithms based on weighted bootstrapping. We do not build on this view of bootstrapping. In comparison, we study non-parametric bootstrapping and justify it without the need for posteriors.

Baransi et al. [7] proposed an ensemble method that approximates the posterior of the arm by multiple bootstrap samples of its history. In round t, the agent chooses one sample uniformly at random for each arm and estimates the value of the arm from it. The observed reward is added with probability 0.5 to each sample of the history. A similar approach was proposed in contextual bandits by Tang et al. [30]. The main difference is that the observed reward is added to all samples of the history with a random Poisson weight that determines its importance. Tang et al. [30] also used SGD to update the model associated with each sample online. McNellis et al. [25] proposed bootstrapping for the exploration of decision trees in contextual bandits. Both McNellis et al. [25] and Tang et al. [30] provide limited theoretical evidence that justify bootstrapping as an approximation to posterior sampling. No regret bound is proved in these papers, and their theoretical results are not strong enough for that. We prove a regret bound and depart from the traditional view that bootstrapping is an approximation to posterior sampling.

An interesting resampling technique for equalizing the histories of arms was studied by Baransi et al. [7]. Let n_1 and n_2 be the number of observations of arms 1 and 2, respectively. Let $n_1 < n_2$. Then the value of arm 1 is estimated by its empirical mean and the value of arm 2 is estimated by the empirical mean of the bootstrap sample of its history of size n_1. Baransi et al. [7] also proved a regret bound in a 2-armed bandit. This approach is different from `Giro` and it is not clear how to generalize it to contextual bandits.

8 Conclusions

We propose a novel exploration strategy for stochastic multi-armed bandits. The key idea is to estimate the value of the arm from the bootstrap sample of its history, where we add pseudo observations after each pull of the arm. We propose an algorithm based on this idea, which we call `Giro`, and derive a $O(K\Delta^{-1}\log n)$ bound on its n-round regret in a Bernoulli bandit with K arms, where Δ is the difference in the expected rewards of the optimal and best suboptimal arms. We extend `Giro` to contextual bandits, and evaluate it extensively on both synthetic and real-world problems. `Giro` is comparable to or better than state-of-the-art algorithms.

We leave open many questions of interest. For instance, our analysis of randomized exploration in Appendix B is extremely general, as we do not make any strong assumptions on the history resampling distribution. The analysis of `Giro` is an instance of this analysis where the resampling distribution is bootstrapping from the history with pseudo observations. The distribution-specific part of the analysis is in Appendix A. We strongly believe that this part can be extended to other bandit problems, such as linear and contextual bandits, and other sampling distributions.

History resampling is computationally demanding, since the history of the arm may need to be resampled in each round. In a Bernoulli bandit (Section 4.1), `Giro` can be implemented efficiently, because the estimated value of the arm can be drawn from a binomial distribution. In general, approximations may be needed to achieve computational efficiency. We suggest and evaluate one such approximation in Section 6.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

[2] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646, 2014.

[3] Shipra Agrawal and Navin Goyal. Further optimal regret bounds for Thompson sampling. In *Artificial Intelligence and Statistics*, pages 99–107, 2013.

[4] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, 2013.

[5] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[6] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[7] Akram Baransi, Odalric-Ambrym Maillard, and Shie Mannor. Sub-sampling for multi-armed bandits. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 115–131. Springer, 2014.

[8] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[9] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[10] Franois Chollet. KERAS. https://github.com/fchollet/keras, 2015.

[11] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *COLT*, pages 355–366, 2008.

[12] J. L. Doob. *Stochastic processes*. Wiley, 1953.

[13] Dean Eckles and Maurits Kaptein. Thompson sampling with the online bootstrap. *arXiv preprint arXiv:1410.4009*, 2014.

[14] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

[15] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.

[16] Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex online problems. In *Proceedings of the 31st International Conference on Machine Learning*, pages 100–108, 2014.

[17] Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable generalized linear bandits: Online computation and hashing. *arXiv preprint arXiv:1706.00136*, 2017.

[18] Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th international conference on Machine learning*, pages 440–447. ACM, 2008.

[19] Jaya Kawale, Hung Bui, Branislav Kveton, Long Tran-Thanh, and Sanjay Chawla. Efficient Thompson sampling for online matrix-factorization recommendation. In *Advances in Neural Information Processing Systems 28*, pages 1297–1305, 2015.

[20] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[21] John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008.

[22] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2018. (to appear).

[23] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

[24] Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. *arXiv preprint arXiv:1703.00048*, 2017.

[25] Ryan McNellis, Adam N. Elmachtoub, Sechan Oh, and Marek Petrik. A practical method for solving contextual bandit problems using decision trees. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*, 2017.

[26] Ian Osband and Benjamin Van Roy. Bootstrapped Thompson sampling and deep exploration. *arXiv preprint arXiv:1507.00300*, 2015.

[27] S Pandey, D Chakrabarti, and D Agarwal. Multi-armed bandit problems with dependent arms. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 721–728, New York, NY, USA, 2007. ACM.

[28] Carlos Riquelme, George Tucker, and Jasper Snoek. Deep Bayesian bandits showdown: An empirical comparison of Bayesian deep networks for Thompson sampling. *arXiv preprint arXiv:1802.09127*, 2018.

[29] Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

[30] Liang Tang, Yexi Jiang, Lei Li, Chunqiu Zeng, and Tao Li. Personalized recommendation via parameter-free contextual bandits. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 323–332. ACM, 2015.

[31] Sharan Vaswani, Branislav Kveton, Zheng Wen, Anup Rao, Mark Schmidt, and Yasin Abbasi-Yadkori. New insights into bootstrapping for bandits. *CoRR*, abs/1805.09793, 2018. URL http://arxiv.org/abs/1805.09793.

[32] Chih-chun Wang, Sanjeev R Kulkarni, and H Vincent Poor. Bandit problems with side observations. *IEEE Transactions on Automatic Control*, 50:338–355, 2005.

[33] Lijun Zhang, Tianbao Yang, Rong Jin, Yichi Xiao, and Zhi-Hua Zhou. Online stochastic linear optimization under one-bit feedback. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 392–401, 2016.

A Upper Bound on the Inverse Probability of Being Optimistic

Our main result is presented in Theorem 2. Roughly speaking, we want to bound

$$W = \int_{x=0}^{\mathbb{E}[X]} \frac{\mathbb{P}\left(X = x\right)}{\mathbb{P}\left(Y \geq \mathbb{E}\left[X\right] \mid X = x\right)} \, \mathrm{d}x$$

from above. From Lemma 2,

$$\mathbb{P}\left(X = x\right) = O(\exp[-2(x - \mathbb{E}\left[X\right])^2]) \, .$$

This result is derived based on tail inequalities for sub-Gaussian random variables. From Lemmas 3 and 4,

$$\mathbb{P}\left(Y \geq \mathbb{E}\left[X\right] \mid X = x\right) = \Omega(\exp[-b(x - \mathbb{E}\left[X\right])^2]) \, ,$$

where $b < 2$ is a tunable parameter. This result relies on the properties of the sampling distribution. Now we combine both results and get an easy to solve integral

$$W \leq \int_{x=0}^{\mathbb{E}[X]} \exp[-(2-b)(x - \mathbb{E}\left[X\right])^2] \, \mathrm{d}x \, .$$

Most of the technical challenges in the proof are due to the fact that we deal with binomial random variables, which behave like normal variables only in very special cases.

Theorem 2. *Let* $m = (2a+1)n$ *and* $b = \dfrac{2a+1}{a(a+1)} < 2$. *Then*

$$W = \sum_{x=0}^{n} B(x; n, p) \left[\sum_{y=\lceil (a+p)n \rceil}^{m} B\left(y; m, \frac{an + x}{m}\right) \right]^{-1} \leq \frac{2e^2\sqrt{2a+1}}{\sqrt{2\pi}} \exp\left[\frac{8b}{2-b}\right] \left(1 + \sqrt{\frac{2\pi}{4-2b}}\right) .$$

Proof. First, we apply the upper bound from Lemma 2 for

$$f(x) = \left[\sum_{y=\lceil (a+p)n \rceil}^{m} B\left(y; m, \frac{an + x}{m}\right) \right]^{-1} .$$

Note that this function decreases in x, as required by Lemma 2, because the probability of observing at least $\lceil (a+p)n \rceil$ ones increases with x, for any fixed $\lceil (a+p)n \rceil$. The resulting upper bound is

$$W \leq \sum_{i=0}^{i_0 - 1} \exp[-2i^2] \left[\sum_{y=\lceil (a+p)n \rceil}^{m} B\left(y; m, \frac{(a+p)n - (i+1)\sqrt{n}}{m}\right) \right]^{-1} + \exp[-2i_0^2] \left[\sum_{y=\lceil (a+p)n \rceil}^{m} B\left(y; m, \frac{an}{m}\right) \right]^{-1} ,$$

where i_0 is the smallest integer such that $(i_0 + 1)\sqrt{n} \geq pn$, as defined in Lemma 2.

Second, we bound both above reciprocals using Lemma 3. The first term is bounded for $x = pn - (i+1)\sqrt{n}$ as

$$\left[\sum_{y=\lceil (a+p)n \rceil}^{m} B\left(y; m, \frac{(a+p)n - (i+1)\sqrt{n}}{m}\right) \right]^{-1} \leq \frac{e^2\sqrt{2a+1}}{\sqrt{2\pi}} \exp[b(i+2)^2] \, .$$

The second term is bounded for $x = 0$ as

$$\left[\sum_{y=\lceil (a+p)n \rceil}^{m} B\left(y; m, \frac{an}{m}\right) \right]^{-1} \leq \frac{e^2\sqrt{2a+1}}{\sqrt{2\pi}} \exp\left[b\frac{(pn + \sqrt{n})^2}{n}\right] \leq \frac{e^2\sqrt{2a+1}}{\sqrt{2\pi}} \exp[b(i_0+2)^2] \, ,$$

where the last inequality is from the definition of i_0. Then we chain the above three inequalities and get

$$W \leq \frac{e^2\sqrt{2a+1}}{\sqrt{2\pi}} \sum_{i=0}^{i_0} \exp[-2i^2 + b(i+2)^2]\,.$$

Now note that

$$2i^2 - b(i+2)^2 = (2-b)\left(i^2 - \frac{4b}{2-b} + \frac{4b^2}{(2-b)^2}\right) - \frac{4b^2}{2-b} - 4b = (2-b)\left(i - \frac{2b}{2-b}\right)^2 - \frac{8b}{2-b}\,.$$

It follows that

$$\begin{aligned}
W &\leq \frac{e^2\sqrt{2a+1}}{\sqrt{2\pi}} \sum_{i=0}^{i_0} \exp\left[-(2-b)\left(i - \frac{2b}{2-b}\right)^2 + \frac{8b}{2-b}\right] \\
&\leq \frac{2e^2\sqrt{2a+1}}{\sqrt{2\pi}} \exp\left[\frac{8b}{2-b}\right] \sum_{i=0}^{\infty} \exp\left[-(2-b)i^2\right] \\
&\leq \frac{2e^2\sqrt{2a+1}}{\sqrt{2\pi}} \exp\left[\frac{8b}{2-b}\right] \left(1 + \int_{u=0}^{\infty} \exp\left[-\frac{u^2}{\frac{2}{4-2b}}\right] \mathrm{d}u\right) \\
&\leq \frac{2e^2\sqrt{2a+1}}{\sqrt{2\pi}} \exp\left[\frac{8b}{2-b}\right] \left(1 + \sqrt{\frac{2\pi}{4-2b}}\right)\,.
\end{aligned}$$

This concludes our proof. ■

Lemma 2. *Let $f(x) \geq 0$ be a decreasing function of x and i_0 be the smallest integer such that $(i_0+1)\sqrt{n} \geq pn$. Then*

$$\sum_{x=0}^{n} B(x;n,p) f(x) \leq \sum_{i=0}^{i_0-1} \exp[-2i^2] f(pn - (i+1)\sqrt{n}) + \exp[-2i_0^2] f(0)\,.$$

Proof. Let

$$\mathcal{X}_i = \begin{cases} \left(\max\left\{pn - \sqrt{n},\, 0\right\},\, n\right], & i = 0\,; \\ \left(\max\left\{pn - (i+1)\sqrt{n},\, 0\right\},\, pn - i\sqrt{n}\right], & i > 0\,; \end{cases}$$

for $i \in [i_0] \cup \{0\}$. Then $\{\mathcal{X}_i\}_{i=0}^{i_0}$ is a partition of $[0,n]$. Based on this observation,

$$\begin{aligned}
\sum_{x=0}^{n} B(x;n,p) f(x) &= \sum_{i=0}^{i_0} \sum_{x=0}^{n} \mathbb{1}\{x \in \mathcal{X}_i\}\, B(x;n,p) f(x) \\
&\leq \sum_{i=0}^{i_0-1} f(pn - (i+1)\sqrt{n}) \sum_{x=0}^{n} \mathbb{1}\{x \in \mathcal{X}_i\}\, B(x;n,p) + f(0) \sum_{x=0}^{n} \mathbb{1}\{x \in \mathcal{X}_{i_0}\}\, B(x;n,p)\,,
\end{aligned}$$

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

$$\sum_{x=0}^{n} \mathbb{1}\{x \in \mathcal{X}_i\}\, B(x;n,p) \leq \mathbb{P}\left(X \leq pn - i\sqrt{n} \,\middle|\, X \sim B(n,p)\right) \leq \exp[-2i^2]\,.$$

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

Lemma 3. *Let $x \in [0, pn]$, $m = (2a+1)n$, and $b = \dfrac{2a+1}{a(a+1)}$. Then for any integer $n > 0$,*

$$\sum_{y=\lceil (a+p)n \rceil}^{m} B\left(y; m, \frac{an+x}{m}\right) \geq \frac{\sqrt{2\pi}}{e^2\sqrt{2a+1}} \exp\left[-b\frac{(pn+\sqrt{n}-x)^2}{n}\right].$$

Proof. By Lemma 4,

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

Now note that

$$\frac{y(m-y)}{m} \leq \frac{1}{m}\frac{m^2}{4} = \frac{(2a+1)n}{4}.$$

Moreover, since $x \in [0, n]$,

$$m\frac{an+x}{m}\frac{(a+1)n-x}{m} \geq m\frac{an}{m}\frac{(a+1)n}{m} = \frac{a(a+1)n}{2a+1} = \frac{n}{b},$$

where b is defined in the claim of this lemma. Now we combine the above three inequalities and have

$$B\left(y; m, \frac{an+x}{m}\right) \geq \frac{2\sqrt{2\pi}}{e^2\sqrt{2a+1}}\frac{1}{\sqrt{n}} \exp\left[-b\frac{(y-an-x)^2}{n}\right],$$

Finally, note the following two facts. First, the above lower bound decreases in y when $y \geq (a+p)n$ and $x \leq pn$. Second, by the pigeonhole principle, there exist at least $\lfloor\sqrt{n}\rfloor$ integers between $(a+p)n$ and $(a+p)n+\sqrt{n}$, starting with $\lceil (a+p)n \rceil$. These observations lead to a trivial lower bound

$$\begin{aligned}\sum_{y=\lceil (a+p)n \rceil}^{m} B\left(y; m, \frac{an+x}{m}\right) &\geq \frac{\lfloor\sqrt{n}\rfloor}{\sqrt{n}}\frac{2\sqrt{2\pi}}{e^2\sqrt{2a+1}} \exp\left[-b\frac{(pn+\sqrt{n}-x)^2}{n}\right] \\ &\geq \frac{\sqrt{2\pi}}{e^2\sqrt{2a+1}} \exp\left[-b\frac{(pn+\sqrt{n}-x)^2}{n}\right].\end{aligned}$$

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

Lemma 4. *For any binomial probability,*

$$B(x; n, p) \geq \frac{\sqrt{2\pi}}{e^2}\sqrt{\frac{n}{x(n-x)}} \exp\left[-\frac{(x-pn)^2}{p(1-p)n}\right].$$

Proof. By Stirling's approximation, for any integer $k \geq 0$,

$$\sqrt{2\pi}k^{k+\frac{1}{2}}e^{-k} \leq k! \leq ek^{k+\frac{1}{2}}e^{-k}.$$

Therefore, any binomial probability can be bounded from below as

$$B(x; n, p) = \frac{n!}{x!(n-x)!}p^x q^{n-x} \geq \frac{\sqrt{2\pi}}{e^2}\sqrt{\frac{n}{x(n-x)}}\left(\frac{pn}{x}\right)^x\left(\frac{qn}{n-x}\right)^{n-x},$$

where $q = 1-p$. Let

$$d(p_1, p_2) = p_1 \log\frac{p_1}{p_2} + (1-p_1)\log\frac{1-p_1}{1-p_2}$$

Algorithm 3 General randomized exploration.

1: **for all** $i \in [K]$ **do** ▷ Initialization
2: $T_{i,0} \leftarrow 0$, $\mathcal{H}_{i,0} \leftarrow ()$
3: **for** $t = 1, \ldots, n$ **do**
4: **for all** $i \in [K]$ **do** ▷ Choose the arm
5: Draw $\theta_{i,t} \sim p(\mathcal{H}_{i,T_{i,t-1}})$
6: $I_t \leftarrow \arg\max_{i \in [K]} \theta_{i,t}$ ▷ Pull the arm
7: Pull arm I_t and observe $V_{i,T_{i,t-1}+1}$
8: **for all** $i \in [K]$ **do** ▷ Update statistics
9: **if** $i = I_t$ **then**
10: $T_{i,t} \leftarrow T_{i,t-1} + 1$
11: $\mathcal{H}_{i,T_{i,t}} \leftarrow \mathcal{H}_{i,T_{i,t-1}} \oplus (V_{i,T_{i,t-1}+1})$
12: **else**
13: $T_{i,t} \leftarrow T_{i,t-1}$

be the KL divergence between Bernoulli random variables with means p_1 and p_2. Then

$$\begin{aligned}
\left(\frac{pn}{x}\right)^x \left(\frac{qn}{n-x}\right)^{n-x} &= \exp\left[x \log\left(\frac{pn}{x}\right) + (n-x)\log\left(\frac{qn}{n-x}\right)\right] \\
&= \exp\left[-n\left(\frac{x}{n}\log\left(\frac{x}{pn}\right) + \frac{n-x}{n}\log\left(\frac{n-x}{qn}\right)\right)\right] \\
&= \exp\left[-nd\left(\frac{x}{n}, p\right)\right] \\
&\geq \exp\left[-\frac{(x-pn)^2}{p(1-p)n}\right],
\end{aligned}$$

where the inequality is from $d(p_1, p_2) \leq \dfrac{(p_1 - p_2)^2}{p_2(1-p_2)}$. Finally, we chain all inequalities and get our claim. ■

B General Randomized Exploration

In this section, we bound the regret of a general bandit algorithm that explores by sampling conditioned on its history. The regret bound of `Giro` in Section 5 is an instance of this result, where we bound sampling-specific artifacts.

Our analysis is for a multi-armed bandit with K arms. The reward distribution of arm i is P_i and its mean is μ_i. Without loss of generality, let arm 1 be optimal, and thus $\mu_1 > \max_{i \neq 1} \mu_i$. The gap of arm i is $\Delta_i = \mu_1 - \mu_i$. As described in Section 4.4 of Lattimore and Szepesvári [22], and without loss of generality, we assume that the rewards of arms are pregenerated and then used when the arm is pulled. The s-th reward of arm i is denoted by $V_{i,s}$ and the pregenerated rewards of arm i in n rounds are $(V_{i,s})_{s=1}^n$.

Our algorithm is presented in Algorithm 3. The number of pulls of arm i after the first t rounds is denoted by $T_{i,t}$, where $T_{i,t} = \sum_{\ell=1}^t \mathbb{1}\{I_\ell = i\}$. The history of arm i after s pulls is $\mathcal{H}_{i,s} = (V_{i,\ell})_{\ell=1}^s$. We define $\mathcal{H}_{i,0} = ()$. A t-round history of the algorithm is $\mathcal{F}_t = (\mathcal{H}_{i,T_{i,t}})_{i=1}^K$, a concatenation of the histories of all arms by the end of round t. The sampling distribution p in line 5 is a function of the history of the arm. For $s \in [n] \cup \{0\}$, let

$$Q_{i,s}(\tau) = \mathbb{P}\left(\theta \geq \tau \,|\, \theta \sim p(\mathcal{H}_{i,s}), \mathcal{H}_{i,s}\right) \tag{5}$$

be the tail probability that a sampled value θ conditioned on history $\mathcal{H}_{i,s}$ is at least τ for some tunable parameter τ. If $p(\mathcal{H}_{i,s})$ concentrates at μ_i as s increases, τ of suboptimal arm i would be chosen from (μ_i, μ_1). More generally, if $p(\mathcal{H}_{i,s})$ concentrates at μ_i', with $\mu_i' \leq \mu_1'$, τ of suboptimal arm i would be chosen from (μ_i', μ_1'). Our regret bound is stated below.

Theorem 3. *For any $\tau_i \in \mathbb{R}$, the expected n-round regret of Algorithm 3 is*

$$R(n) \le \sum_{i=2}^{K} \Delta_i (a_i + b_i)\,,$$

where

$$a_i = \sum_{s=0}^{n-1} \mathbb{E}\left[\min\left\{\frac{1}{Q_{1,s}(\tau_i)} - 1, n\right\}\right], \qquad b_i = \sum_{s=0}^{n-1} \mathbb{P}\left(Q_{i,s}(\tau_i) > 1/n\right) + 1\,.$$

Proof. The proof follows the argument of Agrawal and Goyal [3]. Since arm 1 is optimal, the regret can be written as

$$R(n) = \sum_{i=2}^{K} \Delta_i \mathbb{E}\left[T_{i,n}\right]\,.$$

In the rest of the proof, we bound $\mathbb{E}\left[T_{i,n}\right]$ for each suboptimal arm i. Fix arm $i > 1$. Let $E_{i,t} = \{\theta_{i,t} \le \tau_i\}$ and $\bar{E}_{i,t}$ be the complement of $E_{i,t}$. Then $\mathbb{E}\left[T_{i,n}\right]$ can be decomposed as

$$\mathbb{E}\left[T_{i,n}\right] = \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{1}\{I_t = i\}\right] = \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{1}\{I_t = i, E_{i,t} \text{ occurs}\}\right] + \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{1}\left\{I_t = i, \bar{E}_{i,t} \text{ occurs}\right\}\right]. \tag{6}$$

Term b_i in the Upper Bound

We start with the second term in (6), which corresponds to b_i in our claim. This term can be tightly bounded based on the observation that event $\bar{E}_{t,i}$ is unlikely when $T_{i,t}$ is "large". Let $\mathcal{T} = \{t \in [n] : Q_{i,s}(\tau_i) > 1/n\}$. Then

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{n} \mathbb{1}\left\{I_t = i, \bar{E}_{i,t} \text{ occurs}\right\}\right] &\le \mathbb{E}\left[\sum_{t\in\mathcal{T}} \mathbb{1}\{I_t = i\}\right] + \mathbb{E}\left[\sum_{t\notin\mathcal{T}} \mathbb{1}\left\{\bar{E}_{i,t}\right\}\right] \\
&\le \mathbb{E}\left[\sum_{s=0}^{n-1} \mathbb{1}\{Q_{i,s}(\tau_i) > 1/n\}\right] + \mathbb{E}\left[\sum_{t\notin\mathcal{T}} \frac{1}{n}\right] \\
&\le \sum_{s=0}^{n-1} \mathbb{P}\left(Q_{i,s}(\tau_i) > 1/n\right) + 1\,.
\end{aligned}$$

Term a_i in the Upper Bound

Now we focus on the first term in (6), which corresponds to a_i in our claim. Without loss of generality, we assume that Algorithm 3 is implemented as follows. When arm 1 is pulled for the s-th time, the algorithm generates an infinite i.i.d. sequence $(\theta_\ell^{(s)})_\ell \sim Q_{1,s}$. Then, instead of sampling $\theta_{1,t} \sim Q_{1,s}$ in round t when $T_{1,t-1} = s$, $\theta_{1,t}$ is set to $\theta_t^{(s)}$. Let $M = \{t \in [n] : \max_{j>1} \theta_{j,t} \le \tau_i\}$ be round indices where the values of all suboptimal arms are at most τ_i and

$$A_s = \left\{t \in M : \theta_t^{(s)} \le \tau_i,\, T_{1,t-1} = s\right\}$$

be its subset where the value of arm 1 is at most τ_i and the arm was pulled s times before. Then

$$\sum_{t=1}^{n} \mathbb{1}\{I_t = i,\, E_{i,t} \text{ occurs}\} \le \sum_{t=1}^{n} \mathbb{1}\left\{\max_j \theta_{j,t} \le \tau_i\right\} = \sum_{s=0}^{n-1} \underbrace{\sum_{t=1}^{n} \mathbb{1}\left\{\max_j \theta_{j,t} \le \tau_i,\, T_{1,t-1} = s\right\}}_{|A_s|}\,.$$

In the next step, we bound $|A_s|$. Let

$$\Lambda_s = \min\left\{t \in M : \theta_t^{(s)} > \tau_i,\, T_{1,t-1} \geq s\right\}$$

be the index of the first round in M where the value of arm 1 is larger than τ_i and the arm was pulled at least s times before. If such Λ_s does not exist, we assume that $\Lambda_s = n$. Let

$$B_s = \left\{t \in M \cap [\Lambda_s] : \theta_t^{(s)} \leq \tau_i,\, T_{1,t-1} \geq s\right\}$$

be a subset of $M \cap [\Lambda_s]$ where the value of arm 1 is at most τ_i and the arm was pulled at least s times before.

We claim that $A_s \subseteq B_s$. By contradiction, suppose that there exists $t \in A_s$ such that $t \notin B_s$. Then it must be true that $\Lambda_s < t$, from the definitions of A_s and B_s. From the definition of Λ_s, we know that arm 1 was pulled in round Λ_s, after it was pulled at least s times before. Therefore, it cannot be true that $T_{1,t-1} = s$, and thus $t \notin A_s$. Therefore, $A_s \subseteq B_s$ and $|A_s| \leq |B_s|$. In the next step, we bound $|B_s|$ in expectation.

Let $\mathcal{F}_t = \sigma(\mathcal{H}_{1,T_{1,t}}, \ldots, \mathcal{H}_{K,T_{K,t}}, I_1, \ldots, I_t)$ be the σ-algebra generated by arm histories and pulled arms by the end of round t, for $t \in [n] \cup \{0\}$. Let $P_s = \min\{t \in [n] : T_{1,t-1} = s\}$ be the index of the first round where arm 1 was pulled s times before. If such P_s does not exist, we assume that $P_s = n + 1$. Note that P_s is a stopping time with respect to filtration $(\mathcal{F}_t)_t$. Hence, $\mathcal{G}_s = \mathcal{F}_{P_s - 1}$ is well-defined and thanks to $|A_s| \leq n$, we have

$$\mathbb{E}\left[|A_s|\right] = \mathbb{E}\left[\min\left\{\mathbb{E}\left[|A_s| \,|\, \mathcal{G}_s\right], n\right\}\right] \leq \mathbb{E}\left[\min\left\{\mathbb{E}\left[|B_s| \,|\, \mathcal{G}_s\right], n\right\}\right] .$$

We claim that $\mathbb{E}\left[|B_s| \,|\, \mathcal{G}_s\right] \leq 1/Q_{1,s}(\tau_i) - 1$. First, note that $|B_s|$ can be rewritten as

$$|B_s| = \sum_{t=P_s}^{\Lambda_s} \epsilon_t \rho_t \,,$$

where $\epsilon_t = \mathbb{1}\{\max_{j>1} \theta_{j,t} \leq \tau_i\}$ controls which values $\rho_t = \mathbb{1}\left\{\theta_t^{(s)} \leq \tau_i\right\}$ contribute to the sum. In the language of Doob [12, Chapter VII], $(\sum_{t=P_s}^{t'} \epsilon_t \rho_t)_{t'}$ is an *optional skipping process*, where ϵ_t and ρ_t are independent given their joint past. By the optional skipping theorem in Doob [12, Theorem 2.3 in Chapter VII], the independence of ρ_t is preserved by the skipping process and thus, from the definition of Λ_s, $|B_s|$ has the same distribution as the number of failed independent draws from $\mathrm{Ber}(Q_{1,s})$ until the first success, capped at $n - P_s$. As is well known, the expected value of this quantity, without the cap, is bounded by $1/Q_{1,s} - 1$.

Finally, we chain all inequalities and get

$$\mathbb{E}\left[\sum_{t=1}^{n} \mathbb{1}\{I_t = i, E_{i,t} \text{ occurs}\}\right] \leq \sum_{s=0}^{n-1} \mathbb{E}\left[\min\left\{1/Q_{1,s}(\tau_i) - 1, n\right\}\right] .$$

This concludes our proof. ■