

# Batch Ensemble for Variance Dependent Regret in Stochastic Bandits

Asaf Cassel<sup>1</sup>, Orin Levy<sup>1</sup>, Yishay Mansour<sup>1,2</sup>

<sup>1</sup>Blavatnik School of Computer Science, Tel Aviv University

<sup>2</sup>Google Research, Tel-Aviv

acassel@mail.tau.ac.il, orinlevy@mail.tau.ac.il, mansour.yishay@gmail.com

## Abstract

Efficiently trading off exploration and exploitation is one of the key challenges in online Reinforcement Learning (RL). Most works achieve this by carefully estimating the model uncertainty and following the so-called optimistic model. Inspired by practical ensemble methods, in this work we propose a simple and novel batch ensemble scheme that provably achieves near-optimal regret for stochastic Multi-Armed Bandits (MAB). Crucially, our algorithm has just a single parameter, namely the number of batches, and its value does not depend on distributional properties such as the scale and variance of the losses. We complement our theoretical results by demonstrating the effectiveness of our algorithm on synthetic benchmarks.

## 1 Introduction

Multi-Armed Bandits is a classic framework for sequential decision-making under uncertainty. In this setting, an agent repeatedly interacts with an environment by choosing from a set of  $K$  actions (arms) and subsequently observing a loss signal associated with their choice. The loss signal associated with each arm is a sequence of i.i.d random variables whose mean is unknown to the agent. The agent’s goal is to minimize their cumulative loss, which presents a classic exploration vs. exploitation dilemma, i.e., whether to exploit the current knowledge of the losses or to further explore seemingly sub-optimal actions that may turn out to be better. This trade-off is measured via a notion termed regret, which is the difference between the agent’s performance and that of an oracle who knows the best arm and chooses it throughout the interaction.

Learning algorithms for this setting were extensively studied, and date back to Robbins’ paper (Robbins 1952). In particular, (Lai and Robbins 1985) established that the regret in this problem is lower bounded by  $\Omega(\log T)$ , and there exist learning algorithms that achieve this regret by maximizing a confidence bound modification of the empirical mean. A non-asymptotic analysis was later provided by (Auer, Cesa-Bianchi, and Fischer 2002). Subsequent works obtain bounds that depend on subtler properties of the arms by constructing more elaborate confidence bounds (see e.g.,

Maillard, Munos, and Stoltz (2011) for empirical confidence bounds or UCB-KL).

A commonality of nearly all past works is that they explicitly encode the distributional assumptions on the arms into the algorithm. For example, UCB (Auer, Cesa-Bianchi, and Fischer 2002) builds confidence bounds tailored to distributions bounded in  $[0, 1]$ , UCB-V (Audibert, Munos, and Szepesvári 2009) refines these confidence bounds by incorporating a variance estimate, and KL-UCB (Maillard, Munos, and Stoltz 2011) explicitly uses the KL-divergence to estimate uncertainty and often assume a parametric class to reduce computational complexity. This explicit encoding can (1) be disadvantageous when the arms are misspecified; (2) require delicate parameter tuning; or (3) require prior knowledge of the distributions.

In this work, we show that constructing standard mean estimators from a simple batching scheme and combining them using a min operator, yields an optimistic mean estimator. Choosing greedily with respect to this estimator yields regret bounds that depend on the true concentration properties of the arm distributions.

**Our Contributions.** Our main contribution is a simple MAB algorithm, that does not need tuning of parameters and has low computational overhead. We show that for the Bernoulli r.v. our algorithm achieves an instance-dependent regret that depends on the variances of the arms. We show that our scheme extends to many other distributions, including, distributions that are either symmetric around the mean, have bounded support, or lower bounded variance. Crucially, this adaptation is purely in analysis and does not require modifying the algorithm.

Our scheme easily adapts to a distributed environment. Concretely, instead of explicitly constructing an optimistic mean estimator for each arm, our algorithm may be viewed as separate (distributed) naive bandit algorithms each receiving separate samples, computing the means of each arm, and outputting the best empirical arm together with its empirical mean. The final decision is made by following the decision of the bandit algorithm with the best empirical mean. This interpretation corresponds to practical methods such as (Osband et al. 2016; Tennenholtz et al. 2022), which use ensembles to encourage exploration. We complement our theoretical findings by running experiments on synthetic bench-

marks, showing that our scheme achieves low regret compared to alternative algorithms.

**Related work.** Similar ideas of ensemble and bootstrapping methods have previously been studied. (Ash et al. 2021) introduced anti-concentrated confidence bounds for efficiently approximating the elliptical bonus, using an ensemble of regressors. (Osband et al. 2016) applied bootstrapped DQN in the Arcade Learning Environment and obtained improved learning speed and cumulative performance across most games. (Osband, Van Roy, and Wen 2016) Present randomized least-squares value iteration (RLSVI) - an algorithm designed to explore and generalize via linearly parameterized value functions. Their results established that randomized value functions are a useful tool for efficient exploration along with effective generalization. (Peer et al. 2021) present the Ensemble Bootstrapped Q-Learning (EBQL) algorithm, a natural extension of Double-Q-learning to ensembles that is bias-reduced. They analyze it both theoretically and empirically.

There are bootstrapping methods that add pseudo-rewards, sample the pseudo-rewards, and then run the MAB on the perturbed sequence. These include GIRO (Kveton et al. 2019b) and PHE (Kveton et al. 2019a), which have an instance-dependent regret bound for Bernoulli rewards (but they are not variance-dependent bounds) and Reboot (Wang et al. 2020) which handles Gaussian rewards.

Sub-sampling techniques, combined with a dueling approach, have been first proposed in BESA (Baransi, Maillard, and Mannor 2014) for two arms, and extended in RB-SDA (Baudry, Kaufmann, and Maillard 2020) and SSMC (Chan 2020) for one-parameter exponential distributions.

The most related work to ours is MARS (Moravej Khorasani and Weyer 2023), where they generate optimistic estimates by sampling multiple random subsets and taking the maximum average reward. They show a regret bound for distributions that are continuous and symmetric around the mean. Our methodology can handle non-symmetric distributions and has better computational complexity.

## 2 Preliminaries

**Problem setup.** In a stochastic  $K$ -armed bandit, each arm  $a \in [K]$  is associated with a loss sequence  $\ell_{n,a}$ , ( $n \geq 1$ ) of i.i.d Bernoulli random variables with parameter  $\mu_a \in [0, 1]$ . (Note that  $\mu_a = \mathbb{E}[\ell_{n,a}]$ , for any  $n$ .) Let  $a_\star \in \arg \min_{a \in [K]} \mu_a$  be an optimal action and  $\mu_\star = \mu_{a_\star} = \min_{a \in [K]} \mu_a$  be the optimal value. At each time step  $t = 1, 2, \dots$ , an agent interacts with the bandit by choosing an arm  $a_t \in [K]$  and subsequently observes the random loss  $\ell_t = \ell_{(n_t, a_t), (a_t)}$  where

$$n_{t,a} = \sum_{\tau=1}^t \mathbb{1}_{\{a_\tau = a\}} \quad (1)$$

is the number of times arm  $a \in [K]$  was played up to time  $t$  (inclusive). The agent does not know the problem parameters and must learn them on the fly. We quantify its perfor-

mance via the (pseudo) regret

$$\text{regret}_t = \sum_{\tau=1}^t [\mu_{a_\tau} - \mu_{a_\star}],$$

which measures the performance gap between the agent and the optimal policy that plays an optimal arm at each step. We note that the restriction to Bernoulli arms is mostly for simplicity and we discuss extensions in Section 4.

**Deviation bounds.** In what follows, we require the following fundamental properties. Let  $X_n \in [0, 1]$ , ( $n \geq 1$ ) be i.i.d random variables with  $\mathbb{E}X_n = \mu$  and  $\text{Var}(X_n) = \sigma^2$ . The following is a standard Bernstein inequality for bounded random variables.

**Lemma 1.** Let  $\bar{\mu} = \frac{1}{n} \sum_{n'=1}^n X_{n'}$ . With probability at least  $1 - \delta$

$$\bar{\mu} \geq \mu - \frac{2}{n} \log \frac{1}{\delta} - \sqrt{\frac{\sigma^2}{n} \log \frac{1}{\delta}}.$$

Next, recall that if  $X_n \sim \text{Ber}(\mu)$  then  $\sum_{n'=1}^n X_{n'}$  is Binomial with parameters  $n, \mu$ .

**Lemma 2 (Wiklund (2023), Corollary 1).** If  $\mu \leq 1 - 1/n$  then

$$\Pr(\text{Bin}(n, \mu) \leq n\mu) \geq \frac{1}{4}.$$

## 3 Algorithm and Main Results

At a high level, our Batch Ensemble algorithm works as follows. It splits the samples of each arm into multiple batches. For each batch, it computes an (almost) empirical average. It then computes the minimum of those estimates, which is an optimistic estimator (recall that we are dealing with losses). It then plays the action with the lowest estimate.

In Section 3.1 we analyze the properties of our estimator. The crucial and non-standard property is that with high probability it is an underestimate of the true expected value. The second property is a somewhat standard concentration bound using Bernstein inequality (Lemma 1).

In Section 3.2 we describe the Batch Ensemble algorithm, and state its performance guarantees (Theorem 5). (The proof is deferred to Section 5.) We then discuss a few implementation details, such as a distributed computing view of the algorithm, and the ability to have an *any time* guarantee.

### 3.1 An Optimistic Mean Estimator

Suppose we have observed  $n \geq 0$  samples of an arm  $a \in [K]$ , i.e.,  $\ell_{n',a}, n' \in [n]$ . We build the following mean estimator. First, let,  $l \geq 1$  be a batch number to be determined later. Next, we split the  $n$  samples of arm  $a$  into  $l$  (near-)equal batches<sup>1</sup>

$$\tau_{n,a,l'} = \{n' : n' = l' + i \cdot l \leq n, i \in \mathbb{Z}_{\geq 0}\}, l' \in [l].$$

<sup>1</sup>We used a round-robin schedule to create the batches, but any non-adaptive scheme would work.

Our batch ensemble estimator is

$$\hat{\mu}_{n,a} = \min_{l' \in [l]} \hat{\mu}_{n,a,l'}, \text{ where } \hat{\mu}_{n,a,l'} = \sum_{n' \in \tau_{n,a,l'}} \frac{\ell_{n',a}}{|\tau_{n,a,l'}| + 2}, \quad (2)$$

with the convention that an empty sum is equal to 0. The purpose of adding 2 in the denominator will be made apparent in the proof of the following result, which establishes the optimistic property of our estimator.

**Lemma 3.** *Let  $\delta \in [0, 1]$  and  $l \geq 1$ . Then for any  $n \geq 0$*

$$\Pr(\hat{\mu}_{n,a} \leq \mu_a) \geq 1 - e^{-2l/7}.$$

Notice that choosing  $l = (7/2) \log(1/\delta)$  in Lemma 3 gives the standard high probability optimistic guarantee.

**Proof.** Fix  $l' \in [l]$  and  $a \in [K]$ . We show that  $\Pr(\hat{\mu}_{n,a,l'} \leq \mu_a) \geq 1/4$  for all  $n \geq 0$ . Let  $\tau = |\tau_{n,a,l'}|$  be the number of samples in the  $l'$ -th batch when arm  $a$  has  $n$  samples. If  $\tau = 0$ , the claim holds trivially. If  $\tau = 1$  then for  $\mu_a \geq 1/3$  the claim holds since  $\hat{\mu}_{n,a,l'} \leq 1/3$  and for  $\mu_a < 1/3$  it holds with probability  $1 - \mu_a \geq 2/3$  because  $\hat{\mu}_{n,a,l'} \sim \text{Ber}(\mu_a)$ . Now, assume that  $\tau \geq 2$ . If  $\mu_a \geq 1 - 1/\tau$  then we have that

$$\hat{\mu}_{n,a,l'} \leq \frac{\tau}{\tau + 2} = 1 - \frac{2}{\tau + 2} \leq 1 - \frac{1}{\tau} \leq \mu_a,$$

where the second to last inequality used that  $\tau \geq 2$ . If  $\mu_a \leq 1 - 1/\tau$  then by Lemma 2

$$\Pr(\hat{\mu}_{n,a,l'} \leq \mu_a) \geq \Pr\left(\sum_{n' \in \tau_{n,a,l'}} \ell_{n',a} \leq \tau \mu_a\right) \geq \frac{1}{4}.$$

Now, since the  $\hat{\mu}_{n,a,l'}$  are composed of different variables, they are jointly independent, thus we have

$$\begin{aligned} \Pr(\hat{\mu}_{n,a} > \mu_a) &= \Pr(\hat{\mu}_{n,a,l'} > \mu_a, \forall l' \in [l]) \\ &= \prod_{l' \in [l]} \Pr(\hat{\mu}_{n,a,l'} > \mu_a) \\ &\leq (1 - 1/4)^l \leq e^{-2l/7}. \quad \blacksquare \end{aligned}$$

The following result describes the concentration of our mean estimator. The proof is a straightforward application of Lemma 1.

**Lemma 4.** *Let  $\delta \in [0, 1]$  and  $l \geq 1$ . With probability at least  $1 - \delta$ , simultaneously for all  $n \in [T]$*

$$\mu_a - \hat{\mu}_{n,a} \leq \frac{2}{(n/l) + 1} \log \frac{3T}{\delta} + \sqrt{\frac{\sigma_a^2}{(n/l) + 1} \log \frac{T}{\delta}}.$$

**Proof.** We use Lemma 1 together with a union bound to get that with probability at least  $1 - \delta$ , simultaneously for all  $n \in [T], l' \in [l]$

$$\sum_{n' \in \tau_{n,a,l'}} \frac{\ell_{n',a}}{|\tau_{n,a,l'}|} \geq \mu_a - \frac{2}{|\tau_{n,a,l'}|} \log \frac{T}{\delta} - \sqrt{\frac{\sigma_a^2}{|\tau_{n,a,l'}|} \log \frac{T}{\delta}},$$

Recalling the definition of  $\hat{\mu}_{n,a,l'}$  in Eq. (2), we conclude that

$$\begin{aligned} \hat{\mu}_{n,a,l'} &= \frac{|\tau_{n,a,l'}|}{|\tau_{n,a,l'}| + 2} \sum_{n' \in \tau_{n,a,l'}} \frac{\ell_{n',a}}{|\tau_{n,a,l'}|} \\ &\geq \frac{|\tau_{n,a,l'}|}{|\tau_{n,a,l'}| + 2} \left[ \mu_a - \frac{2}{|\tau_{n,a,l'}|} \log \frac{T}{\delta} - \sqrt{\frac{\sigma_a^2}{|\tau_{n,a,l'}|} \log \frac{T}{\delta}} \right] \\ &\geq \mu_a - \frac{2}{|\tau_{n,a,l'}| + 2} \log \frac{3T}{\delta} - \sqrt{\frac{\sigma_a^2}{|\tau_{n,a,l'}| + 2} \log \frac{T}{\delta}} \\ &\geq \mu_a - \frac{2}{(n/l) + 1} \log \frac{3T}{\delta} - \sqrt{\frac{\sigma_a^2}{(n/l) + 1} \log \frac{T}{\delta}}, \end{aligned}$$

where the second inequality used that  $\mu_a \leq 1$  and the third that  $|\tau_{n,a,l'}| \geq (n/l) - 1$ .  $\blacksquare$

### 3.2 The Batch Ensemble Algorithm

We present the Batch Ensemble algorithm in Algorithm 1. The algorithm receives as input a sequence representing the number of batches to use at each time step, builds a mean estimator as described in Eq. (2), and chooses the arm with the most optimistic (i.e., minimal) estimate. A key distinction from MARS (Moravej Khorasani and Weyer 2023), is that we use a small number of *non-overlapping* batches.

---

#### Algorithm 1: Batch Ensemble for MAB

---

- 1: **input:** number of batches  $l_t$  for all  $t \geq 1$ .
- 2: **initialize:** pull counts  $n_{0,a} = 0$  for all  $a \in [K]$ .
- 3: **for** time step  $t = 1, 2, \dots$  **do**
- 4:   calculate  $\hat{\mu}_{n_{t-1},a}$  as in Eq. (2) with  $l_t$  batches and choose

$$a_t \in \arg \min_{a \in [K]} \hat{\mu}_{n_{t-1},a}. \quad (3)$$

- 5:   observe  $\ell_{(n_{t,a}),a}$  and update  $n_{t,a} = n_{t-1,a} + \mathbb{1}_{\{a_t=a\}}$ .
  - 6: **end for**
- 

The following is our main result, which bounds the regret of the above algorithm (proof in Section 5).

**Theorem 5.** *Suppose we run Algorithm 1 with a fixed number of batches with  $l = (7/2) \log(2T/\delta)$ . With probability at least  $1 - \delta$  the following regret bounds<sup>2</sup> hold simultaneously*

$$\begin{aligned} \text{regret}_T &\leq \sqrt{14T \min\left\{\mu_* K, \sum_{a \neq a_*} \sigma_a^2\right\} \log \frac{6TK}{\delta}} \\ &\quad + 11K \log^2 \frac{6TK}{\delta} \\ \text{regret}_T &\leq \frac{7}{2} \sum_{a \neq a_*} \left(\frac{\sigma_a^2}{\Delta_a} + 2\right) \log^2 \frac{6TK}{\delta} \end{aligned}$$

<sup>2</sup>The second inequality in the last bound holds for arm distributions in  $[0, 1]$  because they satisfy  $\sigma_a^2 \leq \mu_a = \mu_* + \Delta_a$ .

$$\leq \frac{7}{2} \sum_{a \neq a_*} \left( \frac{\mu_*}{\Delta_a} + 3 \right) \log^2 \frac{6TK}{\delta}.$$

**Dependence on the true arm distributions.** We note that our regret bounds depend on the variance due to our use of the Bernstein type inequality in Lemma 1. This choice was made in order to obtain a clear and intuitive result. A tighter bound can be achieved by using the tight Chernoff bound for the empirical mean of each arm, which for Bernoulli r.v depends on the KL divergence. This is true even if different arms come from different distributional families, e.g., some are normal and some are Bernoulli, and does not affect the algorithm or its parameter.

**A distributed view.** We presented Algorithm 1 in the familiar UCB-style index rule. However, a potentially more insightful perspective is to view each batch  $l' \in [l]$  as a separate bandit algorithm that at each round  $t \in [T]$  outputs a prediction  $a_{t,l'} \in \arg \min_{a \in [K]} \hat{\mu}_{(n_{t-1,a}), (a), (l')}$  and its estimate  $\hat{\mu}_{t,l'} = \hat{\mu}_{(n_{t-1,a}), (a_{t,l'}), (l')}$ , and the final decision is made by greedily choosing the best batch, i.e.,  $a_t = a_{t,l_t^*}$  where  $l_t^* \in \arg \min_{l' \in [l]} \hat{\mu}_{t,l'}$ . This decision rule is equivalent to Eq. (3). Importantly, we believe this view could help scale our approach to other problem settings such as MDPs where each batch would output a policy and its value prediction and the final policy is the one associated with the most optimistic batch. We leave this for future research.

We note that, unlike some distributed learning routines that aggregate decisions by averaging them to reduce uncertainty (noise), our approach selects a single “noisy” batch whose decision is followed. This is key to ensuring the optimistic property of our algorithm.

**An anytime expected regret algorithm.** One can always use the doubling trick to obtain an anytime algorithm. However, this leaves the dependence on the confidence level  $\delta$  that, realistically, depends on the time horizon  $T$ . To avoid this, we show that choosing the number of batches as  $l_t = 8 \log t$ , Algorithm 1 has expected regret  $(\mathbb{E} \text{regret}_t, t \geq 1)$  bounded similarly to Theorem 5 but with  $T$  replaced with  $t$  and the dependence on  $\delta$  removed. For details see the supplementary material.

## 4 Beyond Bernoulli Arms

Our results thus far focused on Bernoulli distributed arms. However, this is not explicitly encoded in our algorithm but rather in its analysis. In fact, our algorithm works without change for any arm distributions that satisfy properties akin to Lemmas 1 and 2.

Lemma 1 is a standard concentration bound for bounded random variables. If the random variables are unbounded, Lemma 1 can be replaced with the appropriate Chernoff bound. As long as the distributions are light-tailed (e.g., sub-Gaussian or sub-exponential), this will not change the regret bound significantly. We emphasize that the algorithm does not need to know the tail behavior and thus the bound may be tailored to the true distribution of each arm.

Lemma 2 is a type of anti-concentration result for sums of Bernoulli random variables. We conjecture that all bounded

random variables satisfy this property, but have been unable to prove this. For further details see the conjecture at the end of the section. In what follows, we describe several methods and conditions to satisfy the anti-concentration property for non-Bernoulli arms.

**Bernoulli-fication.** It is well known that arm distributions in  $[0, 1]$  can be converted into Bernoulli arms. To do this one replaces the observed losses of the algorithm  $\ell_{n,a}$  with samples  $\bar{\ell}_{n,a} \sim \text{Ber}(\ell_{n,a})$ . If the distribution is in  $[0, b]$ , one can first scale the losses by dividing with  $b$ . It is straightforward to verify that  $\mathbb{E} \ell_{n,a} = \mathbb{E} \bar{\ell}_{n,a}$ , and thus Theorem 5 holds but with  $\sigma_a$  replaced with the variance of  $\bar{\ell}_{n,a}$ . This does not impact the first-order regret bounds, but can significantly increase the variance-dependent (second-order) bounds (e.g., for deterministic arms).

**Scaled Bernoulli.** It is often the case that arm distributions are not evenly scaled. Most bandit algorithms such as UCB or UCB-V have a single scale parameter, which bounds the worst-case arm. Algorithm 1 does not need to know the scales in advance and automatically enjoys dependence on the true arm scales. To see this, consider arm distributions that are scaled Bernoulli variables with parameters  $b_a \geq \mu_a \geq 0$  such that

$$\ell_{n,a} = \begin{cases} b_a, & \text{w.p } \mu_a/b_a \\ 0, & \text{otherwise.} \end{cases}$$

Notice that if we scale the arms in the analysis, we can still use Lemma 2 to get the optimism claim in Lemma 3. As for concentration, scaling Lemma 1 replaces the  $2/n$  term with  $2b_a/n$ . Propagating this into the analysis of Theorem 5 would modify the bounds such that

$$\begin{aligned} \left[ \frac{\sigma_a^2}{\Delta_a} + 2 \right], \left[ \frac{\mu_*}{\Delta_a} + 3 \right] &\implies \left[ \frac{\sigma_a^2}{\Delta_a} + 2b_a \right], \left[ \frac{\mu_*}{\Delta_a} + 3b_a \right] \\ 11K \log^2 \frac{6TK}{\delta} &\implies 11 \sum_{a \neq a_*} b_a \log^2 \frac{6TK}{\delta}. \end{aligned}$$

Notice that we do not depend on the scale of the optimal arm, which could be meaningful when it is significantly larger than the scale of sub-optimal arms.

**Symmetric distributions** Our algorithm works unchanged for symmetric arm distributions (around their mean). To see this, notice that the sum of symmetric random variables is also symmetric, and thus Lemma 2 holds with probability at least  $1/2$  (instead of  $1/4$ ). We note that unlike (Khorasani and Weyer 2023), we do not require the distributions to be continuous. In particular, the above implies that our algorithm works for Gaussian arm distributions.

**Arms with lower bounded variance.** Recall that the Central Limit Theorem (CLT) implies that any (appropriately scaled) sum of random variables converges in distribution to a Gaussian, which satisfies Lemma 2. Concretely, this implies that even for non-symmetric random variables, the sample mean becomes symmetric as the sample size increases. In the following, we make this informal argument

concrete. Let  $X_i, i \geq 1$  be i.i.d random variables with mean  $\mu$ . Let  $C_\sigma \geq 0$  be a constant such that  $\rho/\sigma^3 \leq C_\sigma$  where  $\sigma^2$  is the variance of  $X_i$  and  $\rho$  is its third absolute central moment

$$\rho = \mathbb{E}[|X_i - \mu|^3].$$

Note that for any sufficiently light-tailed distribution (normal, exponential, bounded),  $C_\sigma$  is bounded (up to a numerical constant) by  $\sigma^{-1}$ . Thus, it suffices to have a lower bound on the variance to bound  $C_\sigma$ . Define  $Y_n = (\sum_{i \in [n]} (X_i - \mu)/(\sigma\sqrt{n}))$  and let  $F_n(\cdot)$  be its Cumulative Distribution Function (CDF). The Berry-Esseen Theorem (see e.g. (Shevtsova 2011)) states that

$$|F_n(x) - \Phi(x)| \leq C_\sigma/2\sqrt{n}, \quad \forall x \in \mathbb{R}, n \geq 1,$$

where  $\Phi$  is the CDF of the standard normal distribution. Taking  $x = 0, n \geq 4C_\sigma^2$  we conclude that

$$\begin{aligned} \Pr\left(\sum_{i \in [n]} X_i \leq n\mu\right) &= \Pr(Y_n \leq 0) \\ &\geq \Phi(0) - C_\sigma/2\sqrt{n} \geq 1/4, \end{aligned}$$

which is the equivalent of Lemma 2. We conclude that Algorithm 1 can work for general distributions by adding a warmup phase that collects  $4C_\sigma^2 \approx 4\sigma^{-2}$  samples for each arm and batch.

**Conjecture.** Notice that for Bernoulli arms we have  $\sigma_a^2 = \mu_a(1 - \mu_a)$ . As such, the above logic would suggest that a long warmup phase is necessary when  $\mu_a$  is close to either 0 or 1. However, Lemma 2 reveals this to be unnecessary. The reason for this gap is that we only require  $F_n(0)$  to be sufficiently large whereas the Berry-Esseen Theorem ensures  $F_n(x)$  converges to  $\Phi(x)$  for all  $x$ , which is much stronger.

Notice that the Bernoulli distribution can be extremely asymmetric when  $\mu_a$  is close to 0 or 1. This leads us to believe that it might be the worst-case distribution for the anti-concentration result in Lemma 2 (among the class of bounded random variables). Thus, we conjecture that Lemma 2 holds for any bounded arm distribution and consequently so do the regret guarantees of Algorithm 1.

## 5 Proof of Theorem 5

Recall that the pseudo-regret may be written as

$$\text{regret}_T = \sum_{a \neq a_*} n_{T,a} \Delta_a, \quad (4)$$

where  $\Delta_a = \mu_a - \mu_*$  is the optimality gap of arm  $a \in [K]$  and  $n_{t,a}$  is defined in Eq. (1). Thus, our goal is to bound  $n_{T,a}$  for each sub-optimal arm. We begin with a standard ‘‘good event’’ over which the regret is bounded deterministically. Suppose that for all  $n \in [T]$  and  $a \neq a_*$  we have

$$\hat{\mu}_{n,a_*} \leq \mu_* \quad (5)$$

$$\hat{\mu}_{n,a} \geq \mu_a - \frac{2}{(n/l) + 1} \log \frac{6TK}{\delta} - \sqrt{\frac{\sigma_a^2}{(n/l) + 1} \log \frac{2TK}{\delta}}. \quad (6)$$

Taking a union bound over Lemma 3 with  $\delta/2$  and Lemma 4 with  $\delta/2K$ , the above holds with probability at least  $1 - \delta$ . Now, suppose that arm  $a$  was played at time  $t$ . Then by the decision rule in Eq. (3), we have that  $\hat{\mu}_{n_{t-1,a},a_*} \geq \hat{\mu}_{n_{t-1,a},a}$ , and thus

$$\mu_* \geq \hat{\mu}_{n_{t-1,a},a_*} \quad (\text{Eq. (5)})$$

$$\begin{aligned} &\geq \hat{\mu}_{n_{t-1,a},a} \\ &\geq \mu_a - \frac{2}{(n_{t-1,a}/l) + 1} \log \frac{6TK}{\delta} \\ &\quad - \sqrt{\frac{\sigma_a^2}{(n_{t-1,a}/l) + 1} \log \frac{2TK}{\delta}}. \end{aligned} \quad (\text{Eq. (6)})$$

Solving this quadratic inequality for  $n_{t-1,a}$ , we have

$$n_{t-1,a} \leq l \left[ -1 + \left( \frac{\sigma_a^2}{\Delta_a^2} + \frac{2}{\Delta_a} \right) \log \frac{6TK}{\delta} \right],$$

Now, let  $t_a$  be the last time arm  $a$  was chosen. Then we have

$$\begin{aligned} n_{T,a} = n_{t_a,a} &= 1 + n_{t_a-1,a} \leq \frac{7}{2} \left( \frac{\sigma_a^2}{\Delta_a^2} + \frac{2}{\Delta_a} \right) \log^2 \frac{6TK}{\delta} \\ &\leq \frac{7}{2} \left( \frac{\mu_a}{\Delta_a^2} + \frac{2}{\Delta_a} \right) \log^2 \frac{6TK}{\delta} \\ &= \frac{7}{2} \left( \frac{\mu_*}{\Delta_a^2} + \frac{3}{\Delta_a} \right) \log^2 \frac{6TK}{\delta}, \end{aligned}$$

where the first inequality used our choice of  $l = (7/2) \log(2T/\delta)$  and the second inequality used that for any random variable in  $[0, 1]$ , we have  $\sigma^2 \leq \mu$ . Plugging this into Eq. (4) concludes the instance-dependent regret bounds.

Next, for instance-independent bounds, we use the standard method of splitting the bound according to the sub-optimality to get that for any  $c > 0$ ,

$$\begin{aligned} n_{T,a} \Delta_a &= n_{T,a} \Delta_a \left[ \mathbb{1}_{\{\Delta_a \leq c^{-1}\}} + \mathbb{1}_{\{\Delta_a^{-1} < c\}} \right] \\ &\leq \frac{n_{T,a}}{c} + \frac{7}{2} (\mu_* c + 3) \log^2 \frac{6TK}{\delta}. \end{aligned}$$

Plugging into Eq. (4) and setting  $c = \sqrt{\frac{2T}{7\mu_* K \log^2(6TK/\delta)}}$  we have

$$\begin{aligned} \text{regret}_T &\leq \sum_{a \neq a_*} \frac{n_{T,a}}{c} + \frac{7}{2} (\mu_* c + 3) \log^2 \frac{6TK}{\delta} \\ &\leq \frac{T}{c} + \frac{7}{2} K (\mu_* c + 3) \log^2 \frac{6TK}{\delta} \\ &\leq \sqrt{14\mu_* TK} \log \frac{6TK}{\delta} + 11K \log^2 \frac{6TK}{\delta}. \end{aligned}$$

Finally, we perform a similar procedure for the variance-dependent bound to get that

$$n_{T,a} \Delta_a \leq \frac{n_{T,a}}{c} + \frac{7}{2} (\sigma_a^2 c + 2) \log^2 \frac{6TK}{\delta},$$

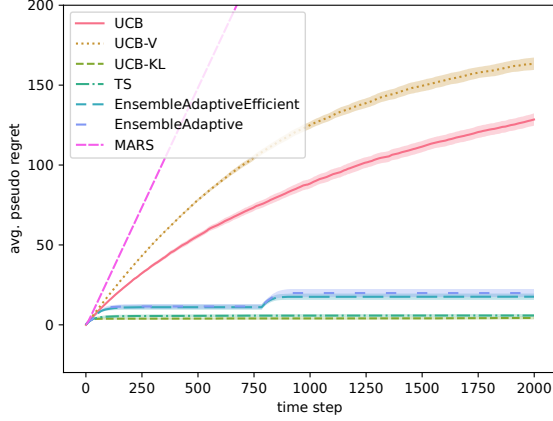


Figure 1: Results for 5 Bernoulli arms with clear, low-variance, best arm. The arms expectations tested are 0.001, 0.15, 0.2, 0.25, 0.3.

and thus setting  $c = \sqrt{\frac{2T}{7(\sum_{a \neq a_*} \sigma_a^2) \log^2(6TK/\delta)}}$ , we have

$$\begin{aligned} \text{regret}_T &\leq \sum_{a \neq a_*} \frac{n_{T,a}}{c} + \frac{7}{2} (\sigma_{a_*}^2 c + 2) \log^2 \frac{6TK}{\delta} \\ &\leq \frac{T}{c} + \frac{7}{2} \left( c \left( \sum_{a \neq a_*} \sigma_a^2 \right) + 2K \right) \log^2 \frac{6TK}{\delta} \\ &\leq \sqrt{14T \sum_{a \neq a_*} \sigma_a^2} \log \frac{6TK}{\delta} + 7K \log^2 \frac{6TK}{\delta}. \end{aligned}$$

## 6 Experiments

In this section, we compare the performance of our ensemble method to other provable methods such as UCB, UCB-V, UCB-KL, TS, and MARS, on synthetic Bernoulli, Gaussian, and beta MAB environments. In our experiments<sup>3</sup>, EnsembleAdaptive refers to running Algorithm 1 with an any-time (adaptive) batch size that scales as  $\log t$ , and EnsembleAdaptiveEfficient refers to a variant that has a better running time. We note that the only difference between the two implementations is that, when increasing the number of batches, instead of redistributing all the samples, the more efficient implementation adds samples to the new (empty) batch until it reaches the size of the remaining batches. In the following, we consider five test cases for MAB environments with five and ten arms. In all test cases, we run 100 simulations each with  $T = 2000$  steps. The performance criterion is the averaged pseudo-regret across all simulations, and we plot its 95% confidence set using the normalized empirical variance.

<sup>3</sup>Code in <https://gist.github.com/asafcassel/17ef6b4cd849fb6f0be4180aac7ed646>

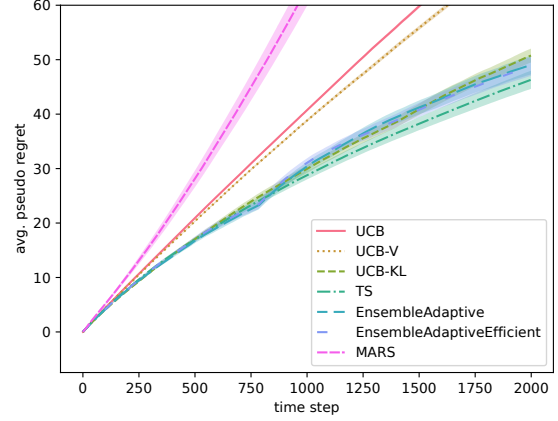


Figure 2: Results for 10 Bernoulli arms with means 0.9, 0.91, 0.92, ..., 0.99.

**Test case 1: Bernoulli arms with clear, low-variance, best arm.** In this test case, we generated a synthetic environment of MAB with five Bernoulli arms, where the means are 0.001, 0.15, 0.2, 0.25, and 0.3 for each arm respectively. Since we work with losses, the first arm is optimal (and has low variance). The other arms are non-optimal and the suboptimality gaps are relatively large, making best-arm identification easier, and demonstrating the superiority of variance-dependent methods.

As seen in Fig. 1, MARS has a long burn-in period with near-linear pseudo regret. On the other hand, the UCB, UCB-V obtain much better regret, with UCB-KL and TS obtaining the lowest regret among all algorithms. However, our Ensemble method implementations are close to UCB-KL, and significantly better than UCB, UCB-V, and MARS. We note that there is a slight jump in the plot of our Ensemble methods results due to the adaptive batch size, which increases the number of batches after approximately 800 steps.

**Test case 2: Bernoulli arms with high expected loss and low variance.** In this test case, we examine the performance of the algorithms in a high-expected loss scenario where the means are 0.9, 0.91, 0.92, ..., 0.99. We note that in this case, identifying the best arm, i.e., the first arm, is hard, since the sub-optimality gaps are relatively low, with the smallest being 0.01. The theory (specifically Lemma 2) suggests this case could challenge our approach, thus making it interesting. As arm variances are low, the variance-dependent algorithms obtain much better results than UCB and MARS. UCB-V is slightly better than UCB, and the best performance is obtained by UCB-KL, TS and our Ensemble methods, which have similar results.

**Test case 3: Bernoulli arms with random means.** This case examines the typical behavior of the algorithms for the classical case of randomly chosen means for Bernoulli arms. In this experiment, in each one of the 100 simulations, we sampled 10 numbers uniformly from the interval  $[0, 1]$ . Each number represents the mean of one Bernoulli arm. We

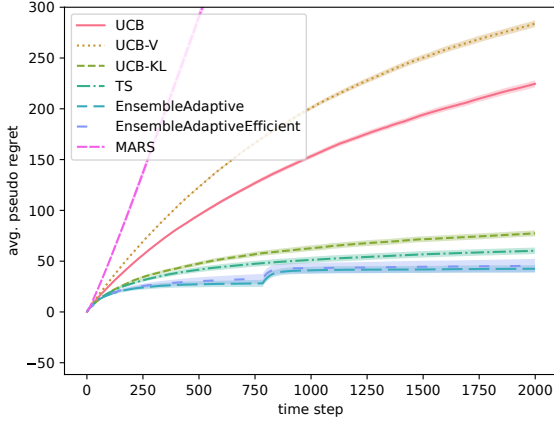


Figure 3: Results for 10 Bernoulli arms with random means.

tested all algorithms using the same sampled means. Fig. 3 presents similar trends. However, here our ensemble method has slightly better performance compared to UCB-KL and TS. This might point to the fact that the previous instances were easier for UCB-KL and TS but that our approach performs better on a randomly sampled instance.

**Test case 4: Gaussian arms with random means and variance 1.** We also consider the case of 10 Gaussian arms with randomly chosen means in  $[0, 1]$  and standard deviation  $\sigma = 1$ . In each one of the 100 simulations, we sampled uniformly at random 10 numbers in  $[0, 1]$ . Each chosen number represents the mean of the related Normal arm. We tested all the compared algorithms using the same means in all simulations. Fig. 4 demonstrates similar trends to those observed in test case 2 but in a higher-variance environment. Again, MARS has near-linear regret, and UCB-V is beaten by UCB, likely due to the high variance of the arms. However, here, our Ensemble method implementations outperform UCB-KL. Another interesting observation is that UCB-KL and our ensemble methods have a relatively high standard deviation in this experiment.

**Test case 5: Beta-distributed arms with random means.** Finally, we simulated 100 runs of MAB with 10 Beta arms that have randomly chosen means in  $[0, 1]$  and  $\alpha$  in  $[0, 5]$ . This represents a scenario where our theoretical guarantees do not hold without a warm-up, thus challenging our conjecture that our results hold even for non-symmetric distributions. Fig. 5 is a positive signal for our conjecture. MARS suffers close-to-linear average pseudo-regret, while our ensemble implementations have the lowest regret, outperforming UCB-KL, UCB and UCB-V.

**Discussion and summary of the results.** Our experiments point to our ensemble implementations having good performance across several scenarios. Our performance is close and sometimes better compared to UCB-KL, TS and consistently better than the remaining methods, especially in scenarios where the arms have low variances. Another ad-

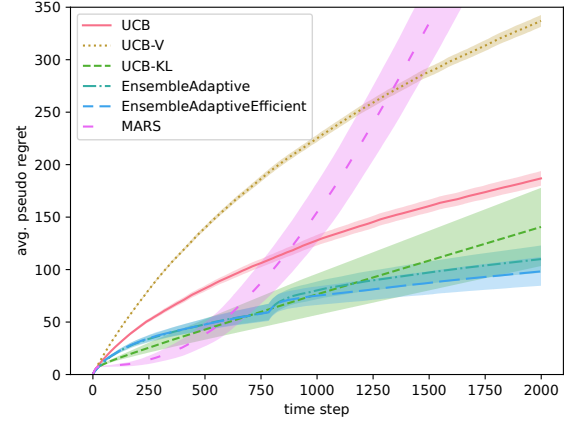


Figure 4: Results for 10 Gaussian arms with random means and the variance 1.

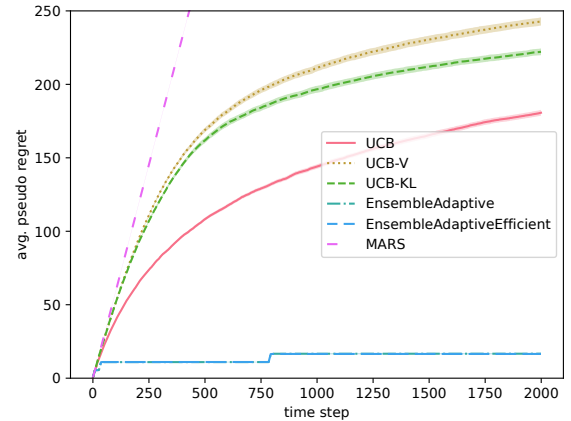


Figure 5: 10 Beta-distributed arms with random means.

vantage of our ensemble algorithms is their running time, which is close to that of standard UCB (constant per step), and memory usage, which scales logarithmically in the time horizon  $T$ . In comparison, MARS has a running time and memory usage of  $O(T^2)$ , and for Bernoulli arms, UCB-KL has to solve an optimization problem using an interior point algorithm at each time step. Overall, this positions our ensemble implementations as promising practical methods. As an added bonus, our improved efficiency variant exhibits similar and sometimes improved performance.

### Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 882396 and grant agreement No. 101078075). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European

Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work received additional support from the Israel Science Foundation (ISF, grant numbers 993/17 and 2549/19), Tel Aviv University Center for AI and Data Science (TAD), the Yandex Initiative for Machine Learning at Tel Aviv University, the Len Blavatnik and the Blavatnik Family Foundation, and by the Israeli VATAT data science scholarship.

## References

- Ash, J. T.; Zhang, C.; Goel, S.; Krishnamurthy, A.; and Kakade, S. 2021. Anti-Concentrated Confidence Bonuses for Scalable Exploration. *arXiv preprint arXiv:2110.11202*.
- Audibert, J.-Y.; Munos, R.; and Szepesvári, C. 2009. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19): 1876–1902.
- Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3): 235–256.
- Baransi, A.; Maillard, O.; and Mannor, S. 2014. Sub-sampling for Multi-armed Bandits. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I*, volume 8724 of *Lecture Notes in Computer Science*, 115–131. Springer.
- Baudry, D.; Kaufmann, E.; and Maillard, O. 2020. Sub-sampling for Efficient Non-Parametric Bandit Exploration. *CoRR*, abs/2010.14323.
- Chan, H. P. 2020. THE MULTI-ARMED BANDIT PROBLEM. *The Annals of Statistics*, 48(1): 346–373.
- Khorasani, M. M.; and Weyer, E. 2023. Maximum average randomly sampled: a scale free and non-parametric algorithm for stochastic bandits. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Kveton, B.; Szepesvári, C.; Ghavamzadeh, M.; and Boutilier, C. 2019a. Perturbed-History Exploration in Stochastic Multi-Armed Bandits. In Kraus, S., ed., *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, 2786–2793. ijcai.org.
- Kveton, B.; Szepesvári, C.; Vaswani, S.; Wen, Z.; Lattimore, T.; and Ghavamzadeh, M. 2019b. Garbage In, Reward Out: Bootstrapping Exploration in Multi-Armed Bandits. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, 3601–3610. PMLR.
- Lai, T. L.; and Robbins, H. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1): 4–22.
- Maillard, O.-A.; Munos, R.; and Stoltz, G. 2011. A Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences. In Kakade, S. M.; and von Luxburg, U., eds., *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, 497–514. Budapest, Hungary: PMLR.
- Moravej Khorasani, M.; and Weyer, E. 2023. Maximum Average Randomly Sampled: A Scale Free and Non-parametric Algorithm for Stochastic Bandits. In *Advances in Neural Information Processing Systems*, volume 36, 58865–58874. Curran Associates, Inc.
- Osband, I.; Blundell, C.; Pritzel, A.; and Van Roy, B. 2016. Deep exploration via bootstrapped DQN. *Advances in neural information processing systems*, 29.
- Osband, I.; Van Roy, B.; and Wen, Z. 2016. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, 2377–2386. PMLR.
- Peer, O.; Tessler, C.; Merlis, N.; and Meir, R. 2021. Ensemble bootstrapping for Q-Learning. In *International Conference on Machine Learning*, 8454–8463. PMLR.
- Robbins, H. 1952. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5): 527–535.
- Shevtsova, I. 2011. On the absolute constants in the Berry-Esseen type inequalities for identically distributed summands. *arXiv preprint arXiv:1111.6554*.
- Tennenholtz, G.; Merlis, N.; Shani, L.; Mannor, S.; Shalit, U.; Chechik, G.; Hallak, A.; and Dalal, G. 2022. Reinforcement learning with a terminator. *Advances in Neural Information Processing Systems*, 35: 35696–35709.
- Wang, C.; Yu, Y.; Hao, B.; and Cheng, G. 2020. Residual Bootstrap Exploration for Bandit Algorithms. *CoRR*, abs/2002.08436.
- Wiklund, T. 2023. Another look at binomial and related distributions exceeding values close to their centre. *arXiv preprint arXiv:2308.05435*.