Multi-armed Bandit Problems with History

Pannaga Shivaswamy
Department of Computer Science
Cornell University, Ithaca NY
pannaga@cs.cornell.edu

Thorsten Joachims
Department of Computer Science
Cornell University, Ithaca NY
tj@cs.cornell.edu

Abstract

In this paper we consider the stochastic multi-armed bandit problem. However, unlike in the conventional version of this problem, we do not assume that the algorithm starts from scratch. Many applications offer observations of (some of) the arms even before the algorithm starts. We propose three novel multi-armed bandit algorithms that can exploit this data. An upper bound on the regret is derived in each case. The results show that a logarithmic amount of historic data can reduce regret from logarithmic to constant. The effectiveness of the proposed algorithms are demonstrated on a large-scale malicious URL detection problem.

1 Introduction

Many real-world problems, ranging from the optimization of advertising revenue in search engines to the scheduling of clinical trials, can be modeled as multi-armed bandit problems. At each time step, the algorithm chooses one of the possible arms (i.e. advertisements, treatments) and observes its rewards. The goal is to maximize the sum of rewards over all time steps, typically expressed as regret compared to the best arm in hindsight. In the conventional formulation of the problem, the algorithm has no prior knowledge about the arms. Many applications, however, provide some data about the arms even before the algorithm starts. For example:

- A search engine company has designed K retrieval functions. Historic data is available from a beta-test on a small sample of paid users, but now the functions should be fielded in the production system as to maximize clickthrough.
- An online movie company has K different recommender functions to suggest movies to a user. When a new user signs up, he is asked to rate a few "pivotal" movies which provides historic data for optimizing the choice of recommender function in the long run.
- A clinical trial experiment was stopped due to a legal hurdle. Now the courts went in favor of continuing the clinical trial but also warn that the losses should be minimum from now on.

More generally, we define historic data as any observations of the arms that are collected before the start of the online learning algorithm. The algorithm itself has no control over the choice of arms in the historic data, nor do all arms have to be sampled uniformly.

The availability of such historic data leads to the question of how online learning algorithms can best use it to reduce regret. This problem is meaningful only for the case of stochastic arms [8, 5], since no amount of historic data can help in the adversarial setting [4].

To our best knowledge, this problem has not been studied in the literature. However, the work by [9] on bandit problem with side information is related. Their work assumes that historic data collected via some policy is available to *evaluate* a mapping from side information to arms. In the absence of side-information, their policy evaluation strategy reduces to choosing the arm with the highest mean reward on the historic data. Related is also the Sleeping Bandits Problem [7], where only a subset of the arms is active at each time step. While it can mimic historic data to some extent (e.g. it allows the addition of a new arm at any time), algorithms and bounds are weaker since they cannot rely on a separation of historic data and online learning.

This paper propose three new online learning algorithms that are able to exploit historic data. We derive

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

upper bounds on the regret for each of the three algorithms, showing that a logarithmic amount of historic data allows them to achieve constant regret. A desirable property of any bandit algorithm with historic obseravations is that the regret is zero with infinite historic data. All the three algorithms that we propose satisfy this property. We also evaluate the algorithms empirically on a malicious URL detection problem, finding that historic data can make a substantial difference on practical problems.

2 Problem Definition and Notation

The stochastic K armed bandit problem considers bounded random variables $X_{j,t} \in [0,1]$ for $1 \leq j \leq K$ and time index $t \geq 1$. Each $X_{j,t}$ denotes the reward that is incurred when the j^{th} arm is pulled the t^{th} time. For arm j, the rewards $X_{j,t}$ are independent and identically distributed with an unknown mean μ_j and an unknown variance σ_j^2. The arm with the largest mean reward is denoted by j^* *i.e.,* $\mu_{j^*} := \max_{1 \leq i \leq K} \mu_i$. Further, for any arm j, Δ_j denotes $\mu_{j^*} - \mu_j$. Often, we replace j^* with * in any notation to denote a quantity that corresponds to j^*.

Historic observations are denoted by $X^h_{j,t} \in [0,1]$ for $1 \leq j \leq K$ and $1 \leq t \leq H_j$ indexing the t^{th} instance of historic reward for arm j. H_j is the number of historic instances available for arm j, and H is defined as $H := \sum_{j=1}^{K} H_j$. The historic rewards for each arm are assumed to be drawn independently from the same distributions as the non-historic rewards.

$T_j(n)$ denotes the number of times the arm j is pulled between times 1 and n (this excludes the pulls of the arm in the historic data). The regret at time n is defined as $R_n := \mu_{j^*} n - \mu_j \sum_{j=1}^{K} \mathbf{E}[T_j(n)]$, where $\mathbf{E}[T_j(n)]$ is the expectation of $T_j(n)$. The per-round regret at time n is defined as R_n/n.

The mean reward from the historic data for arm j is defined as $\bar{X}^h_j := \frac{\sum_{t=1}^{H_j} X^h_{j,t}}{H_j}$. Mean reward of arm j during the execution of the algorithms until its n^{th} pull is defined as $\bar{X}_{j,n} := \frac{\sum_{t=1}^{n} X_{j,t}}{n}$. Analogously, the joint mean reward of arm j incorporating both the historic and the online data is $\bar{X}^h_{j,n} := \frac{\sum_{t=1}^{H_j} X^h_{j,t} + \sum_{t=1}^{n} X_{j,t}}{H_j + n}$. Finally, $V^h_{j,n}$ denotes the sample variance of the rewards for arm j until its n^{th} pull including the historic data and $V_{j,n}$ denotes the sample variance without history.

3 A Naive Algorithm

We first consider the simplest algorithm that makes use of the historic data: pick the arm with the maximum mean reward on the historic data and then to simply play that arm in every iteration. Unfortunately, this is not a very good strategy, since there is a constant probability of suffering regret in each step. By constructing an example, Theorem 1 shows that this algorithm can have regret that grows polynomially with time even if the arms have a logarithmic amount of historic data.

Theorem 1 *Consider a two armed bandit problem. The first arm has a fixed reward* $0.25 + \epsilon$, $0.5 > \epsilon > 0$, *the second arm has a Bernoulli reward with mean* 0.25. *Suppose* $H_2 = \left(3\delta \ln(n)/16\epsilon^2\right)$ *then the naive strategy has regret growing polynomially with* n *for any* $n > \exp(1/\delta)$.

Proof We lower bound the probability that the observed mean reward for the worse (second) arm is higher than the mean reward for the better (first) arm:

$$\begin{aligned}\mathbf{P}[\bar{X}^h_2 > \bar{X}^h_1] &= \mathbf{P}[B > H_2(0.25+\epsilon)] \\ &\geq \mathbf{P}\left[Z > 4\sqrt{H_2}\epsilon/\sqrt{3}\right]. \qquad (1)\end{aligned}$$

In (1) we applied Slud's inequality [1] which states:

$$\mathbf{P}[B > t] \geq \mathbf{P}\left[Z > (t - np) \,/\, \sqrt{np(1-p)}\right],$$

for a binomial random variable B parametrized by n and p such that $p \leq 1/2$, $np \leq t \leq n(1-p)$, and $Z \sim \mathcal{N}(0,1)$ (i.e. standard Gaussian random variable).

Further, for $Z \sim \mathcal{N}(0,1)$, we have from a result in [6]:

$$\mathbf{P}\left[Z > \theta\right] \geq \theta \exp\left(-\theta^2 \,/\, 2\right) \,/\, (\sqrt{2\pi}(1+\theta^2)).$$

For $\theta > 1$, it is easy to verify that $\theta/(1+\theta^2) > \exp(-\theta^2/2)$. Thus, $\mathbf{P}\left[Z > \theta\right] \geq \exp(-\theta^2) \,/\, \sqrt{2\pi}$. Applying this to (1), we get,

$$\mathbf{P}[\bar{X}^h_2 > \bar{X}^h_1] \geq \exp(-16 H_2 \epsilon^2/3) \,/\, \sqrt{2\pi}.$$

Substituting the value of H_2 from the statement of the theorem, we get,

$$\mathbf{P}[\bar{X}^h_2 > \bar{X}^h_1] \geq 1 \,/\, (\sqrt{2\pi} n^{\delta}).$$

Thus, the regret achieved by the Naive algorithm in n steps is at least $\epsilon n^{1-\delta}/\sqrt{2\pi}$. From $\theta > 1$, we get $n > \exp(1/\delta)$. ∎

4 Algorithms and Analysis

In this section, we propose three new algorithms for the stochastic multi-armed bandit problem with historic data. For each algorithm, we prove a logarithmic regret bound. Interestingly, these bounds show

Algorithm 1 – UCB1

At time t play the arm j that maximizes $\bar{X}_{j,n_j} + \sqrt{\frac{2\ln(t)}{n_j}}$, where n_j denotes $T_j(t-1)$.

Algorithm 2 – HUCB1

At time t play the arm j that maximizes $\bar{X}^h_{j,n_j} + \sqrt{\frac{2\ln(H_j+t)}{n_j+H_j}}$, where n_j denotes $T_j(t-1)$.

that a logarithmic amount of historic data is sufficient to allow these algorithms to achieve constant regret. Moreover, as the number of historic observations for every arm tends to infintiy, the regret achieved is zero. In particular, we derive bounds for the expected number of pulls for any suboptimal arm, *i.e.,* $\mathbf{E}[T_j(n)]$. From these, the regret bound can be computed as $\sum_{j:\Delta_j>0} \Delta_j \mathbf{E}[T_j(n)]$.

4.1 HUCB1: UCB1 with Historic Data

Our first algorithm is derived from the UCB1 algorithm [5]. The original UCB1 algorithm is given in Algorithm 1, while our extension of UCB1 for historic data – called HUCB1 – is shown in Algorithm 2.

For a given amount $H_j \geq 0$ of historical data for each arm j, the following theorem provides an upper bound for HUCB1 on the expected number of pulls for any suboptimal arm.

Theorem 2 *The expected number of pulls of any suboptimal arm j, for any time horizon n, satisfies,* $\mathbf{E}[T_j(n)] \leq 1 + l^+ + \frac{\pi^2(1+6H_j)}{6(2H_j+1)^2} + \frac{\pi^2(1+6H_*)}{6(2H_*+1)^2}$., *where,*

$$l^+ = \max\left(0, \frac{8\log(n+H_j)}{\Delta_j^2} - H_j\right). \qquad (2)$$

Proof Define $c^i_{t,s} = \sqrt{(2\ln(t+H_i))/(H_i+s)}$, we then have, for any integer $l > 0$,

$$\begin{aligned}
T_j(n) &= \sum_{t=1}^{n} \{I_t = j\} \leq l + \sum_{t=1}^{n} \{I_t = j, T_j(t-1) \geq l\} \\
&\leq l + \sum_{t=1}^{n} \Big\{ \bar{X}^h_{*,T_*(t-1)} + c^*_{t-1,T_*(t-1)} \\
&\qquad \leq \bar{X}^h_{j,T_j(t-1)} + c^j_{t-1,T_j(t-1)}, T_j(t-1) \geq l \Big\} \\
&\leq l + \sum_{t=1}^{n} \left\{ \min_{0<s<t} \bar{X}^h_{*,s} + c^*_{t-1,s} \leq \max_{l\leq s_j\leq t} \bar{X}^h_{j,s_j} + c^j_{t-1,s_j} \right\} \\
&\leq l + \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{s_j=l}^{t-1} \left\{ \bar{X}^h_{*,s} + c^*_{t,s} \leq \bar{X}^h_{j,s_j} + c^j_{t,s_j} \right\}.
\end{aligned}$$

The event $\left\{ \bar{X}^h_{*,s} + c^*_{t,s} \leq \bar{X}^h_{j,s_j} + c^j_{t,s_j} \right\}$ implies at least one of the following holds:[1]

$$\left.\begin{aligned}
\bar{X}^h_{*,s} &\leq \mu_* - c^*_{t,s}, \\
\bar{X}^h_{j,s_j} &\geq \mu_j + c^j_{t,s_j}, \\
\mu_* &< \mu_j + 2c^j_{t,s_j}.
\end{aligned}\right\} \qquad (3)$$

The derivation so-far is very similar to that in the original UCB1 analysis. However, from this point, having historic data starts to have a significant impact. The probability that the first two inequalities in (3) hold can be bound using Hoeffding's inequality; inclusion of historic data gives significantly tighter bounds:

$$\begin{aligned}
\mathbf{P}\left[\bar{X}^h_{*,s} \leq \mu_* - c^*_{t,s}\right] &\leq e^{-4\log(t+H_*)} = (t+H_*)^{-4}, \\
\mathbf{P}\left[\bar{X}^h_{j,s_j} \geq \mu_j + c^j_{t,s_j}\right] &\leq e^{-4\log(t+H_j)} = (t+H_j)^{-4}.
\end{aligned}$$

Further, for our choice of $l = l^+$ given in (2), the third inequality in (3) is false. We are now ready to bound the expected number of pulls for arm j. We have,

$$\begin{aligned}
\mathbf{E}[T_j(n)] &\leq l^+ + \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{s_j=l^+}^{t-1} \mathbf{P}[\bar{X}^h_{*,s} \leq \mu_* - c^*_{t,s}] \\
&\quad + \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{s_j=l^+}^{t-1} \mathbf{P}[\bar{X}^h_{j,s_j} \geq \mu_j + c^j_{t,s_j}] \\
&\leq l^+ + \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{s_j=l^+}^{t-1} \left((t+H_j)^{-4} + (t+H_*)^{-4}\right) \\
&\leq 1 + l^+ + \frac{\pi^2(1+6H_j)}{6(2H_j+1)^2} + \frac{\pi^2(1+6H_*)}{6(2H_*+1)^2}.
\end{aligned}$$

In the above, we have used the fact that $\frac{m(2m-1)\pi^2}{3(2m+1)^2} \leq \sum_{t=1}^{m}\frac{1}{t^2} \leq \frac{m(2m+2)\pi^2}{3(2m+1)^2}$, to derive an upper bound for $\sum_{t=1}^{\infty}(t+H_j)^{-2}$. ■

First, note that the above bound reduces to the bound for the UCB1 algorithm [5] when $H_j = 0$ for all j. Next, to see how much impact historic data can have on the regret, consider $n = \exp(\frac{H_j\Delta_j^2}{8}) - H_j$. In this situation $\mathbf{E}[T_j(n)] \leq \frac{\pi^2(1+6H_j)}{6(2H_j+1)^2} + \frac{\pi^2(1+6H_*)}{6(2H_*+1)^2}$ for HUCB1 which is inversely related to H_j. However, for UCB1, for the above choice of n, the upper bound on the $\mathbf{E}[T_j(n)]$ is upper bounded by $1 + \frac{8}{\Delta_j}\log\left(\exp\left(\frac{H_j\Delta_j^2}{8}\right) - H_j\right) + \frac{\pi^2}{3}$, which is approximately linear in H_j.

Rather than the confidence interval shown in Algorithm 2, at first glance one might think that

[1] It is easy to check this by negating this claim.

$\sqrt{\frac{2\log(H+t)}{(n_j+H_j)}}$ is the most natural choice to use for historic data. It can be shown that this choice leads to the following bound:

$$\mathbf{E}[T_j(n)] \leq \max\left(0, 8\frac{\log(n+H)}{\Delta_j^2} - H_j\right) + \frac{\pi^2(1+6H)}{3(2H+1)^2}.$$

It therefore has two disadvantages. First, it does not take into account that there could be different numbers of pulls for different arms in the historic data. Second, when H_j is small but H is quite large, the above bound can be worse than the one derived in Theorem 2.

4.2 HUCB3: An ϵ-Greedy Algorithm

Arguably the simplest bandit algorithm is UCB3 [5]. We now explore whether there is a similar algorithm with historic data.

We first present a slightly modified version of UCB3 in Algorithm 3. Instead of having a single rate ϵ for all arms, the following version has a different rate ϵ^j for arm j. Despite this change, the analysis of this algorithm is analogous to that of the original UCB3 algorithm. UCB3 has two parameters, d which is a lower bound on the smallest (non-zero) Δ_j and another parameter $c > 0$, but these two parameters always appear together as c/d^2.

Algorithm 3 – UCB3

Parameters $c > 0$ and $0 < d < \min_{j \neq j^*} \Delta_j$
Define a sequence for each arm: $\epsilon_n^j := \min\left(\frac{1}{K}, \frac{c}{d^2 n}\right)$
At iteration n, let i_n be the arm with the highest average reward (with no historic data), play arm i_n with probability $(1 - \sum_{j=1}^K \epsilon_n^j)$. Play arm j with probability ϵ_n^j.

To derive an algorithm that can exploit historic data, the key is to set the rates ϵ^j in a way that accounts for historic data. It might seem, at first, that replacing the rate $\frac{c}{d^2 n}$ with $\frac{c}{d^2(n+H_j)}$ would work. Unfortunately, this approach does not lead to strong guarantees.

First, observe that in the case of UCB3[2], ϵ^j is $1/K$ until $n \leq cK/d^2$. The amount of exploration done by UCB3 between times $t_0 := cK/d^2 + 1$ to n is lower bounded as follows:

$$\sum_{t=t_0}^{n} \mathbf{P}[I_t = j] = \sum_{t=t_0}^{n} \frac{c}{d^2 n} \geq \frac{c}{d^2}\log\left(\frac{nd^2}{cK}\right) - \mathcal{O}(1)$$

To derive an ϵ-greedy-like algorithm that can exploit historic data, we first find n such that the expected exploration exceeds H_j. This is done by setting H_j equal to the lower bound (we ignore the constant additive term) in the above equation. This gives $n_0 = \frac{cK}{d^2}\exp(\frac{H_j d^2}{c})$. The historic version of the ϵ-greedy algorithm will have the same rates as UCB3 in the first cK/d^2 steps. However, after that, the rate used by the historic algorithm at time step $t > cK/d^2$ will be that of UCB3 at step $(n_0 + t - cK/d^2)$. Based on these ideas, HUCB3 is presented in Algorithm 4. Note that when $H_j = 0$, ϵ_n^j for HUCB3 reduces to $c/d^2 n$, which is exactly the same rate as in UCB3.

We now provide an upper bound on the instantaneous regret of HUCB3 in Theorem 4. The proof of the following theorem is provided as as an appendix due to space constraints. The overall idea of the proof is the same as the corresponding proof for HUCB3. The two differences in our proof are the availability of historic data while applying concentration inequalities and the alternate definition of ϵ_n^j as proposed in Algorithm 4.

Algorithm 4 – HUCB3

Parameters $c > 0$ and $0 < d < 1$
Define a sequence for each arm:
$\epsilon_n^j := 1/K$ for $n \leq cK/d^2$ and
$\epsilon_n^j := \left(K(e^{\frac{H_j d^2}{c}} - 1) + \frac{d^2 n}{c}\right)^{-1}$ for $n > cK/d^2$.
At iteration n, let $j_n = \arg\max_j \bar{X}^h_{j,T_j(n-1)}$.
Play arm j_n with probability $(1 - \sum_{j=1}^K \epsilon_n^j)$. Play arm j with probability ϵ_n^j.

Theorem 3 *For any $n \geq cK/d^2$, where $c \geq 10$, HUCB3 satisfies,*

$$\mathbf{P}[I_n = j] \leq \frac{c}{d^2}\frac{1}{\left(\frac{cK}{d^2}(\exp(\frac{H_j d^2}{c}) - 1) + n\right)} + o\left(\frac{1}{n}\right).$$

The following corollary gives an upper bound on the expected number of pulls of any sub-optimal arm j. It is obtained by summing the instantaneous regrets for arm j given in Theorem 3.

Corollary 4 *HUCB3 admits the following bound for any sub-optimal arm j, for any $n > cK/d^2$,*

$$\mathbf{E}[T_j(n)] \leq \frac{c}{d^2}\log\left(\frac{\frac{cK}{d^2}(\exp(\frac{H_j d^2}{c}) - 1) + n}{\frac{cK}{d^2}\exp(\frac{H_j d^2}{c})}\right) + \mathcal{O}(1).$$

To see how the above bound changes with historic data, suppose $H_j = \frac{c}{d^2}\log(nd^2/cK)$, then $\mathbf{E}[T_j(n)] = \mathcal{O}(1)$. This again shows that a logarithmic amount of historic data suffices to achieve constant regret. It is also easy to see from the proof of Theorem 3 that these additive terms go to zero exponentially with H_j and H_* thus showing that the regret approaches zero as the number of historic observations approaches infinity.

[2]We ignore the floor on cK/d^2 for brevity.

4.3 HUCBV: Exploiting Sample Variance

Our final algorithm is based on a recent version of the UCB algorithm which also incorporates the sample variance of the rewards [2, 3]. In its most basic form, the UCBV algorithm is as shown in Algorithm 5. Audibert *et al.* [2] show that a value of $\theta = 1.2$ is enough for logarithmic convergence. The expected regret of the UCBV algorithm was shown to be upper bounded by $10\sum_{j:\mu_j<\mu_*}\left(\sigma_j^2/\Delta_j + 2\right)\log(n)$. The advantage of UCBV over algorithms that do not incorporate the sample variance is that the regret bound for UCBV involves σ_j^2/Δ_j instead of $1/\Delta_j$. The variance σ_j^2 can be substantially smaller than 1.

Algorithm 5 – UCBV

At time t play the arm j that maximizes $\bar{X}_{j,n_j} + \sqrt{\frac{2\theta V_{j,n_j}\log(t)}{n_j}} + \frac{3\theta\log(t)}{n_j}$.

The historic version of the UCBV algorithm is summarized in Algorithm 6. We will now derive an upper bound on its regret.

Algorithm 6 – HUCBV

At time t play the arm j that maximizes $B_{j,T_j(t-1),t}$ with $B_{j,s,t} = \bar{X}^h_{j,s} + \sqrt{\frac{2\theta V^h_{j,s}\log(t+H_j)}{s+H_j}} + \frac{3\theta\log(t+H_j)}{s+H_j}$.

Theorem 5 *For $\theta = 1.2$, HUCBV satisfies, $\mathbf{E}[T_j(n)] \leq 1 + v_j + \mathcal{O}(1)$ where v_j is defined as:*

$$v_j := \max\left\{8\left(\frac{\sigma_j^2}{\Delta_j^2} + \frac{2}{\Delta_j}\right)\mathcal{E}_j^n - H_j, 2\right\}. \quad (4)$$

$\mathcal{E}_j^n$ denotes $\theta\log(n + H_j)$.

Proof We start with inequality (8) from [3] which holds for any integer $u_j > 1$:

$$\mathbf{E}[T_j(n)] \leq u_j + \sum_{t=u_j+K-1}^{n}\sum_{s=u_j}^{t-1}\mathbf{P}[B_{j,s,t} > \mu_*] + \sum_{t=u_j+K-1}^{n}\mathbf{P}[\exists s : 1 \leq s \leq t-1 \text{ s.t. } B_{*,s,t} \leq \mu_*] \quad (5)$$

Our choice of u_j is the smallest integer greater than v_j defined in (4). Following [3], for $u_j \leq s \leq t$ and $t \geq 2$, our choice of u_j ensures that,

$$\sqrt{\frac{2(\sigma_j^2 + \Delta_j/2)\mathcal{E}_j^t}{s+H_j}} + \frac{3\mathcal{E}_j^t}{s+H_j} \leq \sqrt{\frac{2(\sigma_j^2 + \Delta_j/2)\mathcal{E}_j^n}{u_j+H_j}} + \frac{3\mathcal{E}_j^n}{u_j+H_j} \leq \sqrt{\frac{(2\sigma_j^2+\Delta_j)\Delta_j^2}{8(\sigma_j^2+2\Delta_j)}} + \frac{3\Delta_j^2}{8(\sigma_j^2+2\Delta_j)} \leq \frac{\Delta_j}{2}. \quad (6)$$

Consider the probability in the first term in (5),

$$\begin{aligned}
&\mathbf{P}[B_{j,s,t} > \mu_*] \\
&\leq \mathbf{P}[\bar{X}^h_{j,s} + \sqrt{\frac{2V^h_{j,s}\mathcal{E}_j^t}{s+H_j}} + \frac{3\mathcal{E}_j^t}{s+H_j} > \mu_j + \Delta_j] \\
&\leq \mathbf{P}[\bar{X}^h_{j,s} + \sqrt{\frac{2(\sigma_j^2+\Delta_j/2)\mathcal{E}_j^t}{s+H_j}} + \frac{3\mathcal{E}_j^t}{s+H_j} > \mu_j + \Delta_j] \\
&\quad + \mathbf{P}[V^h_{j,s} \geq \sigma_j^2 + \Delta_j/2] \leq \mathbf{P}[\bar{X}^h_{j,s} - \mu_j > \Delta_j/2] \\
&+ \mathbf{P}[V^h_{j,s} \geq \sigma_j^2 + \Delta_j/2] \leq 2e^{-(s+H_j)\Delta_j^2/(8\sigma_j^2+4\Delta_j/3)}.
\end{aligned}$$

In the above, the second step follows from (6). In the last step, Bernstein's inequality has been used twice and the extra term H_j in the exponent is a result of having historic data for arm j. Summing the above upper bounds from $s = u_j$ to $t-1$ and using the fact that $1 - e^{-x} \geq 2x/3$ for $0 \leq x \leq 3/4$ gives,

$$\sum_{s=u}^{t-1}\mathbf{P}[B_{j,s,t} > \mu_*] \leq \left(\frac{24\sigma_j^2}{\Delta_j^2} + \frac{4}{\Delta_j}\right)e^{-\mathcal{E}_j^n}$$

Now, consider the last term in (5), using Theorem 1 (empirical Bernstein bound) of [3], it can be upper bounded by, $3\sum_{t=u_*+1}^{n}\beta(\mathcal{E}_t^*, t)$, where, $\beta(x,t) := \inf_{1<\alpha\leq 3}\min\left(t, \frac{\log t}{\log\alpha}\right)e^{-x/\alpha}$. Therefore, we can write the upper bound on $\mathbf{E}[T_j(n)]$ as,

$$\mathbf{E}[T_j(n)] \leq 1 + \max\left\{8\left(\frac{\sigma_j^2}{\Delta_j^2} + \frac{2}{\Delta_j}\right)\mathcal{E}_j^n - H_j, 2\right\} + \left(\frac{24\sigma_j^2}{\Delta_j^2} + \frac{4}{\Delta_j}\right)ne^{-\mathcal{E}_j^n} + \sum_{t=u_*+1}^{n}\beta(\mathcal{E}_t^*, t)$$

For the choice, $\theta = 1.2$, $ne^{-\mathcal{E}_j^n}$ in the third term above

becomes, $\frac{n}{(n+H_j)^{1.2}} \leq 1$. Now consider the last term:

$$\sum_{t=u_*+1}^{n} \beta(\mathcal{E}_t^*, t) \leq \sum_{t=3}^{\infty} \beta(\mathcal{E}_t^*, t)$$
$$\leq \sum_{t=3}^{\infty} \min\left(t, \frac{\log t}{\log 1.1}\right) e^{-\theta \log(t+H_*)/1.1}$$
$$\leq \mathcal{O}(1) + \sum_{t=40}^{\infty} \frac{\log t}{\log 1.1} e^{-1.2 \log(t+H_*)/1.1}$$
$$\leq \mathcal{O}(1) + \sum_{t=40}^{\infty} \frac{\log t / \log 1.1}{(t+H_*)^{1.09}} = \mathcal{O}(1).$$

In the second step, we replaced infimum over a range to a specific value in the range. In the third step, we used the fact that $\log t / \log \alpha < t$ for $t \geq 40$ and $\alpha = 1.1$. In the last step, we used the fact that $\sum_{t=1}^{\infty} \frac{\log(t)}{(t+H_*)^{1.09}}$ is a convergent series; it is easy to verify this fact by the integral test. ■

In the case of HUCBV, $\mathbf{E}[T_j(n)] = \mathcal{O}(1)$ when $n = \exp\left(H_j/(9.6(\sigma_j^2/\Delta_j^2 + 2/\Delta_j))\right) - H_j$. Thus, with logarithmic amount of historic data, the regret is constant once again. It can again be seen from the proof that the additive terms approach zero as H_j and H_* approach infinity.

In practice, the performance of HUCBV is significantly better compared to the other versions of the algorithms that we have proposed. This will be a recurring theme in our experiments.

5 Experiments

Experiments were conducted on a large-scale real-world dataset [10] containing about 2.4 million instances. Each instance corresponds to a URL and has more than 3.2 million features associated with it. The label of an instance indicates whether the URL is malicious or not.

Five different SVM classifiers were trained using a subset of twenty thousand examples. The different SVMs corresponded to different C parameter values (which trade-off between margin and slack variables in SVM). Predictions were then obtained on all the remaining instances for all the five classifiers. The instances used in training were not used in the rest of the experiments. The five classifiers were then used as the arms of a multi-armed bandit problem. The reward was simply one when the prediction of the classifier matched the true URL reputation label and zero when it did not. The best arm differed from the second best by about 0.0208. Whereas the best arm differed from the worst by $\Delta := 0.0255$. We show per-round regret expressed as a fraction of Δ (i.e. $R_n/n\Delta$) in our results. Note that these Δ values were estimated from about 2.4 million examples. All the experiments in this section were performed by drawing random samples from this population. Hence the above Δ values denote true values for the underlying distribution from which examples were drawn.

Baselines Obviously, the original UCB algorithms and the NAIVE strategy (Section 3) are baselines in our experiments. However, we also considered three other stronger strategies (called BUCB1, BUCBV and BUCB3). These **stronger strategies (BUCB) cannot be run with arbitrary historic data** and were included merely for a wider perspective. These strategies were as follows. In any experiment, if there were H historic examples (for all the arms together), the corresponding UCB algorithms were run for extra H rounds at the start but the regret accumulated in these first H rounds was simply ignored. Note that the arms pulled in the first H iterations of the BUCB strategies are completely determined by the underlying UCB algorithm. In contrast, our algorithms for historic data can have arbitrary history for any subset of arms.

It is possible to argue that BUCB strategies have higher regret compared to HUCB algorithms. Suppose, UCB1[3] is run for $n + H$ iterations, then the number of pulls of the sub-optimal arm j is $\mathcal{O}(\ln(n + H)/\Delta_j^2)$. The number of pulls in the first H steps is $\mathcal{O}(\ln(H)/\Delta_j^2)$. The worst possible scenario is when $\Theta(\ln(H)/\Delta_j^2)$ pulls are made in the first H steps. Thus ignoring the pulls of arm j in the first H steps would give $\mathcal{O}(\ln(n + H)/\Delta_j^2 - \ln(H)/\Delta_j^2)$ pulls. In contrast, $\mathbf{E}[T_j(n)]$ for HUCB1 is of the order $\mathcal{O}(\ln(n+H_j)/\Delta_j^2 - H_j)$. This shows that the upper bound for our algorithms are much better even though these baseline strategies are stronger than completely ignoring history. Our experiments confirm this finding.

While the regret bounds we proved for the three algorithms prescribe what parameters to use, these parameter choices are often very conservative since the bounds hold for any distribution. We therefore considered variants of the proposed algorithms where the trade-off between exploration and exploitation is tuned empirically. In the case of UCB1, HUCB1, UCBV and HUCBV, we put a weight θ on the confidence interval; in the case of UCB3 and HUCB3, the parameter d was always set at 0.0208; however the parameter c was tuned. To tune the values of these parameters, UCB1, UCB3, and UCBV were run 20 times where the rewards came from a random draw of 3×10^5 instances each time. The parameters corresponding to the smallest average regret from these runs were fixed

[3] We can show similar results for UCB3 and UCBV.

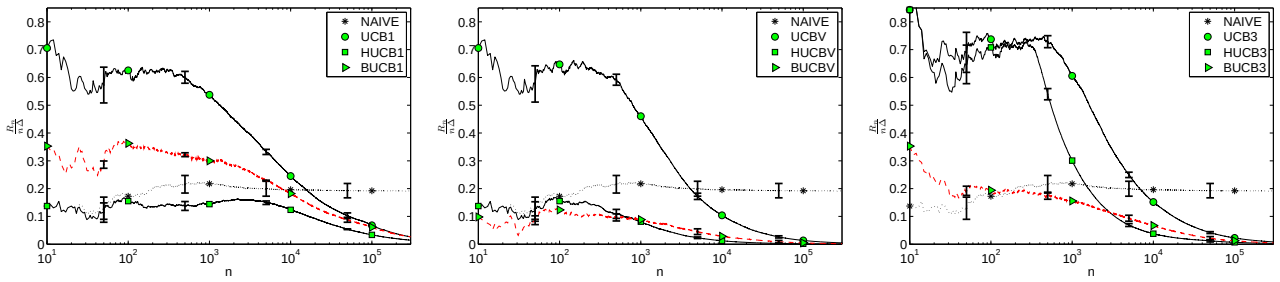

Figure 1: $R_n/n\Delta$ vs iterations with 400 historical examples per-arm.

for the rest of the experiments. For UCB1, θ was determined to be 0.2. In the case of UCB3, the parameter c was found to be 0.03. Finally, in the case of UCBV, θ was equal to 0.04. For our proposed algorithms (e.g. HUCB1) and for the baselines above (e.g. BUCB1), we simply used the same value of parameters found for the corresponding base algorithm (e.g. UCB1).

5.1 How does history affect the regret?

The aim of the first experiment was to study the behavior of regret in the presence of historic data. The total amount of historic data was fixed at 2000, uniformly split into 400 per arm. The algorithms were then run on 3×10^5 instances and the per-round regret was noted after each iteration for each algorithm. The experiments were repeated 200 times by randomly selecting the instances. A different set of historic data was selected for each run.

The results ($R_n/n\Delta$ and error bars) of this experiment are shown in Figure 1. Examining the impact of historical data on the regret, we see that all algorithms that exploit historic data indeed outperform their counterparts. This experiment shows how a comparably small amount of historic data can help achieve a substantial improvement in regret. As expected, the NAIVE algorithm performs poorly whereas, HUCBV has the best performance among all the algorithms. It can also be seen that the HUCB algorithms perform slightly better than the corresponding BUCB strategies (for large n in the case of HUB3 and HUCBV).

5.2 How does regret change with the amount of history?

In this experiment, the amount of historic data is varied to study its effect on the regret. The setup is analogous to the previous experiments and again the historic data is split uniformly among the arms. Per-round regret is measured after $5,000$ iterations.

The results of this experiment are shown in Figure 2. The regret at $5,000$ iterations for UCB1, UCB3, and UCBV is shown as a baseline. As the amount of historic data increases, the regret decreases as expected. Over most amounts of historic data, HUCB1, HUCB3, and HUCBV outperform their conventional counterparts. We once again see a small improvement over BUCB strategies as well. BUCB3 has slightly better performance over HUCB3 with a large amount of history at 5,000 iterations. This is due to the fact that UCB3 algorithms take a longer time to converge (a fact that can be verified from Figure 1 as well) due to constant rates in the beginning. For large amount of historic data, the NAIVE algorithm can reliably pick the best arm using only the historic data. However, when the amount of history is small, the regret from the naive strategy is significantly higher when compared to our algorithms.

5.3 How does the distribution of historic data affect regret?

The final experiment was designed to study the effect of unbalanced amounts of historic data per arm. Since the bounds we derived in this paper showed that the number of times an arm j is pulled depends on H_* and H_j, we fixed the number of instances at 400 for the four non-optimal arms (i.e. $H_j = 400$ when $j \neq j^*$). The number of instances for the optimal arm (H_*) was then varied in steps. The BUCB baselines have an advantage in this experiment since we cannot enforce a distribution of historic data over the arms in that case—the algorithms decide which rewards are revealed to them.

The results of this experiment are shown in Figure 3. We show the behavior of the algorithms at $5,000$ iterations.[4] When H_* is large, the sub-optimal arms are under sampled and they tend to be pulled more often in the beginning. This can be seen by almost flat curves for HUCBV and HUCB3 and by an increase in regret in the case of HUCB1 for larger H_* values. Obviously, the naive algorithm has the opposite behavior

[4] After a large number of rounds (e.g. 10^5) there was hardly any difference in regret for different H_* values.

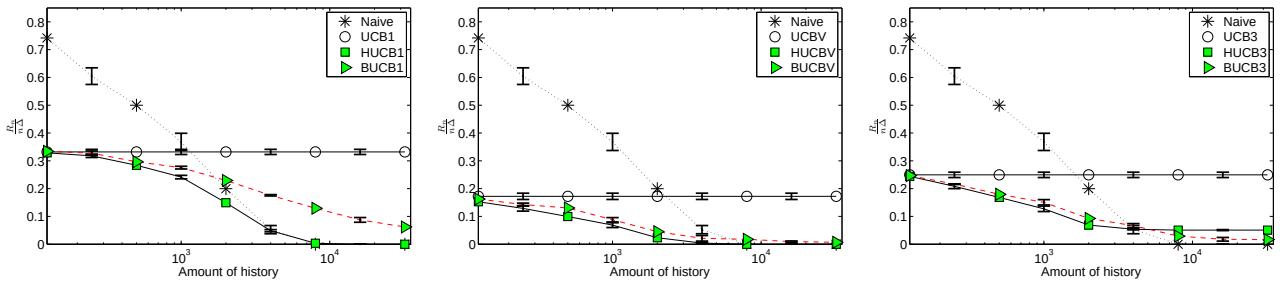

Figure 2: $R_n/n\Delta$ vs the amount of history at 5,000 iterations.

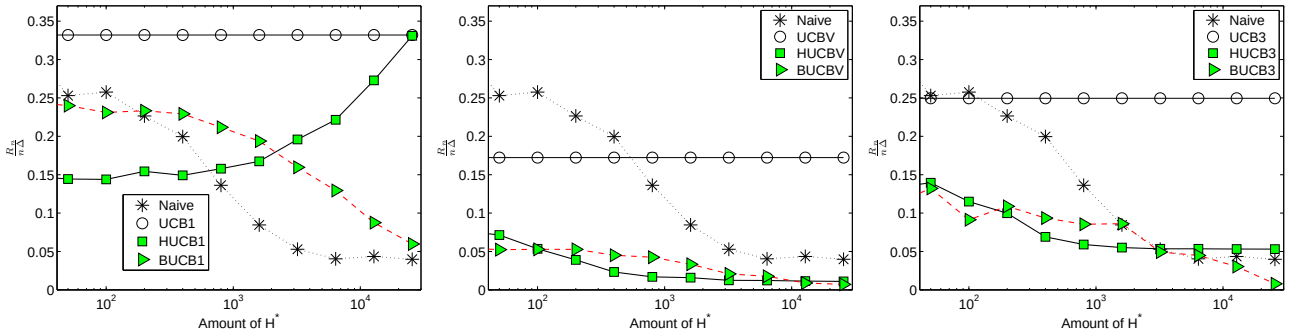

Figure 3: $R_n/n\Delta$ vs the amount of H^* at $5,000$ iterations.

compared to our algorithms since the higher H_*, the more likely it is to choose the best arm. Among the three algorithms, HUCB1 seems to be the most sensitive with respect to unbalanced history.

6 Discussion

As we pointed out in Section 4.1, the naive way of incorporating history is to have $\log(t+H)$ in the confidence interval rather than our choice of $\log(t+H_j)$. We also pointed out that the regret bounds can be significantly better for our choice of confidence interval compared to the naive choice. This leads to an intriguing possibility for the multi-armed bandit problems with no historic data. If we closely examine UCB algorithms (UCB1 for instance), the confidence interval there is $\sqrt{\frac{2\log(t)}{n_j}}$. A natural question is whether it is possible to replace t inside the logarithm such that the per-arm history during a run of UCB1 is better incorporated? An algorithm of this kind will better exploit the per-arm history during a run. Proposing a formal confidence interval and proving rigorous upper bounds in this case seem like interesting directions of research to pursue.

7 Conclusions

We proposed three novel algorithms to exploit historic data in stochastic multi-armed bandit problems. The algorithms themselves have no control over the historic data nor do the arms have to be sampled uniformly. Logarithmic finite-time regret bounds were derived for each of the three proposed algorithms. The bounds showed that already a logarithmic amount of historic data can lead to constant regret with our algorithms. Experiments were conducted on a large-scale dataset. The experiments validated our theory and showed that even a little historic data can make a significant difference in terms of regret. Overall, HUCBV has the best performance among all the algorithms. A properly tuned HUCB3 often performs better than HUCB1.

A future direction in this line of research is to derive algorithms that can exploit historic data also for other stochastic bandit settings, such as bandits with a continuum of arms, dueling bandits, *etc.* While we only showed upper bounds on the performance of the proposed algorithms in this paper, a natural next step is to also prove lower bounds for bandit problems with historic data.

Acknowledgments We thank Bobby Kleinberg for helpful discussions. This work was funded in part under NSF award IIS-0905467.

References

[1] M. Anthony and P.L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, UK, 1999.

[2] J.-Y. Audibert, R. Munos, and Cs. Szepesvári. Tuning bandit algorithms in stochastic environments. In *ALT*, pages 150–165. Springer, 2007.

[3] J.-Y. Audibert, R. Munos, and Cs. Szepesvári. Variance estimates and exploration function in multi-armed bandit. Research report 07-31, Certis - Ecole des Ponts, 2007.

[4] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

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

[6] P. Borjesson and C.-E. Sundberg. Simple approximations of the error function q(x) for communications applications. *IEEE Transactions on Communications*, 27:639–643, 1979.

[7] Robert D. Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. In *COLT*, pages 425–436, 2008.

[8] T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[9] John Langford, Alexander Strehl, and Jennifer Wortman. Exploration scavenging. In *ICML*, 2008.

[10] Justin Ma, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker. Identifying suspicious urls: An application of large-scale online learning. In *ICML*, 2009.

A Appendix: Proof of Theorem 3

Proof Only a sketch of this proof showing the differences with the corresponding steps in a similar derivation for UCB3 are given. The probability that the arm j is chosen at time t is given by:

$$\mathbf{P}[I_n = j] = \epsilon_n^j + (1 - \sum_{j=1}^{K} \epsilon_n^j)\mathbf{P}[\bar{X}^h_{j,T_j(n-1)} \geq \bar{X}^h_{*,T_*(n-1)}]$$

Moreover,

$$\mathbf{P}[\bar{X}_{j,T_j(n)} \geq \bar{X}^*_{T^*(n)}] \leq \mathbf{P}[\bar{X}^h_{j,T_j(n)} \geq \mu_j + \frac{\Delta_j}{2}] + \mathbf{P}[\bar{X}^h_{*,T_*(n)} \leq \mu_* - \frac{\Delta_j}{2}]. \quad (1)$$

Denoting $\frac{1}{2}\sum_{t=1}^{n} \epsilon_t^j$ by x_0^j, it can be shown that the first term above is upper bounded by,

$$\mathbf{P}[\bar{X}^h_{j,T_j(n)} \geq \mu_j + \frac{\Delta_j}{2}] \quad \leq \left(x_0 \mathbf{P}[T_j^R(n) \leq x_0] + \frac{2}{\Delta_j^2} e^{-\Delta_j^2 \lfloor x_0 \rfloor / 2} \right) e^{-H_j \Delta_j^2/2}, \quad (2)$$

where, we get the extra factor $\exp(-H_j\Delta_j^2/2)$ from an application of Hoeffding's inequality incorporating the historic data and $T_j^R(n)$ is the number of times arm j is selected at random in the first n draws. Since $d \leq \Delta_j$ for all j we can replace $\exp(-H_j\Delta_j^2/2)$ with $\exp(-H_j d^2/2)$.

It can further be shown that:

$$\mathbf{P}[T_j^R(n) \leq x_0^j] \leq e^{-x_0^j/5}, \quad (3)$$

using Bernstein's inequality.

Finally, we can lower bound, x_0^j as follows:

$$\begin{aligned} x_0^j &= \frac{1}{2}\sum_{t=1}^{n} \epsilon_t^j \\ &= \frac{1}{2}\sum_{t=1}^{\frac{cK}{d^2}} \frac{1}{K} + \frac{1}{2}\sum_{t=\frac{cK}{d^2}+1}^{n} \frac{c}{d^2(\frac{cK}{d^2}(e^{H_j d^2/c} - 1) + t)} \\ &\geq \frac{c}{2d^2} \log\left(\frac{\frac{ceK}{d^2}(e^{H_j d^2/c} - 1) + ne}{\frac{cK}{d^2} e^{H_j d^2/c}} \right). \end{aligned} \quad (4)$$

Using (1), (2), (3) and (4), it can be shown that:

$$\begin{aligned} \mathbf{P}[I_n = j] \leq & \frac{c}{d^2(\frac{cK}{d^2}(e^{H_j d^2/c} - 1)) + n)} \\ & + \left(\frac{c}{2d^2} P_j^{\frac{c}{10d^2}} \log\left(\frac{1}{P_j}\right) + \frac{2}{d^2} P_j^{\frac{e}{4}} \right) e^{-H_j d^2/2} \\ & + \left(\frac{c}{2d^2} P_*^{\frac{c}{10d^2}} \log\left(\frac{1}{P_*}\right) + \frac{2}{d^2} P_*^{\frac{e}{4}} \right) e^{-H_* d^2/2} \end{aligned} \quad (5)$$

where

$$P_j := \frac{\frac{cK}{d^2} e^{H_j d^2/c}}{\frac{cK}{d^2}(e^{H_j d^2/c} - 1) + n - 1}.$$

Thus, for $c \geq 10$, the last four terms in (5) are $o(\frac{1}{n})$ since $d < 1$. ■