2011 American Control Conference
on O'Farrell Street, San Francisco, CA, USA
June 29 - July 01, 2011

Parametrized Stochastic Multi-armed Bandits with Binary Rewards

Chong Jiang and R. Srikant
Coordinated Science Laboratory and
Dept of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Email: {jiang17,rsrikant}@illinois.edu

***Abstract*— In this paper, we consider the problem of multi-armed bandits with a large number of correlated arms. We assume that the arms have Bernoulli distributed rewards, independent across time, where the probabilities of success are parametrized by known attribute vectors for each arm, as well as an unknown preference vector, each of dimension n. For this model, we seek an algorithm with a total regret that is sub-linear in time and independent of the number of arms. We present such an algorithm, which we call the Three-phase Algorithm, and analyze its performance. We show an upper bound on the total regret which applies uniformly in time. The asymptotics of this bound show that for any $f \in \omega(\log(T))$, the total regret can be made to be $O(n \cdot f(T))$, independent of the number of arms.**

I. INTRODUCTION

A. Motivation

The stochastic multi-armed bandit problem is the following: suppose we are allowed to choose to "pull," or play, any one of m slot machines (also known as one-armed bandits) in each of T timesteps, where each slot machine generates a reward according to its own distribution which is unknown to us. The parameters of the reward distributions are correlated between machines, but the rewards themselves are independent across machines, and independent and identically distributed across time slots. The choice of which arm to pull may be a function of the sequence of past pulls and the sequence of past rewards. If our goal is to maximize the total reward obtained, taking expectation over the randomness of the outcomes, ideally we would pull the arm with the largest mean at all times. However, we do not know in advance which arm has the largest mean, so a certain amount of exploration is required. Too much exploration, though, wastes time that could be spent reaping the reward offered by the best arm. This exemplifies the fundamental trade-off between exploration and exploitation present in a wide class of online machine learning problems.

We consider a model for multi-armed bandit problems in which a large number of arms are present, where the expected rewards of the arms are coupled through an unknown parameter of lower dimension. Now, it is no longer necessary for each arm to be investigated in order to estimate the expected reward from that arm. Instead, we can estimate the underlying parameter; in this way, each pull can yield information about multiple arms. We present a simple algorithm, as well as a bound on the expected total regret as a function of time horizon when using this algorithm. While possibly sub-optimal, this bound is independent of the number of arms.

This model is applicable to certain e-commerce applications: suppose an online retailer has a large number of related products, and wishes to maximize revenue or profit coming from a certain set of customers. If the preferences of this set of customers are known, the list of items which are displayed can be sorted in descending order of expected revenue or profit. However, we may not know a priori what this preference vector is, so we wish to learn online by sequentially presenting each user with an item, observing whether the user buys the item, and then updating an internal estimate of the preference vector.

As a concrete example, imagine an online camera store, with hundreds of different camera models in stock. However, there are perhaps closer to ten features which people will compare when deciding which, if any, to purchase. There are permanent features of the camera itself, such as megapixel count, brand name, and year of introduction, as well as extrinsic features, such as price, review scores, and item popularity. All of these features might be considered by the customer in order to decide whether or not to buy the camera. If bought, the store gains a profit corresponding to the item. A key distinction of our model, when compared to previous work, is the incorporation of this inherently binary choice customers are faced with: to buy or not to buy.

B. Model

Our model consists of a multi-armed bandit with m arms (items) and n underlying parameters (attributes), where $m \geq n$, and potentially $m \gg n$. Each arm i is associated with a constant n-dimensional attribute vector u_i, and we assume that $\text{rank}\,[u_1, \ldots, u_m] = n$. There is also a constant but unknown n-dimensional preference vector $z^* \in \mathbb{R}^n$. The quality $\beta_i = u_i^T z^*$ of arm i is a scalar indicating how desirable the item is to a user. The expected reward of an arm i assuming a given z is defined as

$$\alpha_i(z) = f\left(u_i^T z\right) = \frac{1}{1+\exp\left(-u_i^T z\right)}, \quad \forall\, i \in \{1, \ldots, m\};$$

thus the expected rewards of all of the arms are coupled through z^*. We note that our results are applicable to more general functions f; we will comment more on this later. For notational simplicity, let $\alpha_i^* = \alpha_i(z^*)$. Let $b \in \{1, \ldots, m-1\}$ denote the number of equally best arms,

Research supported in part by AFOSR MURI FA 9550-10-1-0573.

so that $\alpha_1^* = \alpha_2^* = \cdots = \alpha_b^* > \alpha_{b+1}^* \geq \cdots \geq \alpha_m^*$. At each timestep t up to a finite time horizon T, a policy will choose to pull exactly one arm, call this arm C_t, and a reward X_t will be obtained, where $X_t \sim \text{Ber}(\alpha_{C_t}^*)$. We wish to find policies g which maximize the total expected reward, $\sum_{t=1}^{T} X_t$, or equivalently, minimize the expected total regret, $E_g\left[\sum_{t=1}^{T}(\alpha_1^* - X_t)\right] = T \cdot \alpha_1^* - E_g[\sum_{t=1}^{T} \alpha_{C_t}^*]$.

C. Prior Work

For an introduction and survey of classical multi-armed bandit problems and their variations, see Mahajan and Teneketzis [1]. One of the earliest breakthroughs on the classical multi-armed bandit problem came from Gittins and Jones [2], who showed that under geometric discounting, the optimal policy assigns an index to each arm, now known as the Gittins index, and pulls the arm with the largest Gittins index. Other proofs of this optimality have been given later by Weber [3] and Tsitsiklis [4]. Whittle [5] proved that a similar index-based result is nearly optimal in the "restless bandit" variation of this model, where the arms which are not pulled also evolve in time. While these policies greatly simplify a single m-dimensional problem into m 1-dimensional problems, it is still, in general, too computationally complex for online learning.

Lai and Robbins [6] proved an achievable $O(m \cdot \log T)$ lower bound for the expected total regret of the stochastic multi-armed bandit problem in the case of independent arms. Related work by Agrawal *et al.* [7], [8], [9] and Anantharam *et al.* [10], [11] considered similar models with i.i.d. and Markov time dependencies for each arm, and extended the results to include "multiple plays" and "switching costs".

Abe *et al.* [12] and Auer [13] considered models with finite numbers of arms, with reward distributions that are correlated through a multi-variate parameter z of dimension n, and obtained upper bounds on the regret of order $O(\sqrt{mT})$ and $O(\sqrt{nT} \cdot \log T)$, respectively. Mersereau *et al.* [14] considered a model in which the expected rewards are affine functions of a scalar parameter z, but allowed the set of arms to be a bounded, convex region in $\mathbb{R}^n$, in which case m is uncountably infinite. They then derived a policy whose expected total regret is $\Theta(\sqrt{T})$. Rusmevichientong and Tsitsiklis [15] expanded this model to allow for a multi-variate parameter z of dimension n, and showed that the expected total regret (ignoring $\log T$ factors) is $\Theta(n\sqrt{T})$. Dani *et al.* [16] independently considered a nearly identical model, and obtained similar results.

Auer *et al.* [17] considered a non-stochastic version of the multi-armed bandit problem, in which the rewards are no longer drawn from an unknown distribution, but can instead be adversarially generated. The resultant total weak regret, calculated by comparison with the single arm which is best over the entire time horizon, is shown to be $O(\sqrt{mT})$. The change from logarithmic to polynomial regret in this model is due to having rewards which are time-dependent and potentially adversarially generated, instead of being drawn from a time-independent distribution.

Audibert *et al.* [18] considered the problem of best arm identification in a stochastic multi-armed bandit setting, but where the goal is to maximize the probability of determining the best arm at the end of a time horizon, as opposed to the usual goal of minimizing total regret over a time horizon. This model is useful when considering exploration and exploitation as occurring in series, instead of in parallel. The probability of error is shown to be upper bounded by a decaying exponential in T.

Auer *et al.* [19] investigated the finite-time regret of the multi-armed bandit problem, assuming bounded but otherwise arbitrary reward distributions. Using upper confidence bound algorithms, where the confidence interval of an arm shrinks as the arm is subjected to more plays, they achieve a logarithmic upper bound on the regret, uniform over time, that scales with the "gaps" between the expected rewards for the arms.

A common idea used in crafting policies to solve the multi-armed bandit problem is that of the doubling trick [20], [21]. This technique is used to convert a parametrized algorithm which works on a time horizon T, along with its corresponding bound, into a non-parametrized algorithm that runs forever, with an upper bound that holds uniformly over time.

II. ALGORITHM AND MAIN RESULTS

A. Three-Phase Algorithm

We first present an algorithmic description of a policy for the multi-armed bandit problem described in Section I.B. This algorithm, which we call the Three-phase Algorithm, will depend on a scheduling function $g : \mathbb{N}_1 \to \mathbb{N}_0$, such that g is strictly increasing, and that $g(l) \in o\left(\exp\left(k \cdot l\right)\right), \ \forall k > 0$. Since g is not surjective in general, its inverse g^{-1} is not defined over all of $\mathbb{N}_0$; however, the monotonicity of g allows us to define g^{-1} in the following natural way: let $g^{-1}(t) = \max\left\{1, \max\left\{l \in \mathbb{N}_1 : g(l) \leq t\right\}\right\}, \ \forall t \in \mathbb{N}_0$. In Theorem 2.5, we will show that the expected total regret of this policy is $E[R_T] \in O\left(n \cdot g^{-1}(T)\right)$, independent of the number of arms m.

The algorithm requires a selection of n previously determined arms, $\Sigma = \{\sigma(i)\}_{i=1}^{n} \subseteq \{1, \ldots, m\}$, such that $U_\Sigma = \left[u_{\sigma(1)}, \ldots, u_{\sigma(n)}\right]$ has rank n. Such a choice exists since we assume $[u_1, u_2, \ldots, u_m]$ has rank n. The algorithm starts by pulling arms in Σ until each has yielded both a 1 and a 0. Note this takes a random, but a.s. finite number of timesteps (together called Phase 0). After this, the algorithm proceeds in epochs. Epoch l consists of n exploration pulls (called Phase 1), one for each arm in Σ, and $g(l)$ exploitation pulls (called Phase 2). In other words, Phase 1 refines our estimate of z^*, and Phase 2 repeatedly pulls the best arm given that estimate. If we impose a time horizon of T, epochs $1, 2, \ldots$ are appended until the time horizon T has been reached. The three phases are illustrated in Figure 1.

For each timestep t in either Phase 0 or Phase 1, an arm $i \in \Sigma$ is chosen, and the empirical count q_{i,X_t} is incremented by 1. Prior to each Phase 2 timestep during epoch l, there

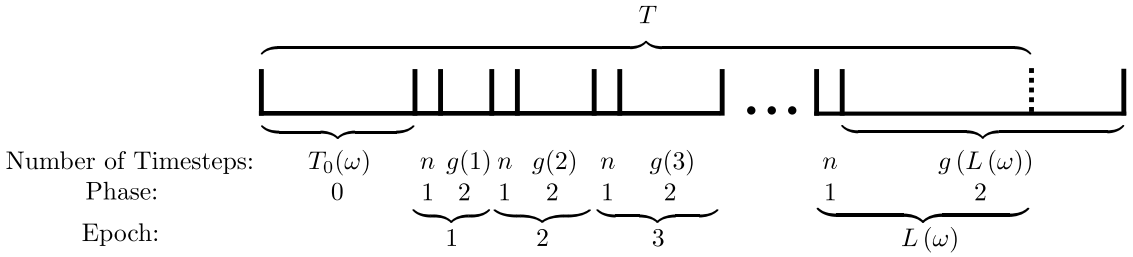

Fig. 1. Given a time horizon T, we partition the T timesteps into Phase 0, Phase 1, and Phase 2 timesteps. The Phase 1 and Phase 2 timesteps are grouped into a total of $L(\omega)$ epochs.

Algorithm 1 Three-phase Algorithm

Require: Scheduling function $g : \mathbb{N}_1 \to \mathbb{N}_0$, such that g is strictly increasing, and $g(l) \in o(\exp(k \cdot l)),\ \forall k > 0$

Require: Set of chosen arms $\Sigma = \{\sigma(i)\}_{i=1}^{n} \subseteq \{1, \ldots, m\}$, such that $U_\Sigma = \left[u_{\sigma(1)}, \ldots, u_{\sigma(n)}\right]$ has rank n.

$t \leftarrow 1$, $l \leftarrow 1$
$q_{i,0} \leftarrow 0$, $q_{i,1} \leftarrow 0\ \forall i \in \Sigma$
while $\exists i \in \Sigma, j \in \{0,1\}$, such that $q_{i,j} = 0$ **do**
  Pull arm $C_t \leftarrow \min\{i \in \Sigma : q_{i,0} = 0 \text{ or } q_{i,1} = 0\}$, obtain reward X_t {Phase 0}
  $q_{C_t,X_t} \leftarrow q_{C_t,X_t} + 1$
  $t \leftarrow t + 1$
end while
loop
  for $i \leftarrow 1$ to n **do**
    Pull arm $C_t \leftarrow \sigma(i)$, obtain reward X_t {Phase 1}
    $q_{C_t,X_t} \leftarrow q_{C_t,X_t} + 1$
    $t \leftarrow t + 1$
  end for
  Form the estimates $\hat{\alpha}_i \leftarrow \dfrac{q_{i,1}}{q_{i,0} + q_{i,1}},\ \forall i \in \Sigma$
  Form the estimate $\hat{z} \leftarrow \left(U_\Sigma^T\right)^{-1} \begin{bmatrix} f^{-1}\left(\hat{\alpha}_{\sigma(1)}\right) \\ \vdots \\ f^{-1}\left(\hat{\alpha}_{\sigma(n)}\right) \end{bmatrix}$
  $C_{(l)} \leftarrow \arg\max_{i \in \{1,\ldots,m\}} \alpha_i(\hat{z})$
  for $s \leftarrow 1$ **to** $g(l)$ **do**
    Pull arm $C_t \leftarrow C_{(l)}$, obtain reward X_t {Phase 2}
    $t \leftarrow t + 1$
  end for
  $l \leftarrow l + 1$
end loop

have already been l Phase 1 pulls. We can form empirical estimates for α_i^* based on only the Phase 0 and Phase 1 timesteps, namely $\hat{\alpha}_{i,l} = \dfrac{q_{i,1}}{q_{i,0} + q_{i,1}},\ \forall i \in \Sigma$. Note that being in Phase 2 implies we have completed Phase 0, which ensures that $q_{i,0} \geq 1$ and $q_{i,1} \geq 1$, and thus $0 < \hat{\alpha}_{i,l} < 1$.

Since f is strictly increasing and continuous, its inverse exists. Since U_Σ is an $n \times n$ matrix with full rank, $\left(U_\Sigma^T\right)^{-1}$ exists. We can now form an estimate for z^*, namely

$$\hat{z} = \left(U_\Sigma^T\right)^{-1} \begin{bmatrix} f^{-1}\left(\hat{\alpha}_{\sigma(1)}\right) \\ \vdots \\ f^{-1}\left(\hat{\alpha}_{\sigma(n)}\right) \end{bmatrix}$$

and choose $C_t = \arg\max_{i \in \{1,\ldots,m\}} \alpha_i(\hat{z})$.

Remark 2.1: In practice, LU decomposition, instead of matrix inversion, can be used to solve for $\hat{z}$. Also, since f is strictly increasing, the estimated best arm in epoch l, $C_{(l)}$, can be computed as $\arg\max_{i \in \{1,\ldots,m\}} \left(u_i^T \hat{z}\right)$.

We shall point out some of the ideas behind this algorithm. First, the algorithm is defined to run indefinitely; to obtain the total regret for any finite time horizon T, we simply terminate the algorithm when timestep T has been reached. This achieves the same outcome as an application of the doubling trick, in that the algorithm is not dependent on a time horizon T. Our algorithm is similar to the algorithm UCB2 of [19]. The main difference is that in our exploration phases, the choice of arm exploits the correlation model that we have assumed in our problem. Furthermore, as we will see later, unlike UCB2, the lengths of the exploitation phases are chosen to grow sub-exponentially in the epoch number in order to obtain a regret bound that grows (slightly larger than) logarithmically in the time horizon. As we gain more information and are able to estimate z^* more accurately, we can spend a greater fraction of timesteps exploiting the arm we think is best; this is achieved by choosing a suitable scheduling function g to control the ratio of the number of exploitation (Phase 2) pulls versus exploration (Phase 1) pulls, as a function of the epoch number l.

B. Main Results

Note that there is only randomness in the outcomes $\{X_t\}_{t=1}^{T}$, since the Three-phase Algorithm is deterministic in the selection of the arm C_t, conditioned on the history. We will use ω to denote the sample-paths of $\{X_t\}_{t=1}^{T}$. Let $L(\omega)$ denote the number of epochs (including partial epochs, as the final one may be truncated) up to timestep T, for a given sample-path ω. Note that given T and $g(l)$, $L(\omega)$ is the same deterministic function of $T_0(\omega)$ for all sample-paths.

Let $R_{i,T}(\omega)$ be the total regret up to timestep T in the Phase i timesteps for a sample-path ω. Now, let $R_T(\omega) = R_{0,T}(\omega) + R_{1,T}(\omega) + R_{2,T}(\omega)$, the total regret up to timestep T for a sample-path ω. Our goal is to find an upper bound on $E[R_T]$, the expected total regret. In particular, we are

interested in the asymptotic behavior of the upper bound as $T \to \infty$.

Lemma 2.2: For the Three-phase Algorithm, we have the following bound on the expected total Phase 0 regret up to timestep T:

$$E\left[R_{0,T}\right] \leq \alpha_1^* \sum_{i \in \Sigma} \left[\frac{1}{\alpha_i^* \left(1-\alpha_i^*\right)} \right].$$

Proof:

Note that $E\left[R_{0,T}\right] \leq \alpha_1^* E\left[\sum_{i \in \Sigma} W_i\right]$, where $W_i \sim \text{Geo}\left(\alpha_i^*\right) + \text{Geo}\left(1-\alpha_i^*\right)$ is the time it takes to first observe a 1 and subsequently observe a 0 from an arm $i \in \Sigma$. Thus,

$$\begin{aligned} E\left[R_{0,T}\right] \leq & \alpha_1^* \sum_{i \in \Sigma} \left[\left(\frac{1}{\alpha_i^*} + \frac{1}{1-\alpha_i^*} \right) \right] \\ = & \alpha_1^* \sum_{i \in \Sigma} \left[\frac{1}{\alpha_i^* \left(1-\alpha_i^*\right)} \right]. \end{aligned}$$

□

Lemma 2.3: For the Three-phase Algorithm, we have the following bound on the expected total Phase 1 regret up to timestep T:

$$E\left[R_{1,T}\right] \leq \alpha_1^* n \cdot E[L].$$

Proof:

$$\begin{aligned} E\left[R_{1,T}\right] \leq & E\left[\sum_{l=1}^{L(\omega)} \sum_{i=1}^{n} \left(\alpha_1^* - \alpha_{\sigma(i)}^* \right) \right] \\ \leq & \alpha_1^* n \cdot E[L]. \end{aligned}$$

□

Lemma 2.4: For the Three-phase Algorithm, for a given choice of scheduling function g, we have the following bound on the expected total Phase 2 regret up to timestep T:

$$E\left[R_{2,T}\right] \leq 2\alpha_1^* n \cdot \sum_{l=1}^{L'-1} \left\{ \exp\left(-l \cdot \gamma\right) g(l) \right\} + \alpha_1^* n \cdot E[L],$$

where L' is a constant which depends on $\{u_i\}_{i=1}^m$ and z^*.

Proof: Recall that $\alpha_i^* = f\left(u_i^T z^*\right)$, where $f(\beta) = \dfrac{1}{1+\exp\left(-\beta\right)}$ is strictly increasing and continuous. Thus f^{-1} is well defined, strictly increasing and continuous. Since $\alpha_1^* = \alpha_2^* = \cdots = \alpha_b^* > \alpha_{b+1}^* \geq \cdots \geq \alpha_m^*$, and because $f\left(u_i^T z\right)$ is continuous in z and defined over $\mathbb{R}^n$, it follows that there exists a neighborhood of z^*, denoted A, such that $\arg\max_{i \in \{1,\ldots,m\}} \alpha_i(z) \in \{1,\ldots,b\},\ \forall z \in A$. Since U_Σ is full rank, A must contain an open parallelotope centered at z^*, $B_{z^*}(\delta) = \left\{ z : \left\| U_\Sigma^T z - U_\Sigma^T z^* \right\|_\infty < \delta \right\}$, where $\delta > 0$ and is largest possible. An example of the problem parameters and the induced region A is shown in Figure 2.

Consider any $z \in B_{z^*}(\delta)$. By definition, $\left|u_i^T z - u_i^T z^*\right| < \delta,\ \forall i \in \Sigma$. This is equivalent to $\left| f^{-1}\left(\alpha_i(z)\right) - f^{-1}\left(\alpha_i^*\right) \right| < \delta,\ \forall i \in \Sigma$. Since f^{-1} is continuous, this is equivalent to having a set of constants $\{\underline{\alpha}_i, \overline{\alpha}_i\}_{i \in \Sigma}$, such that $\underline{\alpha}_i < \alpha_i(z) < \overline{\alpha}_i$, where $\underline{\alpha}_i = f\left(f^{-1}\left(\alpha_i^*\right) - \delta\right)$ and $\overline{\alpha}_i = f\left(f^{-1}\left(\alpha_i^*\right) + \delta\right),\ \forall i \in \Sigma$.

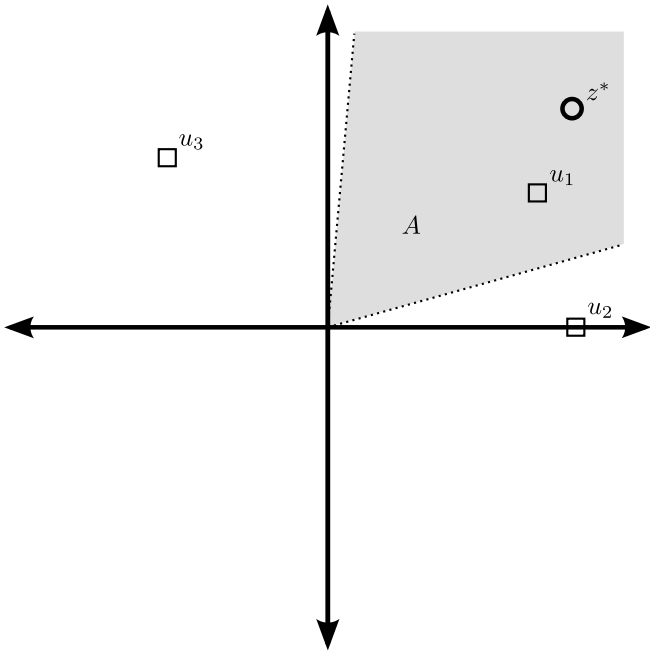

Fig. 2. As an example, consider a scenario with $n = 2$ and $m = 3$. The arms $\{u_i\}_{i=1}^3$ and the preference vector z^* are located at the indicated points. The shaded region is A.

For a Phase 2 timestep during epoch l, the algorithm forms the empirical average rewards $\hat{\alpha}_i,\ \forall i \in \Sigma$. If it is the case that $\underline{\alpha}_i < \hat{\alpha}_i < \overline{\alpha}_i,\ \forall i \in \Sigma$, then by the discussion above, $\hat{z} \in B_{z^*}(\delta) \subseteq A$, and hence, $C_t = \arg\max_{i \in \{1,\ldots,m\}} \{u_i^T \hat{z}\} \in \{1,\ldots,b\}$, and we will have chosen one of the best arms, accumulating zero regret.

Note that during epoch l, we have that $q_{i,0} + q_{i,1} \geq 2 + (l-1) \geq l\ \forall i \in \Sigma$, where the first term is due to the Phase 0 pulls and the second term is due to the Phase 1 pulls. Furthermore, during epoch l, $\hat{\alpha}_i$ is a sum of $q_{i,0} + q_{i,1}$ i.i.d. $\text{Ber}\left(\alpha_i^*\right)$ random variables, $\forall i \in \Sigma$. By the Chernoff bound,

$$\begin{aligned} P\left(\hat{\alpha}_i < \underline{\alpha}_i\right) \leq & \exp\left[-\left(q_{i,0}+q_{i,1}\right) \cdot D\left(\underline{\alpha}_i || \alpha_i^*\right)\right] \\ \leq & \exp\left[-l \cdot D\left(\underline{\alpha}_i || \alpha_i^*\right)\right], \text{ and} \\ P\left(\hat{\alpha}_i > \overline{\alpha}_i\right) \leq & \exp\left[-\left(q_{i,0}+q_{i,1}\right) \cdot D\left(\overline{\alpha}_i || \alpha_i^*\right)\right] \\ \leq & \exp\left[-l \cdot D\left(\overline{\alpha}_i || \alpha_i^*\right)\right],\ \forall i \in \Sigma, \end{aligned}$$

where $D(p||q) = p \cdot \log \dfrac{p}{q} + (1-p) \cdot \log \dfrac{1-p}{1-q}$ is the K-L divergence between two Bernoulli distributions.

Let $\gamma = \min_{i \in \Sigma} \min\left\{D\left(\underline{\alpha}_i || \alpha_i^*\right), D\left(\overline{\alpha}_i || \alpha_i^*\right)\right\}$. Note that from the definitions of $\underline{\alpha}_i$ and $\overline{\alpha}_i$, it follows that $\underline{\alpha}_i < \alpha_i^* < \overline{\alpha}_i,\ \forall i \in \Sigma$. Since $D(p||q) = 0 \iff p = q$, we have that $\gamma > 0$. By the union bound, $P\left(\exists i \in \Sigma : \hat{\alpha}_i \notin \left(\underline{\alpha}_i, \overline{\alpha}_i\right)\right) \leq 2n \exp\left(-l \cdot \gamma\right)$.

Reviewing the chain of implications, we have

$$\begin{aligned}P(\hat{z} \notin A) \leq & P(\hat{z} \notin B_{z^*}(\delta)) \\ = & P\left(\left\|U_\Sigma^T \hat{z} - U_\Sigma^T z^*\right\|_\infty > \delta\right) \\ = & P\left(\exists i \in \Sigma : \left|u_i^T z - u_i^T z^*\right| > \delta\right) \\ = & P\left(\exists i \in \Sigma : \left|f^{-1}(\hat{\alpha}_i) - f^{-1}(\alpha_i^*)\right| > \delta\right) \\ = & P(\exists i \in \Sigma : \hat{\alpha}_i \notin (\underline{\alpha}_i, \overline{\alpha}_i)) \\ \leq & 2n \exp(-l \cdot \gamma).\end{aligned}$$

Then, we have a bound on the expected per-timestep regret $r_{2,l}$ during epoch l:

$$\begin{aligned}E[r_{2,l}] = & E[r_{2,l} | \hat{z} \in A] \cdot P(\hat{z} \in A) \\ & + E[r_{2,l} | \hat{z} \notin A] \cdot P(\hat{z} \notin A) \\ \leq & 0 \cdot P(\hat{z} \in A) + \alpha_1^* \cdot P(\hat{z} \notin A) \\ \leq & 2\alpha_1^* n \exp(-l \cdot \gamma).\end{aligned}$$

Note that the above derivation depends only on the epoch number l, and is independent of the initial duration $T_0(\omega)$. Thus, $E[r_{2,l}] = E[r_{2,l}|T_0(\omega)]$.

We can now find the expected total regret in the phase 2 times up to time T:

$$\begin{aligned}& E[R_{2,T}] \\ = & E\left[\sum_{l=1}^{L(\omega)} r_{2,l} \cdot g(l)\right] \\ \leq & E\left[E\left[\sum_{l=1}^{L'-1} r_{2,l} \cdot g(l) + \sum_{l=L'}^{L(\omega)} r_{2,l} \cdot g(l) | T_0(\omega)\right]\right] \\ \leq & E\left[\sum_{l=1}^{L'-1} E[r_{2,l}|T_0(\omega)] \cdot g(l) + \sum_{l=L'}^{L(\omega)} E[r_{2,l}|T_0(\omega)] \cdot g(l)\right] \\ \leq & \sum_{l=1}^{L'-1} E[r_{2,l}] \cdot g(l) + 2\alpha_1^* n \cdot E\left[\frac{1}{2} L(\omega) | T_0(\omega)\right] \\ = & 2\alpha_1^* n \cdot \sum_{l=1}^{L'-1} \{\exp(-l \cdot \gamma) g(l)\} + \alpha_1^* n \cdot E[L],\end{aligned}$$

where $L' = \max\left\{l : \exp(-l \cdot \gamma) g(l) > \frac{1}{2}\right\}$ is a constant, independent of sample-path, that depends on $\{u_i\}_{i=1}^m$ and z^* (and is therefore unknown to the algorithm). However, since we have assumed $g(l) \in o(\exp(k \cdot l)) \ \forall k > 0$, it follows that $\lim_{l\to\infty} \exp(-l \cdot \gamma) g(l) = 0$, and thus L' is finite. Therefore, the sum $\sum_{l=1}^{L'-1} \{\exp(-l \cdot \gamma) g(l)\}$ is well defined.

□

Theorem 2.5: For the Three-phase Algorithm, we have the following asymptotic bound on the expected total regret up to timestep T: $E[R_T] = O\left(n \cdot g^{-1}(T)\right)$.

Proof:

Let us partition the total time T by Phases, $T = T_0(\omega) + T_1(\omega) + T_2(\omega)$, where $T_i(\omega)$ is the number of timesteps in Phase i for sample-path ω. Note that for all sample-paths ω in which $L(\omega) \geq 2$, we have that $g(L(\omega) - 1) \leq T$, where the left side counts the number of Phase 2 timesteps in the penultimate epoch. Thus, $L(\omega) \leq g^{-1}(T) + 1$ for all sample-paths, and hence, $E[L] \leq g^{-1}(T) + 1$.

Using Lemmas 2.2, 2.3, and 2.4,

$$\begin{aligned}E[R_T] = & E[R_{0,T} + R_{1,T} + R_{2,T}] \\ \leq & K + 2\alpha_1^* n \cdot g^{-1}(T) \\ \in & O\left(n \cdot g^{-1}(T)\right)\end{aligned}$$

where

$$\begin{aligned}K = & \alpha_1^* \sum_{i \in \Sigma} \left[\frac{1}{\alpha_i^*(1 - \alpha_i^*)}\right] \\ & + 2\alpha_1^* n \cdot \left(1 + \sum_{l=1}^{L'-1} [\exp(-l \cdot \gamma) g(l)]\right)\end{aligned}$$

is a constant which depends on n, $\{u_i\}_{i=1}^m$, and z^* (and is therefore unknown to the algorithm), but is finite for any valid set of problem parameters.

□

Corollary 2.6: If $g^{-1}(t) \in \omega(\log(t))$, then g is a valid scheduling function for the Three-phase Algorithm.

Proof: Because $g^{-1}(t) \in \omega(\log(t))$, by definition, $\lim_{t\to\infty} \frac{k_1 \cdot g^{-1}(t)}{\log(t)} > \frac{1}{k_2}$, $\forall k_1, k_2 > 0$.

Since $g : \mathbb{N}_1 \to \mathbb{N}_0$ is strictly increasing by assumption, $\lim_{l\to\infty} g(l) = \infty$. Also, note that by construction, $\forall l \in \mathbb{N}_1$, $g^{-1}(g(l)) = l$. Thus, we can make the substitution $t = g(l)$,

$$\begin{aligned}\lim_{l\to\infty} \frac{k_1 \cdot g^{-1}(g(l))}{\log(g(l))} &= \lim_{t\to\infty} \frac{k_1 \cdot l}{\log(g(l))} \\ &= \lim_{t\to\infty} \frac{\exp(k_1 \cdot l)}{g(l)} > \frac{1}{k_2},\end{aligned}$$

Hence $\lim_{t\to\infty} \frac{g(l)}{\exp(k_1 \cdot l)} < k_2$. Therefore we have the desired result, $g(l) \in o(\exp(k_1 \cdot l)),\ \forall k_1 > 0$, so g is a valid scheduling function.

□

Let $\log^*(x)$, the iterated logarithm function, be defined recursively by

$$\log^*(x) = \begin{cases}0, & \text{if } x \leq 1 \\ 1 + \log^*(\log x), & \text{if } x > 1\end{cases}.$$

Corollary 2.7: The Three-phase Algorithm can achieve $E[R_T] \in O(n \cdot \log(T) \cdot \log^*(T))$.

Proof: Let $g_{LLS}(l) = \max\{t \in \mathbb{N}_1 : \log(t) \cdot \log^*(t) \leq l\}$. Now, $g_{LLS}^{-1}(t) = \lfloor \log(t) \cdot \log^*(t) \rfloor$, so $\lim_{t\to\infty} \frac{g_{LLS}^{-1}(t)}{\log(t)} = \lim_{t\to\infty} \log^*(t) \to \infty$. Thus, $g_{LLS} \in \omega(\log(t))$, and is a valid scheduling function for the Three-phase Algorithm, so an expected total regret of $E[R_T] \in O\left(n \cdot g^{-1}(T)\right) \subseteq O(n \cdot \log(T) \cdot \log^*(T))$ is achievable.

□

Remark 2.8: In accordance with other results, such as [6], we suspect this problem has a lower bound that is asymptotically $c \cdot n \cdot \log(T)$, where c is dependent on the problem parameters $\{u_i\}_{i=1}^m$ and z^*. If this is the case, then by including the term $\log^*(T)$, we are able to obtain an upper bound which is not tight, but within a factor of $\log^*(T)$, while avoiding a dependence on the problem parameters.

C. Generalizations of the Basic Model

1) Arm-dependent Rewards: Suppose that each arm i has a potentially different value of the reward, so that instead of a $\{0,1\}$ reward, it has a $\{0, w_i\}$ reward. Furthermore, suppose that $\{w_i\}_{i=1}^m$ is known. Now, $X_t \sim w_{C_t} \cdot \text{Ber}\left(\alpha^*_{C_t}\right)$ instead of $X_t \sim \text{Ber}\left(\alpha^*_{C_t}\right)$. Let the indices of the arms be sorted by decreasing expected reward $w_i \alpha_i^*$.

Then, Theorem 2.5 generalizes with only minor modifications to the proof, yielding

$$E[R_T] \in O\left(w_1 \alpha_1^* n \cdot g^{-1}(T)\right).$$

Corollary 2.7 also generalizes, so that an expected total regret of $E[R_T] \in O\left(w_1 \alpha_1^* n \cdot \log(T) \cdot \log^*(T)\right)$ is achievable with the Three-phase Algorithm.

2) Generalized Functional Dependency on Quality: If we generalize the definition of the expected reward of an arm i assuming a preference vector z to be $\alpha_i(z) = f_i\left(u_i^T z\right), \ \forall i \in \{1, \ldots, m\}$, with the condition that $f_i(\beta) : \mathbb{R} \to (0,1)$ is strictly increasing and continuous, but otherwise arbitrary, all of the discussion above still holds. The algorithm only needs a slight modification in the formation of the estimate $\hat{z}$, which is now

$$\hat{z} = \left(U_\Sigma^T\right)^{-1} \begin{bmatrix} f_{\sigma(1)}^{-1}\left(\hat{\alpha}_{\sigma(1)}\right) \\ \vdots \\ f_{\sigma(n)}^{-1}\left(\hat{\alpha}_{\sigma(n)}\right) \end{bmatrix}.$$

III. CONCLUSIONS

We have proposed a class of parametrized multi-armed bandit problems, in which the reward distribution is Bernoulli and independent across arms and across time, with a parameter that is a non-linear function of the scalar quality of an arm. The real-valued qualities are inner products between the unknown preference and known attribute vectors. Under this model, we are able to capture the fundamentally binary choice inherent in certain online machine learning problems. Our proposed algorithm can be implemented efficiently, and is nearly optimal in the sense that its asymptotic expected total regret can be made to be $O\left(n \cdot g^{-1}(T)\right)$, for any function $g^{-1}(T) \in \omega\left(\log(T)\right)$. This is in contrast to the $O\left(m \log(T)\right)$ bound of Lai and Robbins, and the $O(n \cdot \sqrt{T})$, large-m bound of Mersereau *et al.*

Several extensions to this work are possible. For example, can small modifications to the algorithm be made in order to obtain $O(n \cdot \sqrt{T})$ regret when given a continuum of arms instead of discrete set of arms? Similarly, slight modifications to this algorithm, in order to allow for arbitrary reward distributions instead of only binary rewards, could also provide a more general application of our nonlinear model. Finally, extensions to multiple plays and having time dependent z^* and $\{u_i\}_{i=1}^m$ would be directly applicable for e-commerce applications.

REFERENCES

[1] A. Mahajan and D. Teneketzis, "Multi-armed bandit problems," in *Foundations and Applications of Sensor Management*, A. O. Hero, D. A. Castañón, D. Cochran, and K. Kastella, Eds. Springer-Verlag, 2007, ch. 6, pp. 121–151.

[2] J. C. Gittins and D. M. Jones, "A dynamic allocation index for the sequential design of experiments," *Progress in Statistics*, vol. 1, pp. 241–266, 1974.

[3] R. Weber, "On the Gittins index for multiarmed bandits," *The Annals of Applied Probability*, vol. 2, no. 4, pp. 1024–1033, Nov. 1992.

[4] J. N. Tsitsiklis, "A short proof of the Gittins index theorem," *The Annals of Applied Probability*, vol. 4, no. 1, pp. 194–199, Feb. 1994.

[5] P. Whittle, "Restless bandits: Activity allocation in a changing world," *Journal of Applied Probability*, vol. 25, pp. 287–298, 1988.

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

[7] R. Agrawal, D. Teneketzis, and V. Anantharam, "Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: Finite parameter space," *IEEE Transactions on Automatic Control*, vol. 34, no. 3, pp. 258–267, Mar. 1989.

[8] ——, "Asymptotically efficient adaptive allocation schemes for controlled Markov chains: Finite parameter space," *IEEE Transactions on Automatic Control*, vol. 34, no. 12, pp. 1249–1259, Dec. 1989.

[9] R. Agrawal, M. Hegde, and D. Teneketzis, "The multi-armed bandit problem with switching cost," in *26th IEEE Conference on Decision and Control, 1987*, vol. 26, Dec. 1987, pp. 1106–1108.

[10] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: I.I.D. rewards," *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 968–976, Nov. 1987.

[11] ——, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part II: Markovian rewards," *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 977–982, Nov. 1987.

[12] N. Abe, A. W. Biermann, and P. M. Long, "Reinforcement learning with immediate rewards and linear hypotheses," *Algorithmica*, vol. 37, no. 4, pp. 263–293, 2003.

[13] P. Auer, "Using confidence bounds for exploitation-exploration tradeoffs," *Journal of Machine Learning Research*, vol. 3, pp. 397–422, March 2003.

[14] A. J. Mersereau, P. Rusmevichientong, and J. N. Tsitsiklis, "A structured multiarmed bandit problem and the greedy policy," *IEEE Transactions on Automatic Control*, vol. 54, no. 12, pp. 2787–2802, Dec. 2009.

[15] P. Rusmevichientong and J. N. Tsitsiklis, "Linearly parameterized bandits," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 395–411, May 2010.

[16] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," in *Proc. of the 21st Annual Conference on Learning Theory*, Helsinki, Finland, July 2008, pp. 363–374.

[17] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2002.

[18] J.-Y. Audibert, S. Bubeck, and R. Munos, "Best arm identification in multi-armed bandits," in *Proc. of the 23rd Annual Conference on Learning Theory*, Haifa, Israel, June 2010, pp. 41–53.

[19] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2, pp. 235–256, 2002.

[20] G. Stoltz, "Incomplete information and internal regret in prediction of individual sequences," Ph.D. dissertation, University of Paris-Sud, Nov. 2005. [Online]. Available: http://eprints.pascal-network.org/archive/00001692/

[21] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. New York, NY: Cambridge University Press, 2006.