Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

A Structured Multiarmed Bandit Problem and the Greedy Policy

Adam J. Mersereau, Paat Rusmevichientong, John N. Tsitsiklis

Abstract— **We consider a multiarmed bandit problem where the expected reward of each arm is a linear function of an unknown scalar with a prior distribution. The objective is to choose a sequence of arms that maximizes the expected total (or discounted total) reward. We demonstrate the effectiveness of a greedy policy that takes advantage of the known statistical correlation structure among the arms. In the infinite horizon discounted reward setting, we show that both the greedy and optimal policies eventually coincide and settle on the best arm, in contrast with the Incomplete Learning Theorem for the case of independent arms. In the total reward setting, we show that the cumulative Bayes risk after T periods under the greedy policy is at most $O(\log T)$, which is smaller than the lower bound of $\Omega\left(\log^2 T\right)$ established by [1] for a general, but different, class of bandit problems. We also establish the tightness of our bounds. Theoretical and numerical results show that the performance of our policy scales independently of the number of arms.**

I. Introduction

In the multiarmed bandit problem, a decision-maker samples sequentially from a set of m arms whose reward characteristics are unknown to the decision-maker. The distribution of the reward of each arm is learned from accumulated experience as the decision-maker seeks to maximize the expected total (or discounted total) reward over a horizon. The problem is a prototypical example of the so-called *exploration versus exploitation* dilemma, where a decision-maker balances the incentive to exploit the arm with the highest expected payoff with the incentive to explore poorly understood arms for information-gathering purposes.

Nearly all previous work on the multiarmed bandit problem has assumed statistically independent arms. This assumption simplifies computation and analysis, leading to multiarmed bandit policies that decompose the problem by arm. Prominent examples are [2] for the infinite horizon problem with discounted rewards, and [1] and [3] for the finite horizon setting.

When the number of arms is large, statistical independence comes at a cost, because it typically leads to policies whose convergence time increases with the number of arms. For instance, most policies require each arm be sampled at least once. At the same time, statistical independence among arms is a strong assumption in practice. In many applications, we expect that information gained by pulling one arm will also impact our understanding of other arms. For example, in a target marketing setting, we might expect *a priori* that similar advertisements will perform similarly. The default approach in such a situation is to ignore any knowledge of correlation structure and use a policy that assumes independence. This seems intuitively inefficient because we would like to use any known statistical structure to our advantage.

Our main thesis is that known statistical structure among arms can be exploited for higher rewards and faster convergence. We show this using a version of the bandit problem where the mean reward of each arm is a known linear function of an unknown scalar on which we have a prior distribution. In the discounted infinite horizon setting, we show that a greedy policy settles on the best arm with probability one, in contrast with the Incomplete Learning Theorem known for the classic independent-arm bandit problem. In the finite horizon setting without discounting, a greedy policy achieves $O(\log T)$ cumulative risk for every horizon T, less than what is known to be possible in the classical case.

Assume that we have m arms indexed by $1, \ldots, m$, where the reward for choosing arm ℓ in period t is given by a random variable X_ℓ^t. We assume that for all $t \geq 1$ and for $\ell = 1, \ldots, m$, X_ℓ^t is given by

$$X_\ell^t = \eta_\ell + u_\ell Z + E_\ell^t \ , \tag{1}$$

where η_ℓ and u_ℓ are known for each arm ℓ, and Z and $\{E_\ell^t : t \geq 1, \ell = 1, \ldots, m\}$ are random variables. We will assume that for any given ℓ, the random variables $\{E_\ell^t : t \geq 1\}$ are identically distributed; furthermore, the random variables $\{E_\ell^t : t \geq 1, \ell = 1, \ldots, m\}$ are independent of each other and of Z.

Our objective is to choose a sequence of arms (one at each period) so as to maximize either the expected total or discounted total rewards. Define the history of the process, H_{t-1}, as the finite sequence of arms chosen and rewards observed through the end of period $t-1$. For each $t \geq 1$, let $\mathcal{H}_{t-1}$ denote the set of possible histories up until the end of period $t-1$. A *policy* $\Psi = (\Psi_1, \Psi_2, \ldots)$ is a sequence of functions such that $\Psi_t : \mathcal{H}_{t-1} \to \{1, 2, \ldots, m\}$ selects an arm in period t based on the history until the end of period $t-1$. For each policy Ψ, the total discounted reward is given by $\mathsf{E}\left[\sum_{t=1}^{\infty} \beta^t X_{J_t}^t\right]$, where $0 < \beta < 1$ denotes the discount factor, and the random variables $J_1, J_2, \ldots$ correspond to the sequence of arms chosen under the policy Ψ, that is, $J_t = \Psi_t(H_{t-1})$. For every $T \geq 1$, we define the T-period cumulative *regret* under Ψ given $Z = z$ as $\text{Regret}(z, T, \Psi) =$

The first author thanks the University of Chicago Graduate School of Business for support. The research of the second and third authors was supported in part by the National Science Foundation through grants DMS-0732196 and ECCS-0701623, respectively.

A. J. Mersereau is with the Kenan-Flagler Business School, University of North Carolina, Chapel Hill, NC 27599, ajm@unc.edu

P. Rusmevichientong is with the School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853, paatrus@cornell.edu

J. N. Tsitsiklis is with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, jnt@mit.edu

$\sum_{t=1}^{T} \mathsf{E}\Big[\max_{\ell=1,\ldots,m}(\eta_\ell + u_\ell z) - (\eta_{J_t} + u_{J_t} z) \;\Big|\; Z = z\Big]$, and the T-period cumulative Bayes risk of the policy Ψ by $\text{Risk}(T,\Psi) = \mathsf{E}_Z[\text{Regret}(Z,T,\Psi)]$.

Although classical formulations of the multiarmed bandit problem often allow for dependence among the arms ([4], [5], [6]), there is relatively little work on the analysis of policies in settings with dependent arms. The papers [7], [8], [9], [10], [11], and [12] analyze multiarmed bandit problems with various forms of arm dependency.

II. Infinite Horizon With Discounted Rewards

In this section, we consider the problem of maximizing the total expected discounted reward. We make the following assumption on the random variables Z and E_ℓ^t.

Assumption 2.1:

(a) The random variable Z is continuous, and $\mathsf{E}\left[Z^2\right] < \infty$. Furthermore, for every t and ℓ, we have $\mathsf{E}[E_\ell] = 0$ and $\gamma_\ell^2 := \mathsf{E}\left[\left(E_\ell^t\right)^2\right] < \infty$.
(b) We have $u_\ell \neq 0$, for every ℓ.
(c) If $k \neq \ell$, then $u_k \neq u_\ell$.

Assumption 2.1(a) places mild moment conditions on the underlying random variables, while Assumption 2.1(b) ensures that the reward of each arm is influenced by the underlying random variable Z. In Section II-C, we will explore the consequences of relaxing this assumption and allow some of the coefficients u_ℓ to be zero. Finally, Assumption 2.1(c) is only introduced for simplicity and results in no loss of generality. Indeed, if the coefficient u_ℓ is the same for several arms, we should only consider playing one with the largest value of η_ℓ, and the others can be eliminated.

A. Complete Learning

Fix an arbitrary policy Ψ, and for every t, let $\mathcal{F}_t$ be the σ-field generated by the history H_t under that policy. Let $Y_t = \mathsf{E}\left[Z \mid \mathcal{F}_t\right]$ and $V_t = \mathsf{E}\left[(Z - Y_t)^2 \mid \mathcal{F}_t\right] = \mathsf{Var}(Z \mid \mathcal{F}_t)$. The following result states that, under Assumption 2.1, we have complete learning for every policy Ψ.

Theorem 1: Under Assumption 2.1, for every policy Ψ, Y_t converges to Z and V_t converges to zero, almost surely.

Proof: Let us fix a policy Ψ, and let $J_1, J_2, \ldots$ be the sequence of arms chosen under Ψ. The sequence $\{Y_t\}$ is a martingale with respect to the filtration $\{\mathcal{F}_t : t \geq 0\}$. Furthermore, since $\mathsf{E}\left[Z^2\right] < \infty$, it is a square integrable martingale. It follows that Y_t converges to a random variable Y, almost surely, as well as in the mean-square sense. Furthermore, Y is equal to $\mathsf{E}[Z \mid \mathcal{F}_\infty]$, where $\mathcal{F}_\infty$ is the smallest σ-field containing $\mathcal{F}_t$ for all t ([13]).

We wish to show that $Y = Z$. For this, it suffices to show that Z is $\mathcal{F}_\infty$-measurable. To this effect, we define

$$\hat{Y}_t = \frac{1}{t}\sum_{\tau=1}^{t} \frac{X_{J_\tau}^\tau - \eta_{J_\tau}}{u_{J_\tau}} = Z + \frac{1}{t}\sum_{\tau=1}^{t} \frac{E_{J_\tau}^\tau}{u_{J_\tau}}.$$

Then, $\mathsf{Var}(\hat{Y}_t - Z) = \frac{1}{t^2}\sum_{\tau=1}^{t} \frac{\gamma_{J_\tau}^2}{u_{J_\tau}^2} \leq \frac{\max_\ell(\gamma_\ell^2/u_\ell^2)}{t}$. It follows that $\hat{Y}_t$ converges to Z in the mean square. Since $\hat{Y}_t$ belongs to $\mathcal{F}_\infty$ for every t, it follows that its limit, Z, also belongs to $\mathcal{F}_\infty$. This completes the proof of convergence of Y_t to Z.

The definition of V_t implies that $V_t = \mathsf{E}\left[(Z - Y_t)^2 \mid \mathcal{F}_t\right] = \mathsf{E}\left[Z^2 \mid \mathcal{F}_t\right] - Y_t^2$, so that V_t is a nonnegative supermartingale. Therefore, V_t converges almost surely (and thus, in probability) to some random variable V. Since $\lim_{t\to\infty} \mathsf{E}[V_t] = 0$, V_t also converges to zero in probability. Therefore, $V = 0$ with probability one. ∎

In our problem, the rewards of the arms are correlated through a single random variable Z to be learned, and thus, we intuitively have only a "single" arm in our setting. Because uncertainty is univariate, we have complete learning under any policy. This is contrast with the classical multi-armed bandit case, where the Incomplete Learning Theorem (see, for example, [14]) states that no policy is guaranteed to find the best arm. As a consequence of Theorem 1, we will show in Theorem 3 in Section II-B that an optimal policy will settle on the best arm with probability one.

B. A Greedy Policy

From Theorem 1, the posterior mean of Z, under any policy, converges to the true value of Z almost surely. This suggests that a simple greedy policy – one whose decision at each period is based solely on the posterior mean – might perform well. A greedy policy is a policy whose sequence of decisions $\left(J_1^G, J_2^G, \ldots\right)$ is defined by: for each $t \geq 1$,

$$J_t^G = \arg\max_{\ell=1,\ldots,m} \left\{\eta_\ell + u_\ell \mathsf{E}\left[Z \mid \mathcal{F}_{t-1}^G\right]\right\},$$

where $\left\{\mathcal{F}_t^G : t \geq 1\right\}$ denotes the corresponding filtration; for concreteness, we assume that ties are broken in favor of arms with lower index. Note that the decision J_t^G is a myopic one, based only on the conditional mean of Z given the past observations up until the end of period $t-1$.

Intuitively, the quality of the greedy decision will depend on the variability of Z relative to the difference between the expected reward of the best and second best arms. To make this concept precise, we introduce the following definition. For any μ, let $\Delta(\mu)$ denote the difference between the reward of the best and the second best arms, that is,

$$\Delta(\mu) = \max_{\ell=1,\ldots,m}\{\eta_\ell + \mu u_\ell\} - \max_{\ell=1,\ldots,m:\, \ell\neq\pi^G(\mu)}\{\eta_\ell + \mu u_\ell\},$$

where $\pi^G(\mu) = \arg\max_{\ell=1,\ldots,m}\{\eta_\ell + \mu u_\ell\}$. Figure 1 shows an example of the function $\Delta(\cdot)$ in a setting with 4 arms. Note that $\Delta(\cdot)$ is a continuous and nonnegative function. As seen from Figure 1, $\Delta(\mu)$ may be zero for some μ. However, given our assumption that the coefficients u_ℓ are distinct, one can verify that $\Delta(\mu)$ has at most $m-1$ zeros.

The next theorem shows that, under any policy, if the posterior standard deviation is small relative to the the mean difference between the best and second best arms, then it is optimal to use a greedy decision.

Theorem 2: Under Assumption 2.1, there exists a constant δ that depends only on β and the coefficients u_ℓ, with the

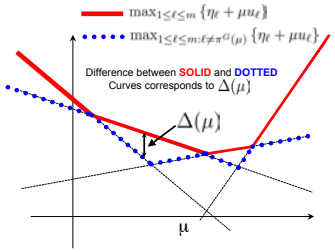

Fig. 1. An example of $\Delta(\cdot)$ with 4 arms.

following property. If we follow a policy Ψ until some time $t-1$, and if

$$\Delta\left(\mathsf{E}\left[Z \mid \mathcal{F}_{t-1}\right]\right)/\sqrt{\mathsf{Var}\left[Z \mid \mathcal{F}_{t-1}\right]} \;>\; \delta\;,$$

then it is optimal to apply the greedy policy in period t. (Here, $\mathcal{F}_{t-1}$ is the σ-field generated by the history H_{t-1}.)

Proof: Let us fix a policy Ψ and some $t \geq 1$, and define $\mu_{t-1} = \mathsf{E}\left[Z \mid \mathcal{F}_{t-1}\right]$. Let J^* and R^* denote the greedy decision and the corresponding reward in period t: $J^* = \arg\max_{\ell=1,\ldots,m}\{\eta_\ell + u_\ell\mu_{t-1}\}$ and $R^* = \eta_{J^*} + u_{J^*}\mu_{t-1}$. We will first establish a lower bound on the total expected discounted reward (from time t onward) associated with a policy that uses a greedy decision in period t and thereafter. For each $s \geq t-1$, let $M_s^G = \mathsf{E}\left[Z \mid \mathcal{F}_s^G\right]$ denote the conditional mean of Z under this policy, where $\mathcal{F}_s^G$ is the σ-field generated by the history of the process when policy Ψ is followed for up to time $t-1$, and the greedy policy is followed thereafter, so that $\mathcal{F}_{t-1}^G = \mathcal{F}_{t-1}$. Under this policy, the expected reward at each time $s \geq t$ is

$$\mathsf{E}\left[\max_{\ell=1,\ldots,m}\{\eta_\ell + u_\ell M_{s-1}^G\} \;\middle|\; \mathcal{F}_{t-1}\right] \geq R^*,$$

where we first used Jensen's inequality and the fact that the sequence M_s^G, $s \geq t-1$, forms a martingale. Thus, the present value at time t of the expected discounted reward under a strategy that uses a greedy decision in period t and thereafter is at least $R^*/(1-\beta)$.

Now, consider any policy that differs from the greedy policy at time t, and plays some arm $k \neq J^*$. Let $R_k = \eta_k + u_k\mathsf{E}\left[Z \mid \mathcal{F}_{t-1}\right] = \eta_k + u_k\mu_{t-1}$ denote the immediate expected reward in period t. The future rewards under this policy are upper bounded by the expected reward under the best arm. Thus, under this policy, the expected total discounted reward from t onward is upper bounded by $R_k + \frac{\beta}{1-\beta}\mathsf{E}\left[\max_{\ell=1,\ldots,m}\{\eta_\ell + u_\ell Z\} \;\middle|\; \mathcal{F}_{t-1}\right]$. Using the fact that $\max_\ell\{\eta_\ell + u_\ell Z\} \leq \max_\ell\{\eta_\ell + u_\ell\mu_{t-1}\} + \max_\ell\{u_\ell(Z-\mu_{t-1})\} = R^* + \max_\ell\{u_\ell(Z-\mu_{t-1})\}$, we can show that the future rewards under this policy are upper bounded by

$$R_k + \frac{\beta}{1-\beta}R^* + \frac{\beta}{1-\beta}\mathsf{E}\left[\max_{\ell=1,\ldots,m}\{u_\ell\left(Z-\mu_{t-1}\right)\} \;\middle|\; \mathcal{F}_{t-1}\right].$$

Note that $\mathsf{E}\left[\max_\ell\{u_\ell\left(Z-\mu_{t-1}\right)\}\middle|\mathcal{F}_{t-1}\right] \leq \left(\max_\ell|u_\ell|\right)\left(\mathsf{Var}(Z \mid \mathcal{F}_{t-1})\right)^{1/2}$, which implies that the present value at time t of the total discounted reward is upper bounded by

$$R_k + \frac{\beta}{1-\beta}R^* + \frac{\beta}{1-\beta}\left(\max_{\ell=1,\ldots,m}|u_\ell|\right)\sqrt{\mathsf{Var}\left(Z \mid \mathcal{F}_{t-1}\right)}$$

Recall that the total discounted reward under the greedy policy is at least $R^*/(1-\beta) = R^* + \beta R^*/(1-\beta)$. Moreover, for any arm $k \neq J^*$, $R^* \geq R_k + \Delta(\mu_{t-1})$, and thus,

$$\frac{R^*}{(1-\beta)} \;\geq\; R_k + \Delta\left(\mu_{t-1}\right) + \frac{\beta}{1-\beta}R^*$$

Comparing the expected discounted rewards of the two policies, we see that a greedy policy is better than any policy that takes a non-greedy action if

$$\Delta(\mu_{t-1}) \;>\; \frac{\beta}{1-\beta}\left(\max_{\ell=1,\ldots,m}|u_\ell|\right)\sqrt{\mathsf{Var}\left(Z \mid \mathcal{F}_{t-1}\right)}\;,$$

which is the desired result. ■

It is straightforward to combine Theorems 1 and 2 to prove that the greedy and optimal policies both settle on the best arm with probability one. For the sake of brevity, we state the following result without proof.

Theorem 3: Under Assumption 2.1, an optimal policy eventually agrees with the greedy policy, and both settle on the best arm with probability one.

C. Relaxing Assumption 2.1(b) Can Lead to Incomplete Learning

In this section, we briefly discuss the consequences of allowing the coefficients u_ℓ to be zero for some arms. For the remainder of the section, we restrict our attention to a setting where the underlying random variables Z and E_ℓ^t are normally distributed. Given this assumption, it is straightforward to formulate the problem as a Markov Decision Process (MDP) with state characterized by (μ, σ), where μ and σ are the posterior mean and the posterior standard deviation, respectively. This state can be updated based on observations using well-known formulas. The expected reward of pulling arm ℓ in state (μ, σ) is $r\left((\mu,\sigma),\ell\right) = \eta_\ell + u_\ell\mu$. Although the reward function is unbounded, it follows from Theorem 1 in [15] that a stationary policy is optimal.

If we restrict ourselves to stationary policies, when an arm ℓ is played with $u_\ell = 0$, the state remains the same and the policy will keep playing the same arm forever. Thus, an arm with $u_\ell = 0$ can be viewed as a "retirement option". When such a retirement option exists (without loss of generality, we may assume that this option corresponds to the arm $\ell = 1$ with $u_1 = 0$), we can construct examples where either an optimal or a greedy policy will retire on the wrong arm with positive probability. We have the following result.

Theorem 4: If the random variables Z and E_ℓ^t are normally distributed, and if $\eta_1 > \max_{\ell:u_\ell\neq 0}\{\eta_\ell + u_\ell\mu\}$ for some $\mu \in \mathbb{R}$, then the optimal and greedy policies disagree forever with positive probability. Furthermore, under either the optimal or the greedy policy, there is positive probability of retiring even though arm 1 is not the best arm.

III. Finite Horizon With Undiscounted Rewards

We now consider a finite horizon version of the problem, under the expected total reward criterion, and focus on identifying a policy with small cumulative Bayes risk. As in Section II, a simple greedy policy performs well in this setting. We introduce the following assumption on the coefficients u_ℓ and on the error random variables E_ℓ^t.

Assumption 3.1:

(a) There exist positive constants b, and λ such that for every ℓ and $x \geq 0$, $\Pr\left(|E_\ell^t| \geq x\right) \leq be^{-\lambda x}$,

(b) There exist positive constants $\underline{u}$ and $\overline{u}$ such that for every ℓ, $\underline{u} \leq |u_\ell| \leq \overline{u}$.

We view b, λ, $\underline{u}$ and $\overline{u}$ as absolute constants, which are the same for all instances of the problem under consideration. Our subsequent bounds will depend on these constants, although the dependence will not be made explicit. The first part of Assumption 3.1 states that the tails of the random variables $|E_\ell^t|$ decay exponentially. It is equivalent to an assumption that all $|E_\ell^t|$ are stochastically dominated by a shifted exponential random variable. We use the above observations to derive an upper bound on the moment generating function of E_ℓ^t, and then a lower bound on the corresponding large deviations rate function, ultimately resulting in tail bounds for the estimators Y_t given in the following theorem.

Theorem 5: Under Assumption 3.1, there exist positive constants f_1 and f_2 depending only on the parameters b, λ, $\underline{u}$, and $\overline{u}$, such that for every $t \geq 1$, $a \geq 0$, and $z \in \mathbb{R}$,

$$\Pr\left(|Y_t - z| > a \mid Z = z\right) \leq e^{-f_1 ta} + e^{-f_1 ta^2},$$

$$\mathsf{E}\left[(Y_t - z)^2 \middle| Z = z\right] \leq \frac{f_2}{t}, \text{ and } \mathsf{E}\left[|Y_t - z| \middle| Z = z\right] \leq \frac{f_2}{\sqrt{t}} .$$

The second part of Assumption 3.1 requires, in particular, the coefficients u_ℓ to be nonzero. It is imposed because if some u_ℓ is zero, then, the situation is similar to the one encountered in Section II-C: a greedy policy may settle on a non-optimal arm, with positive probability, resulting in a cumulative regret that grows linearly with time.

We will study the following variant of a greedy policy.

Greedy Policy for Finite Horizon Undiscounted Rewards
Initialization: Set $Y_0 = 0$.
Description: For periods $t = 1, 2, \ldots$

1) Sample arm $J_t = \arg\max_{\ell=1,\ldots,m} \{\eta_\ell + Y_{t-1} u_\ell\}$, with ties broken arbitrarily.
2) Let $X_{J_t}^t$ denote the observed reward from arm J_t.
3) Update the estimate Y_t by letting $Y_t = (1/t) \sum_{s=1}^t \left(X_{J_s}^s - \eta_{J_s}\right) / u_{J_s}$.

Output: A sequence of arms played $\{J_t : t = 1, 2, \ldots\}$.

The two main results of this section are stated in the following theorems. The first provides an upper bound on the regret $\text{Regret}\,(z, T, \text{GREEDY})$ under the GREEDY policy. The proof is given in Section III-A.

Theorem 6: Under Assumption 3.1, there exist positive constants c_1 and c_2 that depend only on the parameters b, λ, $\underline{u}$, and $\overline{u}$, such that for every $z \in \mathbb{R}$ and $T \geq 1$,

$$\text{Regret}\,(z, T, \text{GREEDY}) \leq c_1|z| + c_2\sqrt{T} .$$

Furthermore, the above bound is tight in the sense that there exists a problem instance involving two arms and a positive constant c_3 such that, for every policy Ψ and $T \geq 2$, there exists $z \in \mathbb{R}$ with

$$\text{Regret}\,(z, T, \Psi) \geq c_3\sqrt{T} .$$

On the other hand, for every problem instance that satisfies Assumption 3.1, and every $z \in \mathbb{R}$, the infinite horizon regret under the GREEDY policy is bounded; that is, $\lim_{T\to\infty} \text{Regret}\,(z, T, \text{GREEDY}) < \infty$.

From the regret bound of Theorem 6, and by taking expectation with respect to Z, we obtain an easy upper bound on the cumulative Bayes risk, namely, $\text{Risk}(T, \text{GREEDY}) = O(\sqrt{T})$. Furthermore, the tightness results suggest that this bound may be the best possible. Surprisingly, as established by the next theorem, if Z is continuous and its prior distribution has a bounded density function, the cumulative Bayes risk only grows at the rate of $O(\log T)$, independent of the number of arms. The proof is given in Section III-B.

Theorem 7: Under Assumption 3.1, if Z is a continuous random variable whose density function is bounded above by A, then exist positive constants d_1 and d_2 that depend only on A and the parameters b, λ, $\underline{u}$, and $\overline{u}$, such that for every $T \geq 1$,

$$\text{Risk}\,(T, \text{GREEDY}) \leq d_1 \mathsf{E}\left[|Z|\right] + d_2 \ln T .$$

Furthermore, this bound is tight in the sense that there exists a problem instance with two arms and a positive constant d_3 such that for every $T \geq 1$, and every policy Ψ,

$$\text{Risk}\,(T, \Psi) \geq d_3 \ln T .$$

The above risk bound is smaller than the lower bound of $\Omega\left(\log^2 T\right)$ established by [1]. To understand why this is not a contradiction, let X_ℓ denote the mean reward associated with arm ℓ, that is, $X_\ell = \eta_\ell + u_\ell Z$, for all ℓ. Then, for any $i \neq \ell$, X_i and X_ℓ are perfectly correlated, and the conditional distribution $\Pr\{X_\ell \in \cdot \mid X_i = x_i\}$ of X_ℓ given $X_i = x_i$ is degenerate, with all of its mass at a single point. In contrast, the $\Omega\left(\log^2 T\right)$ lower bound of [1] assumes that the cumulative distribution function of X_ℓ, conditioned on X_i, has a continuous and bounded derivative over an open interval, which is not the case in our model.

A. Regret Bounds: Proof of Theorem 6

In this section, we will establish an upper bound on the regret, conditioned on any particular value z of Z, and establish its tightness. Let us first introduce some notation. We define a reward function $g : \mathbb{R} \to \mathbb{R}$, as follows: for every $z \in \mathbb{R}$, we let

$$g(z) = \max_{\ell=1,\ldots,m} \{\eta_\ell + u_\ell z\}.$$

Note that $g(\cdot)$ is convex. Let $g^+(z)$ and $g^-(z)$ be the right-derivative and left-derivative of $g(\cdot)$ at z, respectively. These directional derivatives exist at every z, and by Assumption 3.1, $\max\{|g^+(z)|, |g^-(z)|\} \leq \overline{u}$ for all z. Both left and right derivatives are nondecreasing with $g^+(-\infty) = g^-(-\infty) =$

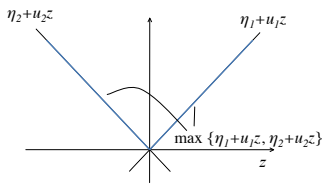

Fig. 2. The two-arm instance with $(\eta_1, u_1) = (0, 1)$ and $(\eta_2, u_2) = (0, -1)$, used to prove the tightness result in Theorem 6.

$\min_\ell u_\ell$ and $g^+(\infty) = g^-(\infty) = \max_\ell u_\ell$. (We define $g^+(\infty) = \lim_{z\to\infty} g^+(z)$, etc.) We define a measure μ on $\mathbb{R}$ as follows: for any $b \in \mathbb{R}$, let

$$\mu\big((-\infty, b]\big) = g^+(b) - g^+(-\infty). \tag{2}$$

It is easy to check that if $a \le b$, $\mu([a, b]) = g^+(b) - g^-(a)$. Note that this measure is finite with $\mu(\mathbb{R}) \le 2\overline{u}$.

Consider a typical time period. Let z be the true value of the parameter, and let y be an estimate of z. A best arm j^* is such that $g(z) = \eta_{j^*} + u_{j^*}z$. Given the estimate y, a greedy policy selects an arm j such that $g(y) = \eta_j + u_j y$. In particular, $\eta_j + u_j y \ge \eta_{j^*} + u_{j^*} y$, which implies that $\eta_{j^*} - \eta_j \le -(u_{j^*} - u_j)y$. Therefore, the instantaneous regret, which we denote by $r(z, y)$, can be bounded as follows:

$$r(z, y) = \eta_{j^*} + u_{j^*}z - \eta_j - u_j z \quad \le \quad (u_{j^*} - u_j)(z - y).$$

Since $g^-(y) \le u_j \le g^+(y)$ and $g^-(z) \le u_{j^*} \le g^+(z)$,

$$\begin{aligned} r(z, y) &\le \big(g^+(z \vee y) - g^-(z \wedge y)\big) \cdot |z - y| \\ &= \mu\big([z \wedge y, z \vee y]\big) \cdot |z - y|. \end{aligned} \tag{3}$$

Since $\mu(\mathbb{R}) \le 2\bar{u}$, given an estimate Y_t of Z, the instantaneous regret in period $t + 1$ is bounded above by

$$\mathsf{E}[r(z, Y_t) \mid Z = z] \quad \le \quad 2\overline{u}\mathsf{E}[|z - Y_t| \mid Z = z] \le \frac{c_4}{\sqrt{t}}$$

for some constant c_4, where the last inequality follows from Theorem 5. Since $\sum_{t=1}^{T-1} 1/\sqrt{t} \le 2\sqrt{T}$, it follows that the cumulative regret until time T is bounded above by

$$\text{Regret}(z, T, \textsc{Greedy}) \le 2\overline{u}|z| + 2c_4\sqrt{T},$$

where we use the fact that the instantaneous regret incurred in period 1 is bounded above by $2\overline{u}|z|$ because $Y_0 = 0$. This proves the upper bound on regret given in Theorem 6.

To establish the tightness result, we consider a problem instance with two arms with $(\eta_1, u_1) = (0, 1)$ and $(\eta_2, u_2) = (0, -1)$, as illustrated in Figure 2. Fix a policy Ψ and $T \ge 2$. Let $z_0 = 1/\sqrt{T}$. Also, let $\Pr_{z_0}\{\cdot\}$ and $\Pr_{-z_0}\{\cdot\}$ denote $\Pr\{\cdot \mid Z = z_0\}$ and $\Pr\{\cdot \mid Z = -z_0\}$, respectively. Then,

$$\begin{aligned} &\max\{\text{Regret}(z_0, T, \Psi), \text{Regret}(-z_0, T, \Psi)\} \\ &= 2z_0 \max\left\{\sum_{t=1}^{T} \Pr_{z_0}\{J_t = 2\}, \sum_{t=1}^{T} \Pr_{-z_0}\{J_t = 1\}\right\} \\ &\ge 2z_0 \sum_{t=1}^{T} \frac{1}{2}\Big(\Pr_{z_0}\{J_t = 2\} + \Pr_{-z_0}\{J_t = 1\}\Big), \end{aligned} \tag{4}$$

where the inequality follows from the fact that the maximum of two numbers is lower bound by their average. For this problem instance, we assume that the random variables E_ℓ^t have a standard normal distribution. We recognize the right-hand side in Eq. (4) as the Bayesian risk in a finite horizon Bayesian variant of our problem, where Z is equally likely to be z_0 or $-z_0$. This can be formulated as a (partially observable) dynamic programming problem whose information state is Y_t (because Y_t is a sufficient statistic, given past observations). Since we assume that the random variables E_ℓ^t have a standard normal distribution, the distribution of Y_t, given either value of Z, is always normal, with mean Z and variance $1/t$, independent of the sequence of actions taken. Thus, we are dealing with a problem in which actions do not affect the distribution of future information states; under these circumstances, a greedy policy that myopically maximizes the expected instantaneous reward at each step is optimal. Hence, it suffices to prove a lower bound for the right-hand side of Eq. (4) under the greedy policy.

By symmetry, under the greedy policy, we have

$$\begin{aligned} &2z_0 \sum_{t=1}^{T} \frac{1}{2}\Big(\Pr_{z_0}\{J_t = 2\} + \Pr_{-z_0}\{J_t = 1\}\Big) \\ &= \frac{2}{\sqrt{T}} \sum_{t=1}^{T} \Pr_{z_0}\{J_t = 2\} = \frac{2}{\sqrt{T}} \sum_{t=1}^{T} \Pr_{z_0}(Y_t < 0). \end{aligned}$$

Let W denote a standard normal random variable. Since $z_0 = 1/\sqrt{T}$, we have, for $t \le T$,

$$\begin{aligned} \Pr_{z_0}(Y_t < 0) = \Pr\left(z_0 + \frac{W}{\sqrt{t}} < 0\right) &= \Pr\left(W < -\frac{\sqrt{t}}{\sqrt{T}}\right) \\ \ge \Pr(W < -1) &\ge 0.15. \end{aligned}$$

It follows that $\text{Regret}(z_0, T, \textsc{Greedy}) \ge 0.3\sqrt{T}$. This implies that for any policy Ψ, there exists a value of Z (either z_0 or $-z_0$), for which $\text{Regret}(z, T, \Psi) \ge 0.3\sqrt{T}$.

We finally prove the last statement in Theorem 6. Fix some $z \in \mathbb{R}$, and let j^* be an optimal arm. There is a minimum distance $d > 0$ such that the greedy policy will pick an inferior arm $j \ne j^*$ in period $t+1$ only when our estimate Y_t differs from z by at least d (that is, $|z - Y_t| \ge d$). By Theorem 5, the expected number of times that we play an inferior arm j is bounded above by $\sum_{t=1}^{\infty} \Pr\{|z - Y_t| \ge d \mid Z = z\} \le 2\sum_{t=1}^{\infty}\left(e^{-f_1 td} + e^{-f_1 td^2}\right) < \infty$. Thus, the expected number of times that we select suboptimal arms is finite.

B. Bayes Risk Bounds: Proof of Theorem 7

We assume that the random variable Z is continuous, with a probability density function $p_Z(\cdot)$, which is bounded above by A. The argument below involves integrals with respect to the measure μ introduced in Section III-A (see Equation (2)).

Consider an arbitrary time $t + 1$ at which we make a decision based on the estimate Y_t computed at the end of period t. It follows from the upper bound on the instantaneous regret (see Equation (3)) that the instantaneous Bayes risk at time $t + 1$ is bounded above by $\mathsf{E}\big[(g^+(Z) - g^-(Y_t))(Z - Y_t)\mathbb{1}_{Y_t \le Z}\big] + \mathsf{E}\big[(g^+(Y_t) - g^-(Z))(Y_t - Z)\mathbb{1}_{Z \le Y_t}\big]$. We will derive a bound just on the term $\mathsf{E}[(g^+(Y_t) - g^-(Z))(Y_t - Z)\mathbb{1}_{Z \le Y_t}]$. The same bound is obtained for the other term,

through an identical argument. Since $\mu([a,b]) = g^+(b) - g^-(a)$ whenever $a \leq b$, we have

$$\begin{aligned}
&\mathsf{E}\big[(g^+(Y_t) - g^-(Z))(Y_t - Z)\mathbb{1}_{Z\leq Y_t}\big] \\
&\quad = \mathsf{E}\Big[\int_{q\in[Z,Y_t]} (Y_t - Z)\mathbb{1}_{Z\leq Y_t}\, d\mu(q)\Big] \\
&\quad = \mathsf{E}\Big[\int \mathbb{1}_{Z\leq q}\mathbb{1}_{Y_t\geq q}(Y_t - Z)\mathbb{1}_{Z\leq Y_t}\, d\mu(q)\Big] \\
&\quad = \int \mathsf{E}\big[\mathbb{1}_{Z\leq q}\mathbb{1}_{Y_t\geq q}(Y_t - Z)\big] d\mu(q).
\end{aligned}$$

The interchange of the integration and the expectation is justified by Fubini's Theorem, because $\mathbb{1}_{Z\leq q}\mathbb{1}_{Y_t\geq q}(Y_t - Z) \geq 0$. Though we omit the details of the argument, we can use Theorem 5 to show that for any $q \in \mathbb{R}$,

$$\mathsf{E}\big[\mathbb{1}_{Z\leq q}\mathbb{1}_{Y_t\geq q}(Y_t - Z)\big] \leq \frac{d_4}{t},$$

for some constant d_4 that depends only on the parameters $\underline{u}$, $\overline{u}$, b, and λ of Assumption 3.1. Since $\int d\mu(q) = \mu(\mathbb{R}) \leq 2\overline{u}$, it follows that the instantaneous Bayes risk incurred in period $t+1$ is at most $2\overline{u}d_4/t$. Then, the cumulative Bayes risk is bounded above by

$$\text{Risk}(T, \textsc{Greedy}) \leq 2\overline{u}\mathsf{E}\left[|Z|\right] + 2\overline{u}d_4 \ln T.$$

It remains to establish the tightness of our bound. We consider again the two-arm example of Figure 2, and also assume that Z is uniformly distributed on $[-2, 2]$. Consider an arbitrary time $t \geq 2$ and suppose that $z = 1/\sqrt{t}$, so that arm 1 is the best one. In the proof of Theorem 6, we have shown that the expected instantaneous regret in period t under the GREEDY policy is at least $0.30/\sqrt{t}$. A simple modification of this argument shows that for any z between $1/\sqrt{t}$ and $2/\sqrt{t}$, the expected instantaneous regret in period t is at least $d_5/\sqrt{t}$, where d_5 is a positive number (easily determined from the normal tables). Since $\Pr(1/\sqrt{t} \leq Z \leq 2/\sqrt{t}) = 1/(4\sqrt{t})$, we see that the instantaneous Bayes risk at time t is at least $d_5/(4t)$. Consequently, the cumulative Bayes risk satisfies

$$\text{Risk}(T, \textsc{Greedy}) \geq d_6 \ln T,$$

for some new numerical constant d_6. As we have argued in the proof of Theorem 6, for this two-arm problem instance, the greedy policy is optimal. It follows that that the lower bound we have established actually applies to all policies.

IV. Numerical Results

We summarize a numerical study comparing our greedy policy with a policy that assumes independent arms. We choose the well-known independent-arm multiarmed bandit policy of [1], to be referred to as "Lai87", which provides performance guarantees for a wide range of priors, including priors that allow for dependence between the arm rewards [1]. However, Lai87 policy tracks separate statistics for each arm, and thus does not take advantage of the known values of the coefficients η_ℓ and u_ℓ.

We consider problem instances with 5 arms, where all of the coefficients η_ℓ and u_ℓ are generated randomly and independently, according to a uniform distribution on $[-1, 1]$. We assume that the random variables E_ℓ^t are normally distributed, with mean zero and variance $\gamma_\ell^2 = 1$. For each of 5000 instances, we sample a value z from the standard normal distribution and compute arm rewards according to Equation (1) for $T = 100$ time periods. We compare the two policies with an oracle policy that knows the true value of z and always chooses the best arm. In Figure 3, we plot for the case $m = 5$, instantaneous rewards $X_{J_t}^t$ and per-period average cumulative regret, $\frac{1}{t}\sum_{s=1}^t (X_{oracle}^s - X_{J_s}^s)$, both averaged over the 5000 paths.

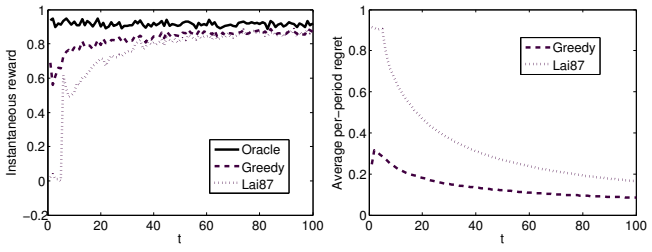

Fig. 3. Instantaneous rewards and per-period average cumulative regret for randomly generated problem instances with $m = 5$, averaged over 5000 paths. Differences between the policies in the right-hand plot are all significant at the 95% level.

We observe that the greedy policy appears to converge faster than Lai87, and we found the difference to be greater for larger m, supporting the insight from Theorem 7, that Bayes risk under our greedy policy is independent of m.

References

[1] T. L. Lai, "Adaptive treatment allocation and the multi-armed bandit problem," *Ann. Stat.*, vol. 15, no. 3, pp. 1091–1114, 1987.

[2] J. Gittins and D. M. Jones, "A dynamic allocation index for the sequential design of experiments," in *Progress in Statistics*, J. Gani, Ed. Amsterdam: North-Holland, 1974, pp. 241–266.

[3] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, pp. 4–22, 1985.

[4] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, pp. 285–294, 1933.

[5] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, vol. 58, pp. 527–535, 1952.

[6] D. Berry and B. Fristedt, *Bandit Problems: Sequential Allocation of Experiments*. London: Chapman and Hall, 1985.

[7] D. Feldman, "Contributions to the "two-armed bandit" problem," *Ann. Math. Stat.*, vol. 33, pp. 847–856, 1962.

[8] R. Keener, "Further contributions to the "two-armed bandit" problem," *Ann. Stat.*, vol. 13, no. 1, pp. 418–422, 1985.

[9] E. L. Pressman and I. N. Sonin, *Sequential Control With Incomplete Information*. London: Academic Press, 1990.

[10] J. Ginebra and M. K. Clayton, "Response surface bandits," *J. Roy. Stat. Soc. B*, vol. 57, no. 4, pp. 771–784, 1995.

[11] S. Pandey, D. Chakrabarti, and D. Agrawal, "Multi-armed bandit problems with dependent arms," in *Proceedings of the 24th International Conference on Machine Learning*, 2007.

[12] A. Tewari and P. L. Bartlett, "Optimistic linear programming gives logarithmic regret for irreducible MDPs," in *Advances in Neural Information Processing Systems 20*, 2008.

[13] R. Durrett, *Probability: Theory and Examples*. Belmont: Duxbury Press, 1996.

[14] M. Brezzi and T. L. Lai, "Incomplete learning from endogenous data in dynamic allocation," *Econometrica*, vol. 68, no. 6, pp. 1511–1516, 2000.

[15] S. A. Lippman, "On dynamic programming with unbounded rewards," *Management Sci.*, vol. 21, no. 11, pp. 1225–1233, 1975.