A Structured Multiarmed Bandit Problem and the Greedy Policy

Adam J. Mersereau, Paat Rusmevichientong, and John N. Tsitsiklis, *Fellow, IEEE*

***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 the greedy and optimal policies eventually coincide, and both settle on the best arm. This is 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 $\Omega(\log^2 T)$ established by Lai [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.**

***Index Terms*—Markov decision process (MDP).**

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 has garnered significant attention as 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.

Manuscript received February 28, 2008; revised July 05, 2008, December 23, 2008, and March 23, 2009. First published November 03, 2009; current version published December 09, 2009. This work was supported by the University of Chicago Graduate School of Business and the University of North Carolina Kenan-Flagler Business School, by the National Science Foundation through Grants DMS-0732196, CMMI-0746844, and ECCS-0701623. Recommended by Associate Editor R. Braatz.

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

P. Rusmevichientong is with the School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853 USA (e-mail: paatrus@cornell.edu).

J. N. Tsitsiklis is with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: jnt@mit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2009.2031725

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. The landmark result of Gittins and Jones [2], assuming an infinite horizon and discounted rewards, shows that an optimal policy always pulls the arm with the largest "index," where indices can be computed independently for each arm. In their seminal papers, Lai and Robbins [3] and Lai [1] further show that index-based approaches achieve asymptotically optimal performance in the finite horizon setting when the objective is to maximize total expected rewards.

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.

We study a fairly specific model that exemplifies a broader class of bandit problems where there is a known prior functional relationship among the arms' rewards. Our main thesis is that known statistical structure among arms can be exploited for higher rewards and faster convergence. Our assumed model is sufficient to demonstrate this thesis using a simple greedy approach in two settings: infinite horizon with discounted rewards, and finite horizon undiscounted rewards. In the discounted reward setting, we show that greedy and optimal policies eventually coincide, and both settle on the best arm in finite time. This differs from the classical multiarmed bandit case, where the Incomplete Learning Theorem [4]–[6] states that no policy is guaranteed to find the best arm. In the finite horizon setting, we show that the cumulative Bayes risk over T periods (defined below) under the greedy policy is bounded above by $O(\log T)$ and is independent of the number of arms. This is in contrast with the classical multiarmed bandit case where the risk over T periods is at least $\Omega(\log^2 T)$ (see [1]), and typically scales linearly with the number of arms. We outline our results and contributions in more detail in Section I-B.

Our formulation assumes that the mean reward of each arm is a linear function of an unknown scalar on which we have a prior

distribution. 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 throughout the paper 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 up 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, $J_t = \Psi_t(H_{t-1})$. We say that a policy is *optimal* if it maximizes the future total discounted reward, at every time and every possible history.

For every $T \geq 1$, we define the T-period cumulative *regret* under Ψ given $Z = z$, denoted by $\mathrm{Regret}(z, T, \Psi)$, as follows:

$$\mathrm{Regret}(z,T,\Psi) = \sum_{t=1}^{T} \mathsf{E}\left[\max_{\ell=1,\ldots,m} (\eta_\ell + u_\ell z) - (\eta_{J_t} + u_{J_t} z) \,|Z = z\right]$$

and the T-period cumulative Bayes risk of the policy Ψ by

$$\mathrm{Risk}(T,\Psi) = \mathsf{E}_Z\left[\mathrm{Regret}(Z,T,\Psi)\right].$$

We note that maximizing the expected total reward over a finite horizon is equivalent to minimizing Bayes risk.

Although this is not our focus, we point out an application of our model in the area of dynamic pricing with demand learning. Assume that we are sequentially selecting from a finite set $\mathcal{P}$ of prices with the objective of maximizing revenues over a horizon. When the price $p_\ell \in \mathcal{P}$ is selected at time t, we assume that sales S_ℓ^t are given by the linear demand curve

$$S_\ell^t = a - b p_\ell + \epsilon_\ell^t$$

where a is a known intercept but the slope b is unknown. The random variable ϵ_ℓ^t is a noise term with mean zero. The revenue is then given by $R_\ell^t = p_\ell S^t = a p_\ell - p_\ell^2 b - p_\ell \epsilon_\ell^t$, which is a special case of our model. We mention this dynamic pricing problem as an example application, though our model is more generally applicable to a range of control situations involving a linear function to be estimated. Example application domains include drug dosage optimization (see [7]), natural resource exploration, and target marketing.

A. Related Literature

As discussed in Section I, work on the multiarmed bandit problem has typically focused on the case where the arm rewards are assumed to be statistically independent. The literature can be divided into two streams based on the objective function: maximizing the expected total discounted reward over an infinite horizon and minimizing the cumulative regret or Bayes risk over a finite horizon. Our paper contributes to both streams of the literature. In the discounted, infinite horizon setting, the landmark result of Gittins and Jones [2] shows that an index-based policy is optimal under geometric discounting. Several alternative proofs of the so-called Gittins Index result exist (for example, [8]–[11]); see [12] for a summary and review. The classical Gittins assumptions do not hold in our version of the problem because statistical dependence among arms does not allow one to compute indices for each arm in a decomposable fashion. In the discounted setting, it is known (see [4]–[6]) that learning is incomplete when arms are independent. That is, an optimal policy has a positive probability of never settling on the best arm.

A second stream of literature has sought to maximize the expected total undiscounted reward or, equivalently, to minimize regret, defined as expected underperformance of a policy relative to the policy that knows and always chooses the best arm. A full characterization of an optimal policy given this objective appears to be difficult, and most authors have concerned themselves with rates of convergence of particular policies. The seminal work of Lai and Robbins [3] gives an asymptotic $\Omega(\log T)$ lower bound on the regret as a function of time. It also provides a policy based on "upper confidence bounds" on the arm rewards, whose performance asymptotically matches the lower bound. Lai [1] extends these results and shows, among other things, that in a Bayesian finite horizon setting, and under a fairly general set of assumptions, the cumulative Bayes risk must grow at least as fast as $\Omega(\log^2 T)$. Subsequent papers along this line include [13]–[15].

Interest in bandit problems under an assumption of dependent arms has a long history. Thompson [16], in what is widely regarded as the original paper on the multiarmed bandit problem, allows for correlation among arms in his initial formulation, though he only analyzes a special case involving independent arms. Robbins [17] formulates a continuum-armed bandit regression problem that subsumes our model, but does not provide an analysis of regret or risk. The formulation in Chapter 2 of [18] allows for correlation among arms (though most of the book concerns cases with independent arms). There has been relatively little analysis, however, of bandit problems with dependent arms. Two-armed bandit problems with two hidden states are considered in [19], [20]. A formulation with an arbitrary number of hidden states can be found in [21], along with a detailed analysis of the case with two hidden states. "Response surface bandits," multiarmed bandit problems whose arm rewards are linked through an assumed functional model, are formulated

in [22], and a simple tunable heuristic is proposed. Bandit problems where arm dependence is represented via a hierarchical model are studied in [23]. Although our model can be viewed as a special case of the formulation considered in [24], we obtain stronger results by exploiting the special structure of our model. Our regret and risk bounds (Theorems 3.1 and 3.2) are independent of the number of arms and apply to settings with infinite arms. In contrast, the regret bound in [24] scales linearly with the number of arms, and their model requires a finite number of arms. Seeking to maximizing average reward in an irreducible (but unknown) Markov decision process (MDP), [25] includes a policy that admits logarithmic regret but scales linearly with the number of actions and states of the underlying MDP.

Our model can be viewed as a special case of an online convex optimization problem, by considering randomized decisions. Let $\mathcal{S} = \{p \in [0,1]^m : \sum_{\ell=1}^m p_\ell = 1\}$ denote an m-dimensional simplex, where each $p = (p_1, \ldots, p_m) \in \mathcal{S}$ can be interpreted as the probabilities of playing the m arms. Given $Z = z$, the expected reward under a decision $p \in \mathcal{S}$ is given by $\sum_{\ell=1}^m p_\ell(\eta_\ell + u_\ell z)$, which is linear in p. For a bounded linear reward function on an m-dimensional decision space, [26] includes a policy whose cumulative regret over T periods is at most $O(m^{5/3}T^{2/3})$ (see also [27], [28]). This result has been generalized to convex cost functions (see [29], [30]), obtaining policies whose T-period regret is $O(m\ T^{3/4})$. Nearly all of the work in this area focuses on minimizing regret, and all known policies have regret that scales with the dimension of the problem space (corresponding to the number of arms m in our setting). By exploiting the specific structure of our reward function, however, we can get a stronger result and obtain a policy whose cumulative regret over T periods is only $O(\sqrt{T})$. Moreover, our regret bound is independent of m (Theorem 3.1 in Section III). We also consider the discounted reward and cumulative Bayes risk criteria.

We presented in Section I an application of our model to dynamic pricing with learning. Although this is not a focus of the current paper, we mention that there is a growing literature on this specific topic. See [31]–[34] for examples of recent work. All of these models are distinguished from ours in their objectives and in the specific demand and inventory situations treated.

B. Contributions and Organization

We view our main contributions to be 1) a model of statistical dependence among the arm rewards, 2) analysis of such a model under both expected discounted and undiscounted reward objectives, and 3) demonstration that prior knowledge of the statistical dependence of the different arms can improve performance and scalability. To the best of our knowledge, this is the first paper to provide detailed theoretical analysis of a multiarmed bandit model where the arm rewards are correlated through a continuous random variable with known prior distribution.

Section II includes our analysis of the infinite-horizon setting with geometric discounting. Theorem 2.1 establishes our main result on "complete learning." When every arm depends on the underlying random variable Z (that is, if $u_\ell \neq 0$ for all ℓ), the posterior mean of Z converges to its true value. We also show that a greedy decision is optimal when the variance of the posterior distribution is sufficiently small (Theorem 2.2). These two results together imply that eventually an optimal policy coincides with the greedy policy, and both settle on the best arm (Theorem 2.3). As mentioned previously, the latter result relies on the assumed correlation structure among the arms and is in contrast to the Incomplete Learning Theorem for the classical multiarmed bandit setting. We conclude Section II by examining the case where some of the coefficients u_ℓ are allowed to be zero. We argue that the corresponding arms can be interpreted as "retirement options," and prove that when retirement options are present, the optimal and greedy policies may never coincide, and that learning is generally incomplete.

In Section III, we analyze a similar greedy policy in the finite horizon setting, under the expected reward, or equivalently, cumulative Bayes risk criterion. We focus first on measuring the regret of the greedy policy. We show in Theorem 3.1 that the cumulative regret over T periods admits an $O(\sqrt{T})$ upper bound and that this bound is tight. Although this leads to an immediate $O(\sqrt{T})$ upper bound on the cumulative Bayes risk, we show that we can achieve an even smaller, $O(\log T)$, cumulative Bayes risk bound, under mild conditions on the prior distribution of Z (Theorem 3.2). The $O(\log T)$ risk bound is smaller than the known $\Omega(\log^2 T)$ lower bound of Lai [1], and we explain why our framework represents an exception to the assumptions required in [1]. Theorem 3.2 also shows that Bayes risk scales independently of the number of arms m. This result suggests that when the number of arms is large, we would expect significant benefits from exploiting the correlation structure among arms. Numerical experiments in Section IV support this finding.

II. Infinite Horizon With Discounted Rewards

In this section, we consider the problem of maximizing the total expected discounted reward. For any policy Ψ, the expected total discounted reward is defined as $\mathsf{E}[\sum_{t=1}^\infty \beta^t X_{J_t}^t]$, where $0 < \beta < 1$ denotes the discount factor and J_t denotes the arm chosen in period t under the policy Ψ. 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 $\mathsf{E}[Z^2] < \infty$. Furthermore, for every t and ℓ, we have $\mathsf{E}[E_\ell^t] = 0$ and $\gamma_\ell^2 := \mathsf{E}[(E_\ell^t)^2] < \infty$.

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 consequence of relaxing this assumption and allow some of the coefficients u_ℓ to be zero. Finally, since we focus on maximizing the expected reward, if the coefficient u_ℓ is the same for several arms, then we should only consider playing one with the largest value of η_ℓ, because it will give the maximum expected reward. Thus, Assumption 2.1(c) is introduced primarily to simplify our exposition.

In the next section, we show that "complete learning" is possible, under Assumption 2.1. In Theorem 2.1, we show that the posterior mean of Z converges to its true value, under any policy. This result then motivates us to consider in Section II-B a greedy policy that makes a myopic decision based only on the

current posterior mean of Z. We establish a sufficient condition for the optimality of the greedy policy (Theorem 2.2), and show that both the greedy and optimal policies eventually settle on the best arm, with probability one (Theorem 2.3). In contrast, when we allow some of the coefficients u_ℓ to be zero, it is possible for the greedy and optimal policies to disagree forever, with positive probability (Theorem 2.5 in Section II-C).

A. Complete Learning

Let us 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 be the posterior mean of Z, that is,

$$Y_t = \mathsf{E}[Z|\mathcal{F}_t]$$

and let V_t be the conditional variance, that is

$$V_t = \mathsf{E}\left[(Z - Y_t)^2|\mathcal{F}_t\right] = \mathsf{Var}(Z|\mathcal{F}_t).$$

The following result states that, under Assumption 2.1, we have complete learning, for every policy Ψ.

Theorem 2.1 (Complete Learning): 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 $\mathsf{E}[Z^2] < \infty$, 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 $\mathsf{E}[Z|\mathcal{F}_\infty]$, where $\mathcal{F}_\infty$ is the smallest σ-field containing $\mathcal{F}_t$ for all t (see [35]).

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

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

Then

$$\mathsf{Var}(\hat{Y}_t - Z) = \frac{1}{t^2}\sum_{\tau=1}^{t} \frac{\gamma^2_{J_\tau}}{u^2_{J_\tau}} \leq \frac{\max_\ell \left(\gamma^2_\ell/u^2_\ell\right)}{t}.$$

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.

Concerning the conditional variance, the definition of V_t implies that $V_t = \mathsf{E}[(Z - Y_t)^2|\mathcal{F}_t] = \mathsf{E}[Z^2|\mathcal{F}_t] - Y_t^2$, 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. Because uncertainty is univariate, we have complete learning under any policy, in contrast to the Incomplete Learning Theorem for the classical multiarmed bandit problems. As a consequence of Theorem 2.1, we will show in Section II-B (Theorem 2.3) that an optimal policy will settle on the best arm with probability one.

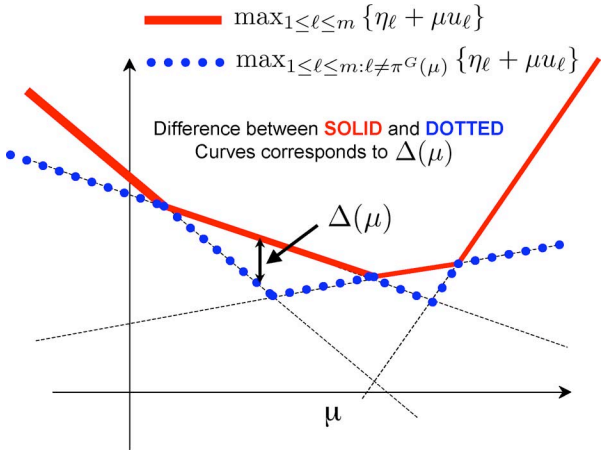

Fig. 1. Example of $\Delta(\cdot)$ with four arms.

B. A Greedy Policy

From Theorem 2.1, 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 $(J_1^G, J_2^G, \ldots)$ is defined by: for each $t \geq 1$

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

where $\{\mathcal{F}_t^G : t \geq 1\}$ 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 that difference between the expected 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 $\pi^G(\mu) = \arg\max_{\ell=1,\ldots,m}\{\eta_\ell + \mu u_\ell\}$. Fig. 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 Fig. 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 mean difference between the best and second best arms, then it is optimal to use a greedy policy. This result provides a sufficient condition for optimality of greedy decisions.

Theorem 2.2 (Optimality of Greedy Decisions): Under Assumption 2.1, there exists a constant δ that depends only on β

and the coefficients u_ℓ, with the following property. If we follow a policy Ψ until some time $t-1$, and if the sample path satisfies

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

then an optimal policy must make a greedy decision at time t. (Here, $\mathcal{F}_{t-1}$ is the σ-field generated by the history H_{t-1}. If h_{t-1} denotes the realized history up to the end of period $t-1$, then the above ratio is equal to$\Delta(\mathsf{E}[Z|H_{t-1}=h_{t-1}])/\sqrt{\mathsf{Var}[Z|H_{t-1}=h_{t-1}]}$.)

Proof: Let us fix a policy Ψ and some $t \geq 1$, and define $\mu_{t-1} = \mathsf{E}[Z|\mathcal{F}_{t-1}]$, which is the posterior mean of Z given the observations until the end of period $t-1$. Let J^* and R^* denote the greedy decision and the corresponding expected reward in period t, that is

$$\begin{aligned} J^* &= \arg\max_{\ell=1,\ldots,m} \{\eta_\ell + u_\ell \mathsf{E}[Z|\mathcal{F}_{t-1}]\} \\ &= \arg\max_{\ell=1,\ldots,m} \{\eta_\ell + u_\ell \mu_{t-1}\} \\ \text{and} \quad R^* &= \eta_{J^*} + u_{J^*}\mu_{t-1}. \end{aligned}$$

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 $M_s^G = \mathsf{E}[Z|\mathcal{F}_s^G]$ 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 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 (conditioned on $\mathcal{F}_{t-1}$) at each time $s \geq t$ is

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

where we first used Jensen's inequality, and then 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 (conditioned on $\mathcal{F}_{t-1}$) 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 $R_k = \eta_k + u_k \mathsf{E}[Z|\mathcal{F}_{t-1}] = \eta_k + u_k\mu_{t-1}$ denote the immediate expected reward in period t. The future expected 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

$$\begin{aligned} &R_k + \frac{\beta}{1-\beta}\mathsf{E}\left[\max_{\ell=1,\ldots,m}\{\eta_\ell + u_\ell Z\}|\mathcal{F}_{t-1}\right] \\ &\quad = R_k + \frac{\beta}{1-\beta} \\ &\qquad \times \mathsf{E}\left[\max_{\ell=1,\ldots,m}\{\eta_\ell + u_\ell\mu_{t-1} + u_\ell(Z-\mu_{t-1})\}|\mathcal{F}_{t-1}\right] \\ &\quad \leq R_k + \frac{\beta}{1-\beta}\max_{\ell=1,\ldots,m}\{\eta_\ell + u_\ell\mu_{t-1}\} \\ &\qquad + \frac{\beta}{1-\beta}\mathsf{E}\left[\max_{\ell=1,\ldots,m}\{u_\ell(Z-\mu_{t-1})\}|\mathcal{F}_{t-1}\right] \\ &\quad = R_k + \frac{\beta}{1-\beta}R^* + \frac{\beta}{1-\beta} \\ &\qquad \times \mathsf{E}\left[\max_{\ell=1,\ldots,m}\{u_\ell(Z-\mu_{t-1})\}|\mathcal{F}_{t-1}\right]. \end{aligned}$$

Note that

$$\begin{aligned} &\mathsf{E}\left[\max_{\ell=1,\ldots,m}\{u_\ell(Z-\mu_{t-1})\}|\mathcal{F}_{t-1}\right] \\ &\quad \leq \left(\max_{\ell=1,\ldots,m}|u_\ell|\right)\mathsf{E}\left[|Z-\mu_{t-1}||\mathcal{F}_{t-1}\right] \\ &\quad \leq \left(\max_{\ell=1,\ldots,m}|u_\ell|\right)\sqrt{\mathsf{E}\left[(Z-\mu_{t-1})^2|\mathcal{F}_{t-1}\right]} \\ &\quad = \left(\max_{\ell=1,\ldots,m}|u_\ell|\right)\sqrt{\mathsf{Var}(Z|\mathcal{F}_{t-1}).} \end{aligned}$$

Thus, under this policy the expected total discounted reward from time t onward is upper bounded by

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

Recall that the total expected discounted reward under the greedy policy is at least $R^*/(1-\beta)$. Moreover, for any arm $k \neq J^*$

$$\begin{aligned} \frac{R^*}{(1-\beta)} &= R^* + \frac{\beta}{1-\beta}R^* \\ &\geq (\eta_k + u_k\mu_{t-1}) + \Delta(\mu_{t-1}) + \frac{\beta}{1-\beta}R^* \\ &= R_k + \Delta(\mu_{t-1}) + \frac{\beta}{1-\beta}R^* \end{aligned}$$

where the inequality follows from the definition of $\Delta(\cdot)$. Thus, 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

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

which is the desired result. ■

As a consequence of Theorem 2.2, we can show that greedy and optimal policies both settle on the best arm with probability one.

Theorem 2.3: Under Assumption 2.1, if a policy is optimal, then it eventually agrees with the greedy policy, and both settle on the best arm with probability one.

Proof: Let J^* denote the best arm for each Z, that is, $J^* = \arg\max_\ell\{\eta_\ell + u_\ell Z\}$. Since Z is a continuous random variable and $\Delta(\cdot) = 0$ at finitely many points, we can assume that J^* is unique.

For the greedy policy, since $\mathsf{E}[Z|\mathcal{F}_t^G]$ converges to Z almost surely by Theorem 2.1, it follows that the greedy policy will eventually settle on the arm J^* with probability one.

Consider an optimal policy Π^*. Let $M_t^* = \mathsf{E}[Z|\mathcal{F}_t^*]$, where $\mathcal{F}_t^*$ denotes the filtration under Π^*. By Theorem 2.1, M_t^* converges to Z almost surely, and thus, $\Delta(M_t^*)$ converges to a positive number, almost surely. Also, $\mathsf{Var}(Z|\mathcal{F}_t^*)$ converges to zero

by Theorem 2.1. Thus, the condition in Theorem 2.2 is eventually satisfied, and thus Π^* eventually agrees with the greedy policy, with probability one. ∎

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

In this section, we explore the consequences of relaxing Assumption 2.1(b), and allow the coefficients u_ℓ to be zero, for some arms. We will show that, in contrast to Theorem 2.3, there is a positive probability that the greedy and optimal policies disagree forever. To demonstrate this phenomenon, we will restrict our attention to a setting where the underlying random variables Z and E_ℓ^t are normally distributed.

When Z and all E_ℓ^t are normal, the posterior distribution of Z remains normal. We can thus formulate the problem as a Markov Decision Process (MDP) whose state is characterized by $(\mu,\theta) \in \Re \times \Re_+$, where μ and θ denote the posterior mean and the inverse of the posterior variance, respectively. The action space is the set of arms. When we choose an arm ℓ at state (μ,θ), the expected reward is given by $r((\mu,\theta),\ell) = \eta_\ell + u_\ell\mu$. Moreover, the updated posterior mean $\mu'((\mu,\theta),\ell)$ and the inverse of the posterior variance $\theta'((\mu,\theta),\ell)$ are given by

$$\theta'((\mu,\theta),\ell) = \theta + \frac{u_\ell^2}{\gamma_\ell^2}$$

$$\mu'((\mu,\theta),\ell) = \frac{\theta}{\theta + u_\ell^2/\gamma_\ell^2}\mu + \frac{u_\ell^2/\gamma_\ell^2}{\theta + u_\ell^2/\gamma_\ell^2}\left(\frac{X_\ell^t - \eta_\ell}{u_\ell}\right)$$

where X_ℓ^t denotes the observed reward in period t. We note that these update formulas can also be applied when $u_\ell = 0$, yielding $\mu'((\mu,\theta),\ell) = \mu$ and $\theta'((\mu,\theta),\ell) = \theta$. The reward function $r((\cdot,\cdot),\cdot)$ in our MDP is unbounded because the state space is unbounded. However, as shown in the following lemma, there exists an optimal policy that is stationary. The proof of this result appears in Appendix A.

Lemma 2.4 (Existence of Stationary Optimal Policies): When the random variables Z and E_ℓ^t are normally distributed, then a deterministic stationary Markov policy is optimal; that is, there exists an optimal policy $\pi^* : \Re\times\Re_+ \to \{1,\ldots,m\}$ that selects a deterministic arm $\pi^*(\mu,\theta)$ for each state (μ,θ).

It follows from the above lemma that we can restrict our attention to stationary policies. If $u_\ell = 0$ for some ℓ, then there is a single such ℓ, by Assumption 2.1(c), and we can assume that it is arm $\ell = 1$. Since we restrict our attention to stationary policies, when arm 1 is played, the information state remains the same, and the policy will keep playing arm 1 forever, for an expected discounted reward of $\eta_1/(1-\beta)$. Thus, arm 1, with $u_1 = 0$, can be viewed as a "retirement option."

Note that in this setting a greedy policy $\pi^G : \Re \times \Re_+ \to \{1,\ldots,m\}$ is defined as follows: for every $(\mu,\theta) \in \Re \times \Re_+$:

$$\pi^G(\mu,\theta) = \arg\max_\ell\{\eta_\ell + u_\ell\mu\}$$

with ties broken arbitrarily. We have the following result.

Theorem 2.5 (Incomplete Learning): 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 $\mu \in \Re$, 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.

Proof: Under the assumption $\eta_1 > \max_{\ell:u_\ell\neq 0}\{\eta_\ell + u_\ell\mu\}$ for some $\mu \in \Re$, there is an open interval $I = (a,b)$ with $a < b$ such that whenever $\mu \in I$, the greedy policy must retire, that is, $\pi^G(\mu,\theta) = 1$ for all $\mu \in I$ and $\theta \in \Re_+$. Outside the closure of I, the greedy policy does not retire. Outside the closure of I, an optimal policy does not retire either because higher expected rewards are obtained by first pulling arm ℓ with $\eta_\ell + u_\ell\mu > \eta_1$. Without loss of generality, let us assume that $u_\ell > 0$ for some ℓ. A similar argument applies if we assume $u_\ell < 0$ for some ℓ.

Fix some $\epsilon \in (0, b-a)$, and let $A_\epsilon = (b-\epsilon, b)$ be an open interval at the right end of I. There exists a combination of sufficiently small ϵ_0 and θ_0 (thus, large variance) such that when we consider the set of states $(\mu,\theta) \in A_{\epsilon_0} \times (0,\theta_0)$, the expected long-run benefit of continuing exceeds the gain from retiring, as can be shown with a simple calculation. The set of states $A_{\epsilon_0} \times (0,\theta_0)$ will be reached with positive probability. When this happens, the greedy policy will retire. On the other hand, the optimal policy will choose to explore rather than retire.

Let M_t denote the posterior mean in period t under the optimal policy. We claim that once an optimal policy chooses to explore (that is, play an arm other than arm 1), there is a positive probability that *all* posterior means in future periods will exceed b, in which case the optimal policy never retires. To establish the claim, assume that $M_{t_0} > b$ for some t_0. Let τ be the stopping time defined as the first time after t_0 that $M_t \leq b$. We will show that $\Pr\{\tau = \infty\} > 0$, so that M_t stays outside I forever, and the optimal policy never retires.

Suppose, on the contrary, that $\Pr\{\tau < \infty\} = 1$. Since M_t is a square integrable martingale, it follows from the Optional Stopping Theorem that $\mathsf{E}[M_\tau] = M_{t_0}$, which implies that $b < M_{t_0} = \mathsf{E}[M_\tau] = \mathsf{E}[M_\tau; \tau < \infty] \leq b$, where the last inequality follows from the definition of τ. This is a contradiction, which establishes that $\Pr\{\tau = \infty\} > 0$, and therefore the greedy policy differs from the optimal one, with positive probability.

For the last part of the theorem, we wish to show that under either the optimal or the greedy policy, there is positive probability of retiring even though arm 1 is not the best arm. To establish this result, consider the interval $I' = (a + ((b-a)/3), a + (2(b-a)/3)) \subset I$, representing the middle third of the interval I. There exists θ' sufficiently large (thus, small variance) such that when we consider the states $(\mu,\theta) \in I' \times [\theta',\infty)$, the expected future gain from exploration is outweighed by the decrease in immediate rewards. These states are reached with positive probability, and at such states, the optimal policy will retire. The greedy policy also retires at such states because $I' \subseteq I$. At the time of retirement, however, there is positive probability that arm 1 is not the best one. ∎

We note that when $\eta_1 \leq \max_{\ell:u_\ell\neq 0}\{\eta_\ell + u_\ell\mu\}$ for all $\mu \in \Re$, one can verify that, as long as ties are never broken in favor of retirement, neither the greedy or the optimal policy will ever retire, so we can ignore the retirement option.

III. Finite Horizon With Total 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 Sec-

tion II, a simple greedy policy performs well in this setting. Before we proceed to the statement of the policy and its analysis, we introduce the following assumption on the coefficients u_ℓ associated with the arms 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 $x \geq 0$

$$\Pr\left(\left|E_\ell^t\right| \geq x\right) \leq be^{-\lambda x}.$$

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

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 this dependence will not be made explicit. The first part of Assumption 3.1 simply 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.

The second part of the 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. More sophisticated policies, with active experimentation, are needed in order to guarantee sublinear growth of the cumulative regret, but this topic lies outside the scope of this paper.

We will study the following variant of a greedy policy. It makes use of suboptimal (in the mean squared error sense) but simple estimators Y_t of Z, whose tail behavior is amenable to analysis. Indeed, it is not clear how to establish favorable regret bounds if we were to define Y_t as the posterior mean of Z.

GREEDY POLICY FOR FINITE HORIZON TOTAL UNDISCOUNTED REWARDS

Initialization: Set $Y_0 = 0$, representing our initial estimate of the value of Z.

Description: For periods $t = 1, 2, \ldots$

1) Sample arm J_t, where $J_t = \arg\max_{\ell=1,\ldots,m}\{\eta_\ell + Y_{t-1}u_\ell\}$, 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

$$Y_t = \frac{1}{t}\sum_{s=1}^{t}\frac{X_{J_s}^s - \eta_{J_s}}{u_{J_s}}.$$

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

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

Theorem 3.1: 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 $z \in \Re$ and $T \geq 1$

$$\mathrm{Regret}(z, T, \textsc{Greedy}) \leq c_1|z| + c_2\sqrt{T}.$$

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 $z \in \Re$ with

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

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

Let us comment on the relation and differences between the various claims in the statement of Theorem 3.1. The first claim gives an upper bound on the regret that holds for all z and T. The third claim states that for any fixed z, the cumulative regret is finite, but the finite asymptotic value of the regret can still depend on z. By choosing unfavorably the possible values of z (e.g., by letting $z = 1/\sqrt{T}$ or $z = -1/\sqrt{T}$, as in the proof in Section III-B), the regret can be made to grow as $\sqrt{t}$ for $t \leq T$, before it stabilizes to a finite asymptotic value, and this is the content of the second claim. We therefore see that the three claims characterize the cumulative regret in our problem under different regimes.

It is interesting to quantify the difference between the regret achieved by our greedy policy, which exploits the problem structure, and the regret under a classical bandit algorithm that assumes independent arms (see [36] and [37] for notions of relative or "external" regret). Theorem 3.1 shows that the cumulative regret of the greedy policy, for fixed z, is bounded. Lai and Robbins [3] establish a lower bound on the cumulative regret of any policy that assumes independent arms, showing that the regret grows as $\Omega(\log T)$. Thus, accounting for the problem structure in our setting results in a $\Omega(\log T)$ benefit. Similarly, the regret of our greedy policy scales independently of m, while typical independent-arm policies, such as UCB1 [15] or the policy of [1], sample each arm once. The difference in cumulative regret between the two policies thus grows linearly with m.

From the regret bound of Theorem 3.1, and by taking expectation with respect to Z, we obtain an easy upper bound on the cumulative Bayes risk, namely, $\mathrm{Risk}(T, \textsc{Greedy}) = O(\sqrt{T})$. 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 resulting 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-C.

Theorem 3.2: Under Assumption 3.1, if Z is a continuous random variable whose density function is bounded above by A, then there 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 $T \geq 1$

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

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 $T \geq 2$, and every policy Ψ

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

The above risk bound is smaller than the lower bound of $\Omega(\log^2 T)$ established by Lai [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 $\Pr\{X_\ell \in \cdot | X_i = x_i\}$ of X_ℓ given $X_i = x_i$ is degenerate, with all of its mass at a single point. In contrast, the $\Omega(\log^2 T)$ 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.

We finally note that our formulation and most of the analysis easily extends to a setting involving an infinite number of arms, as will be discussed in Section III-D.

A. Discussion of Assumption 3.1(a) and Implications on the Estimator

In this section, we reinterpret Assumption 3.1(a), and record its consequences on the tails of the estimators Y_t. Let $x_0 > 0$ be such that $b \leq e^{\lambda x_0}$. Then, Assumption 3.1(a) can be rewritten in the form

$$\Pr\left(\left|E_\ell^t\right| \geq x\right) \leq \min\left\{1, e^{-\lambda(x - x_0)}\right\}, \qquad \forall \ell, \forall x \geq 0.$$

Let U be an exponentially distributed random variable, with parameter λ, so that

$$\Pr(U + x_0 \geq x) = \min\left\{1, e^{-\lambda(x - x_0)}\right\}.$$

Thus

$$\Pr\left(\left|E_\ell^t\right| \geq x\right) \leq \Pr(U + x_0 \geq x)$$

which implies that each random variable E_ℓ^t is stochastically dominated by the shifted exponential random variable $U + x_0$; see [38] for the definition and properties of stochastic dominance.

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. The proof is given in Appendix B.

Theorem 3.3: 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 $z \in \Re$

$$\max\{\Pr(Y_t - z > a | Z = z), \Pr(Y_t - z < -a | Z = z)\} \leq e^{-f_1 t a} + e^{-f_1 t a^2}$$

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

B. Regret Bounds: Proof of Theorem 3.1

In this section, we will establish an upper bound on the regret, conditioned on any particular value z of Z, and the tightness of our regret bound. Consider a typical time period. Let z be the true value of the parameter, and let y be an estimate of z. The best arm j^* is such that $\eta_{j^*} + u_{j^*} z = \max_{\ell=1,\ldots,m} \eta_\ell + u_\ell z$ Given the estimate y, a greedy policy selects an arm j such that $\eta_j + u_j y = \max_{\ell=1,\ldots,m} \eta_\ell + u_\ell y$. In particular, $\eta_j + u_j y \geq \eta_{j^*} + u_{j^*} y$, which implies that $\eta_{j^*} - \eta_j \leq -(u_{j^*} - u_j) y$. Therefore, the instantaneous regret, due to choosing arm j instead of the best arm j^*, which we denote by $r(z, y)$, can be bounded as follows:

$$\begin{aligned} r(z, y) &= \eta_{j^*} + u_{j^*} z - \eta_j - u_j z \\ &\leq -(u_{j^*} - u_j) y + u_{j^*} z - u_j z \\ &= (u_{j^*} - u_j)(z - y) \leq 2\overline{u} |z - y| \end{aligned} \tag{2}$$

where the last inequality follows from Assumption 3.1.

At the end of period t, we have an estimate Y_t of Z. Then, the instantaneous regret in period $t + 1$ is given by

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

for some constant c_4, where the last inequality follows from Theorem 3.3. It follows that the cumulative regret until time T is bounded above by

$$\begin{aligned} \mathrm{Regret}(z, T, \textsc{Greedy}) &= \sum_{t=1}^{T} \mathsf{E}\left[r(z, Y_{t-1}) | Z = z\right] \\ &\leq 2\overline{u}|z| + c_4 \sum_{t=1}^{T-1} \frac{1}{\sqrt{t}} \\ &\leq 2\overline{u}|z| + 2c_4\sqrt{T} \end{aligned}$$

where the last inequality follows from the fact $\sum_{t=1}^{T-1} 1/\sqrt{t} \leq 2\sqrt{T}$. We also used 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 the regret given in Theorem 3.1.

To establish the first tightness result, we consider a problem instance with two arms, and parameters $(\eta_1, u_1) = (0, 1)$ and $(\eta_2, u_2) = (0, -1)$, as illustrated in Fig. 2. For this problem instance, we assume that the random variables E_ℓ^t have a standard normal distribution. Fix a policy Ψ and $T \geq 2$. Let $z_0 = 1/\sqrt{T}$. By our construction

$$\begin{aligned} &\max\{\mathrm{Regret}(z_0, T, \Psi), \mathrm{Regret}(-z_0, T, \Psi)\} \\ &\quad = 2z_0 \max\left\{\sum_{t=1}^{T} \Pr\{J_t = 2 | Z = z_0\}, \right. \\ &\qquad \left. \sum_{t=1}^{T} \Pr\{J_t = 1 | Z = -z_0\}\right\} \\ &\quad \geq 2z_0 \sum_{t=1}^{T} \frac{1}{2} \left(\Pr\{J_t = 2 | Z = z_0\} \right. \\ &\qquad \left. + \Pr\{J_t = 1 | Z = -z_0\}\right) \end{aligned} \tag{3}$$

where the inequality follows from the fact that the maximum of two numbers is lower bounded by their average. We recognize the right-hand side in (3) as the Bayes 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)

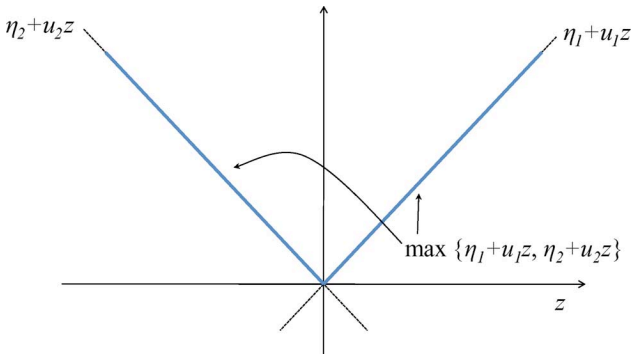

Fig. 2. Two-arm instance, with $(\eta_1, u_1) = (0, 1)$ and $(\eta_2, u_2) = (0, -1)$, used to prove the tightness result in Theorem 3.1.

dynamic programming problem whose information state is Y_t (because Y_t is a sufficient statistic, given past observations). [1]

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 (3) under the greedy policy.

Indeed, under the greedy policy, and using the symmetry of the problem, we have

$$\begin{aligned} &2z_0 \sum_{t=1}^{T} \frac{1}{2} \left(\Pr\{J_t = 2 | Z = z_0\} + \Pr\{J_t = 1 | Z = -z_0\} \right) \\ &= \frac{2}{\sqrt{T}} \sum_{t=1}^{T} \Pr\{J_t = 2 | Z = z_0\} \\ &= \frac{2}{\sqrt{T}} \sum_{t=1}^{T} \Pr(Y_t < 0 | Z = z_0). \end{aligned}$$

Since $z_0 = 1/\sqrt{T}$, we have, for $t \leq T$

$$\begin{aligned} \Pr(Y_t < 0 | Z = z_0) &= \Pr(z_0 + W/\sqrt{t} < 0) \\ &= \Pr\left(W < -\frac{\sqrt{t}}{\sqrt{T}} \right) \geq \Pr(W < -1) \\ &\geq 0.15 \end{aligned}$$

[1]The standard definition of the information state in this context is the posterior distribution of the "hidden" state Z, or, equivalently, the posterior probability $p_t = \Pr\{Z = z_0 | H_t\}$, where H_t denotes the history of the process until the end of period t. Let $\phi(\cdot)$ denote the density function of the standard normal random variable. The posterior probability p_t depends on H_t only through the likelihood ratio $\Pr\{H_t | Z = z_0\} / \Pr\{H_t | Z = -z_0\}$, which is equal to

$$\begin{aligned} \frac{\prod_{s=1}^{t} \phi\left(X_{J_s}^s - u_{J_s} z_0\right)}{\prod_{s=1}^{t} \phi\left(X_{J_s}^s + u_{J_s} z_0\right)} &= \prod_{s=1}^{t} \frac{\exp\left\{-\left(X_{J_s}^s - z_0/u_{J_s}\right)^2/2\right\}}{\exp\left\{-\left(X_{J_s}^s + z_0/u_{J_s}\right)^2/2\right\}} \\ &= e^{2z_0 \sum_{s=1}^{t} X_{J_s}^s / u_{J_s}} = e^{2 z_0 t Y_t} \end{aligned}$$

where the first equality follows from the fact that $u_{J_s} \in \{-1, +1\}$ for all s. Thus, p_t is completely determined by Y_t, so that Y_t can serve as an alternative information state.

where W is a standard normal random variable. It follows that $\mathrm{Regret}(z_0, T, \textsc{Greedy}) \geq 0.3\sqrt{T}$. This implies that for any policy Ψ, there exists a value of Z (either z_0 or $-z_0$), for which $\mathrm{Regret}(z, T, \Psi) \geq 0.3\sqrt{T}$.

We finally prove the last statement in Theorem 3.1. Fix some $z \in \Re$, and let j^* be an optimal arm. There is a minimum distance $d > 0$ such that the greedy policy will pick an inferior arm $j \neq j^*$ in period $t + 1$ only when our estimate Y_t differs from z by at least d (that is, $|z - Y_t| \geq d$). By Theorem 3.3, the expected number of times that we play an inferior arm j is bounded above by

$$\sum_{t=1}^{\infty} \Pr\{|z - Y_t| \geq d | Z = z\} \leq 2 \sum_{t=1}^{\infty} \left(e^{-f_1 t d} + e^{-f_1 t d^2} \right) < \infty.$$

Thus, the expected number of times that we select suboptimal arms is finite.

C. Bayes Risk Bounds: Proof of Theorem 3.2

We assume that the random variable Z is continuous, with a probability density function $p_Z(\cdot)$, which is bounded above by A. Let us first introduce some notation. We define a reward function $g : \Re \to \Re$, as follows: for every $z \in \Re$, 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

$$\begin{aligned} g^+(-\infty) &= g^-(-\infty) \\ &= \lim_{z \to -\infty} g^+(z) = \lim_{z \to -\infty} g^-(z) = \min_\ell u_\ell \end{aligned}$$

and

$$g^+(\infty) = g^-(\infty) = \lim_{z \to \infty} g^+(z) = \lim_{z \to \infty} g^-(z) = \max_\ell u_\ell.$$

Define a measure ρ on $\Re$ as follows: for any $b \in \Re$, let

$$\rho\left((-\infty, b]\right) = g^+(b) - g^+(-\infty). \tag{4}$$

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

Consider a typical time period. Let z be the true value of the parameter, and let y be an estimate of z. The greedy policy chooses the arm j such that $g(y) = \eta_j + u_j y$, while the true best arm j^* is such that $g(z) = \eta_{j^*} + u_{j^*} z$. We know from (2) that the instantaneous regret $r(z, y)$, due to choosing arm j instead of the best arm j^*, is bounded above by

$$\begin{aligned} r(z, y) &\leq (u_{j^*} - u_j)(z - y) \\ &\leq \left(g^+(z \vee y) - g^-(z \wedge y) \right) \cdot |z - y| \\ &= \rho\left([z \wedge y, z \vee y]\right) \cdot |z - y| \end{aligned}$$

where the second inequality follows from the fact that $g^-(y) \leq u_j \leq g^+(y)$ and $g^-(z) \leq u_{j^*} \leq g^+(z)$. The final equality follows from the definition of the measure ρ in (4).

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 above bound on the instantaneous regret that the instantaneous Bayes risk at time $t+1$ is bounded above by

$$\begin{aligned}&\mathsf{E}\left[\left(g^+(Z\vee Y_t)-g^-(Y_t\wedge Z)\right)\cdot|Z-Y_t|\right]\\&\quad=\mathsf{E}\left[\left(g^+(Z)-g^-(Y_t)\right)(Z-Y_t)\mathbb{1}_{Y_t\le Z}\right]\\&\qquad+\mathsf{E}\left[\left(g^+(Y_t)-g^-(Z)\right)(Y_t-Z)\mathbb{1}_{Z\le Y_t}\right].\end{aligned}$$

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 $\rho([a,b])=g^+(b)-g^-(a)$ whenever $a\le b$, we have

$$\begin{aligned}&\mathsf{E}\left[\left(g^+(Y_t)-g^-(Z)\right)(Y_t-Z)\mathbb{1}_{Z\le Y_t}\right]\\&\quad=\mathsf{E}\left[\int_{q\in[Z,Y_t]}(Y_t-Z)\mathbb{1}_{Z\le Y_t}d\rho(q)\right]\\&\quad=\mathsf{E}\left[\int\mathbb{1}_{Z\le q}\mathbb{1}_{Y_t\ge q}(Y_t-Z)\mathbb{1}_{Z\le Y_t}d\rho(q)\right]\\&\quad=\mathsf{E}\left[\int\mathbb{1}_{Z\le q}\mathbb{1}_{Y_t\ge q}(Y_t-Z)d\rho(q)\right]\\&\quad=\int\mathsf{E}\left[\mathbb{1}_{Z\le q}\mathbb{1}_{Y_t\ge q}(Y_t-Z)\right]d\rho(q).\end{aligned}$$

The interchange of the integration and the expectation is justified by Fubini's Theorem, because $\mathbb{1}_{Z\le q}\mathbb{1}_{Y_t\ge q}(Y_t-Z)\ge 0$. We will show that for any $q\in\Re$

$$\mathsf{E}\left[\mathbb{1}_{Z\le q}\mathbb{1}_{Y_t\ge q}(Y_t-Z)\right]\le\frac{d_4}{t}$$

for some constant d_4 that depends only on the parameters $\underline{u}$, $\overline{u}$, b, and λ of Assumption 3.1. Since $\int d\rho(q)=\rho(\Re)\le 2\overline{u}$, 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

$$\begin{aligned}\mathrm{Risk}(T,\textsc{Greedy})&\le 2\overline{u}\mathsf{E}\left[|Z|\right]+2\overline{u}d_4\sum_{t=1}^{T-1}\frac{1}{t}\\&\le 2\overline{u}\mathsf{E}\left[|Z|\right]+4\overline{u}d_4\ln T\end{aligned}$$

where the last inequality follows from the fact that $\sum_{t=1}^{T-1}1/t\le 2\ln T$.

Thus, it remains to establish an upper bound on $\mathsf{E}[\mathbb{1}_{Z\le q}\mathbb{1}_{Y_t\ge q}(Y_t-Z)]$. Without loss of generality (only to simplify notation and make the argument a little more readable), let us consider the case $q=0$. Using Theorem 3.3 and the fact that, for $v>0$, $(v+|Z|)^2\ge v^2+|Z|^2$ in the inequality below, we have

$$\begin{aligned}&\mathsf{E}\left[\mathbb{1}_{Z\le 0}\mathbb{1}_{Y_t\ge 0}(Y_t-Z)\right]\\&\quad=\mathsf{E}\left[\mathbb{1}_{Z\le 0}\mathbb{1}_{Y_t\ge 0}Y_t\right]+\mathsf{E}\left[\mathbb{1}_{Z\le 0}\mathbb{1}_{Y_t\ge 0}|Z|\right]\\&\quad=\mathsf{E}\left[\mathbb{1}_{Z\le 0}\mathsf{E}\left[\mathbb{1}_{Y_t\ge 0}Y_t|Z\right]\right]+\mathsf{E}\left[\mathbb{1}_{Z\le 0}|Z|\mathsf{E}\left[\mathbb{1}_{Y_t\ge 0}|Z\right]\right]\\&\quad=\mathsf{E}\left[\mathbb{1}_{Z\le 0}\int_0^\infty\Pr(Y_t>v|Z)dv\right]\\&\qquad+\mathsf{E}\left[\mathbb{1}_{Z\le 0}|Z|\Pr(Y_t\ge 0|Z)\right]\\&\quad\le\mathsf{E}\left[\mathbb{1}_{Z\le 0}\int_0^\infty e^{-f_1(v+|Z|)t}dv\right]\\&\qquad+\mathsf{E}\left[\mathbb{1}_{Z\le 0}\int_0^\infty e^{-f_1\left(v^2+|Z|^2\right)t}dv\right]\\&\qquad+\mathsf{E}\left[\mathbb{1}_{Z\le 0}|Z|e^{-f_1|Z|t}\right]\\&\qquad+\mathsf{E}\left[\mathbb{1}_{Z\le 0}|Z|e^{-f_1Z^2t}\right].\end{aligned}$$

We will now bound each one of the four terms, denoted by $C_1,\ldots,C_4$, in the right-hand side of the above inequality. We have

$$\begin{aligned}C_1&=\mathsf{E}\left[\mathbb{1}_{Z\le 0}e^{-f_1|Z|t}\right]\int_0^\infty e^{-f_1vt}dv\\&\le\int_0^\infty e^{-f_1vt}dv=\frac{1}{f_1t}.\end{aligned}$$

Furthermore

$$\begin{aligned}C_2&\le\int_{-\infty}^0 p_Z(z)e^{-f_1|z|^2t}dz\cdot\int_0^\infty e^{-f_1v^2t}dv\\&\le A\int_{-\infty}^0 e^{-f_1|z|^2t}dz\cdot\int_0^\infty e^{-f_1v^2t}dv=\frac{A\pi}{4f_1t}.\end{aligned}$$

For the third term, we have

$$C_3\le A\int_0^\infty ze^{-f_1zt}dz=\frac{A}{f_1^2t^2}\le\frac{A}{f_1^2t}.$$

Finally

$$C_4\le A\int_0^\infty ze^{-f_1z^2t}dz=\frac{A}{2f_1t}.$$

Since all of these bounds are proportional to $1/t$, our claim has been established, and the upper bound in Theorem 3.2 has been proved.

To complete the proof of Theorem 3.2, it remains to establish the tightness of our logarithmic cumulative risk bound. We consider again the two-arm example of Fig. 2(a), and also assume that Z is uniformly distributed on $[-2,2]$. Consider an arbitrary time $t\ge 2$ and suppose that $z=1/\sqrt{t}$, so that arm 1 is the best one. Under our assumptions, the estimate Y_t is normal with mean z and standard deviation $1/\sqrt{t}$. Thus, $\Pr(Y_t<0)=\Pr(z+W/\sqrt{t}<0)=\Pr(W<-1)\ge 0.15$, where W is a standard normal random variable. Whenever $Y_t<0$, the inferior arm 2 is chosen, resulting in an instantaneous regret of $2z=2/\sqrt{t}$. Thus, the expected instantaneous regret in period t is at least $0.30/\sqrt{t}$. A simple modification of the above 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 $\Pr(1/\sqrt{t}\le Z\le 2/\sqrt{t})=1/(4\sqrt{t})$, we see that

the instantaneous Bayes risk at time t is at least $d_5/(4t)$. Consequently, the cumulative Bayes risk satisfies

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

for some new numerical constant d_6.

For the particular example that we studied, it is not hard to show that the greedy policy is actually optimal: since the choice of arm does not affect the quality of the information to be obtained, there is no value in exploration, and therefore, the seemingly (on the basis of the available estimate) best arm should always be chosen. It follows that that the lower bound we have established actually applies to all policies.

D. The Case of Infinitely Many Arms

Our formulation generalizes to the case where we have infinitely many arms. Suppose that ℓ ranges over an infinite set, and define, as before, $g(z) = \sup_\ell\{\eta_\ell + u_\ell z\}$. We assume that the supremum is attained for every z. With this model, it is possible for the function g to be smooth. If it is twice differentiable, the measure μ is absolutely continuous and

$$g^+(b) - g^-(a) = \int_a^b g''(z)dz$$

where g'' is the second derivative of g. The proofs of the $O(\sqrt{T})$ and $O(\ln T)$ upper bounds in Theorems 3.1 and 3.2 apply without change, and lead to the same upper bounds on the cumulative regret and Bayes risk.

Recall that the lower bounds in Theorem 3.1 involve a choice of z that depends on the time of interest. However, when infinitely many arms are present, a stronger tightness result is possible involving a fixed value of z for which the regret is "large" for all times. The proof is given in Appendix C.

Proposition 3.4: For every $\epsilon \in (0, 1/2)$, there exists a problem instance, involving an infinite number of arms, and a value $z \in \Re$ such that for all $T \geq 2$

$$\text{Regret}(z, T, \text{GREEDY}) \geq \alpha(\epsilon) \cdot T^{0.5-\epsilon}$$

for some function $\alpha(\cdot)$.

IV. Numerical Results

We have explored the theoretical behavior of the regret and risk, under our proposed greedy policy, in Section III. We now summarize a numerical study intended to quantify its performance, compared with a policy that assumes independent arms. For the purposes of this comparison, we have chosen the well-known independent-arm multiarmed bandit policy of [1], to be referred to as "Lai87". We note that [1] provides performance guarantees for a wide range of priors, including priors that allow for dependence between the arm rewards. Lai87, however, 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 view Lai87 as an example of a policy that does not account for our assumed problem structure. We also implemented and tested a variant of the UCB1 policy of [15], modified slightly to account for the known arm variance in our problem. We found the performance of this UCB1-based policy to be substantially similar to Lai87. Our implementation of Lai87 is the same as the one in the original paper. In particular, Lai87 requires a scalar function h satisfying certain technical properties; we use the same h that was used in the original paper's numerical results.

We consider two sets of problem instances. In the first, 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 $m = 3, 5, 10$, we generate 5000 such problem instances. For each instance, we sample a value z from the standard normal distribution and compute arm rewards according to (1) for $T = 100$ time periods. In addition to the greedy policy and Lai87, we also compare with an optimistic benchmark, namely an oracle policy that knows the true value of z and always chooses the best arm. In Fig. 3, we plot for the case $m = 5$, instantaneous rewards $X_{J_t}^t$ and per-period average cumulative regret

$$\frac{1}{t}\sum_{s=1}^{t}\left(X_{oracle}^s - X_{J_s}^s\right)$$

both averaged over the 5000 paths. We include average cumulative regret plots for randomly-generated 3- and 10-arm problems in Fig. 4.

We observe that the greedy policy appears to converge faster than Lai87 in all three problem sets, with the difference being greater for larger m. This supports the insight from Theorem 3.2, that Bayes risk under our greedy policy is independent of m.

One practical advantage of the greedy policy over Lai87 can be seen in the left-hand plot of Fig. 5, which illustrates a randomly generated 5-arm problem instance included in our simulation. Each line in the graph represents the expected reward of an arm as a function of the unknown random variable Z. Thus the optimal arm for a given value of z is the maximum among these line segments. We observe that when arms are randomly generated according to our procedure, several arms can often be eliminated from consideration *a priori* because they will never achieve the maximum for any realization of z. The greedy policy will never choose such arms, though Lai87 may. On the other hand, recall that the greedy policy's performance is expected to depend on the constants $\overline{u}$ and $\underline{u}$ in Assumption 3.1, which measure the magnitude and relative sizes of the slopes u_ℓ. (For example, the proof of Theorem 3.1 indicates that the constants involved in the upper bound are proportional to $\overline{u}/\underline{u}$.) For randomly selected problems, there will be instances in which the worst-case ratio $\max_{k,\ell}|u_k/u_\ell|$ is large so that $\overline{u}/\underline{u}$ is also large, resulting in less favorable performance bounds.

The second set of problem instances is inspired by the dynamic pricing problem formulated in Section I. We assume that the sales S_ℓ^t at time t under the price p_ℓ are of the form $S_\ell^t = 2 - p_\ell b + \epsilon_\ell^t$, where b is normally distributed with mean $\mu = 1$ and standard deviation $\sigma = 0.25$. Thus, the revenue is $R_\ell^t = 2p_\ell - p_\ell^2 b + p_\ell\epsilon_\ell^t = (2p_\ell - p_\ell^2\mu) - \sigma p_\ell Z + p_\ell\epsilon_\ell^t$, where Z is a standard normal random variable. We also assume that the errors

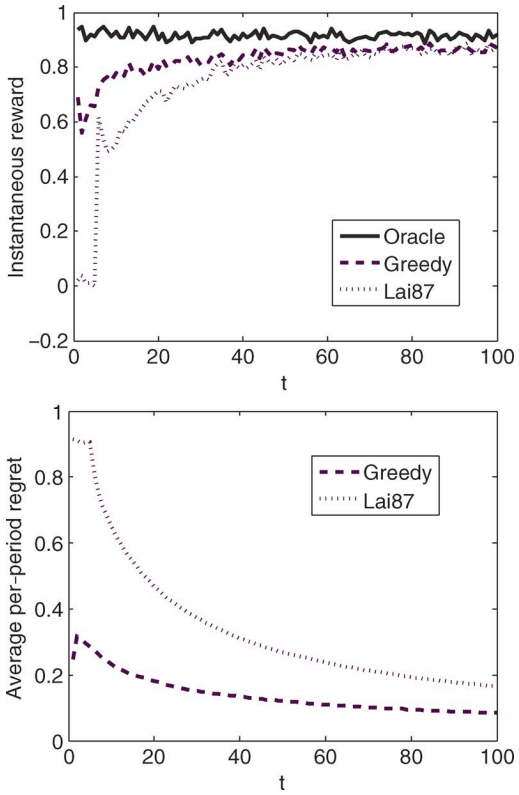

Fig. 3. Instantaneous rewards and per-period average cumulative regret for randomly generated problem instances with $m = 5$, averaged over 5000 paths. Differences between the policies in the right-hand plot are all significant at the 95% level.

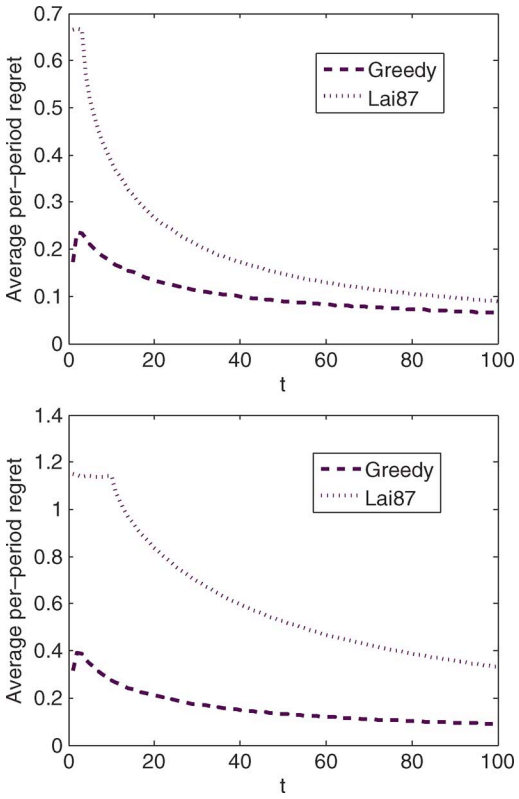

Fig. 4. Per-period average cumulative regret for randomly generated problem instances for $m = 3$ and $m = 10$. Differences between the policies in both plots are significant at the 95% level.

ϵ_ℓ^t are normally distributed with mean zero and variance 0.1. We set $m = 5$, corresponding to five prices: 0.75, 0.875, 1.0, 1.125, 1.25. The expected revenue as a function of z for each of the five arms/prices is shown in the right-hand side plot of Fig. 5. We see that in this instance, in contrast to the randomly generated instance in the left-hand side plot, every arm is the optimal arm for some realization of z.

We simulate 5000 runs, each involving a different value z, sampled from the standard normal distribution, and we apply each one of our three policies: greedy, Lai87, and oracle. Fig. 6 gives the instantaneous rewards and per-period average cumulative regret, both averaged over the 5000 runs. Inspection of Fig. 6 suggests that the greedy policy performs even better relative to Lai87 in the dynamic pricing example than in the randomly generated instances. Our greedy policy is clearly better able to take advantage of the inherent structure in this problem.

V. Discussion and Future Research

We conclude by highlighting our main findings. We have removed the typical assumption made when studying multiarmed bandit problems, that the arms are statistically independent, by considering a specific statistical structure underlying the mean rewards of the different arms. This setting has allowed us to demonstrate our main conjecture, namely, that one can take advantage of known correlation structure and obtain better performance than if independence were assumed. At the same time, we have specific results on the particular problem, with univariate uncertainty, that we have considered. Within this setting, simple greedy policies perform well, independent of the number of arms, for both discounted and undiscounted objectives.

We believe that our paper opens the door to development of a comprehensive set of policies that account for correlation structures in multiarmed bandit and other learning problems. While correlated bandit arms are plausible in a variety of practical settings, many such settings require a more general problem setup than we have considered here. Of particular interest are correlated bandit policies for problems with multivariate uncertainty and with more general correlation structures.

Appendix A
Proof of Lemma 2.4

Proof: We will show that there exists a dominating function $w : \Re \times \Re_+ \to \Re_+$ such that

$$\sup_{(\mu,\theta)\in\Re\times\Re_+} \frac{|\max_\ell r(\mu,\theta,\ell)|}{w(\mu,\theta)} < \infty$$

and for each $(\mu,\theta) \in \Re \times \Re_+$

$$\max_\ell \mathsf{E}\left[w\left(\mu'(\mu,\theta,\ell),\theta'(\mu,\theta,\ell)\right)\right] \leq w(\mu,\theta) + \frac{\max_i |u_i|}{\min\left\{|u_i/\gamma_i| : u_i \neq 0\right\}}.$$

The desired result then follows immediately from Theorem 1 in [39].

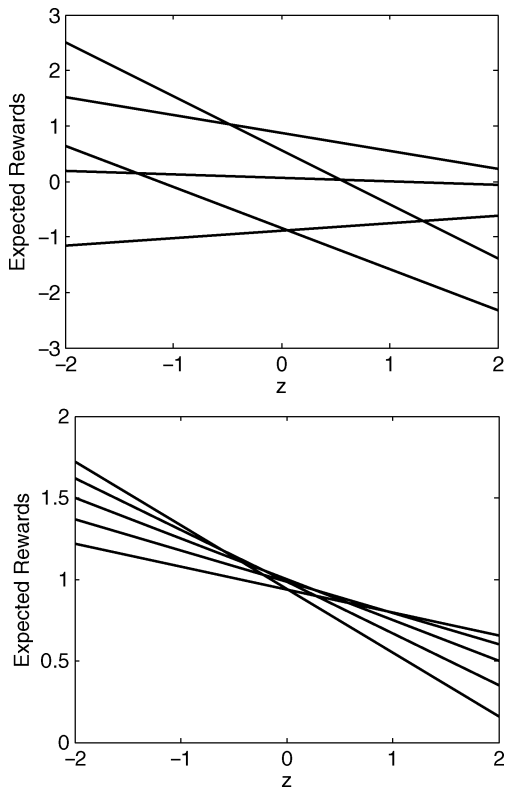

Fig. 5. Mean reward of each arm as a function of z for a randomly generated problem (left) and a dynamic pricing problem (right).

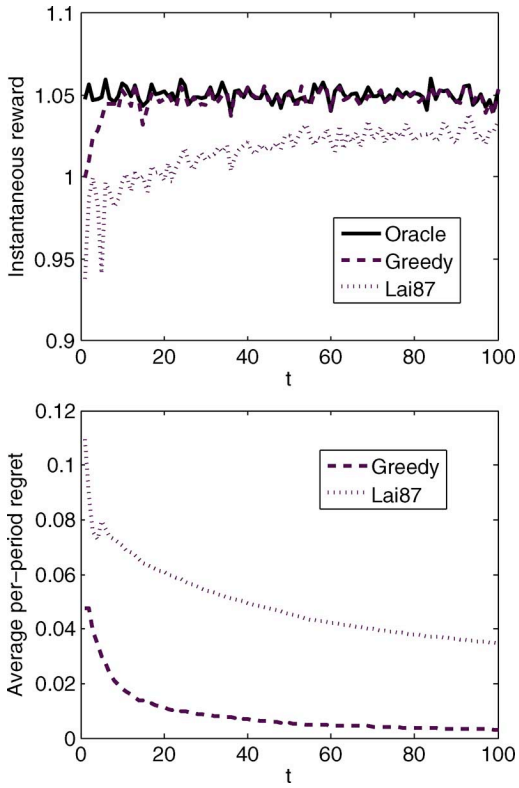

Fig. 6. Per-period average cumulative regret for the dynamic pricing problem with 5 candidate prices. Differences between the policies are significant at the 95% level.

Let $\bar{\eta} = \max_\ell |\eta_\ell|$ and $\bar{u} = \max_\ell |u_\ell|$. For each $(\mu, \theta) \in \Re \times \Re$, let

$$w(\mu, \theta) = \bar{\eta} + \bar{u}\left(|\mu| + \frac{1}{\sqrt{\theta}}\right).$$

The first condition is clearly satisfied because for each state $(\mu, \theta) \in \Re \times \Re_+$ and arm ℓ

$$\frac{|r(\mu, \theta, \ell)|}{w(\mu, \theta)} \leq \frac{|\eta_\ell| + |u_\ell||\mu|}{w(\mu, \theta)} \leq 1.$$

To verify the second condition, note that if $u_\ell = 0$, then $\mu'(\mu, \theta, \ell) = \mu$ and $\theta'(\mu, \theta, \ell) = \theta$, with probability one and the inequality is trivially satisfied. So, suppose that $u_\ell \neq 0$. It follows from the definition of $w(\cdot, \cdot)$ that

$$\begin{aligned}
&\mathsf{E}\left[w\left(\mu'(\mu, \theta, \ell), \theta'(\mu, \theta, \ell)\right)\right] \\
&\quad = \bar{\eta} + \bar{u}\mathsf{E}\left|\mu'(\mu, \theta, \ell)\right| + \frac{\bar{u}}{\sqrt{\theta'(\mu, \theta, \ell)}} \\
&\quad = \bar{\eta} + \bar{u}\mathsf{E}\left|\mu'(\mu, \theta, \ell)\right| + \frac{\bar{u}}{\sqrt{\theta + u_\ell^2/\gamma_\ell^2}} \\
&\quad \leq \bar{\eta} + \bar{u}\mathsf{E}\left|\mu'(\mu, \theta, \ell)\right| + \frac{\bar{u}}{\min\left\{|u_i/\gamma_i| : u_i \neq 0\right\}}
\end{aligned}$$

To establish the desired result, it thus suffices to show that $\mathsf{E}|\mu'(\mu, \theta, \ell)| \leq |\mu| + 1/\sqrt{\theta}$.

Since $X_\ell^t = \eta_\ell + u_\ell Z + E_\ell^t$, $Z \sim \mathcal{N}(\mu, 1/\theta)$, and $E_\ell^t \sim \mathcal{N}(0, \gamma_\ell^2)$, it follows that $(X_\ell^t - \eta_\ell)/u_\ell$ has the same distribution as $\mu + W\sqrt{(1/\theta) + (\gamma_\ell^2/u_\ell^2)}$, where W is an independent standard normal random variable. It follows from the definition of $\mu'((\mu, \theta), \ell)$ that

$$\begin{aligned}
\mathsf{E}\left|\mu'(\mu, \theta, \ell)\right| &= \mathsf{E}\left|\mu + W\frac{\left(u_\ell^2/\gamma_\ell^2\right)\sqrt{\frac{1}{\theta} + \frac{\gamma_\ell^2}{u_\ell^2}}}{\theta + u_\ell^2/\gamma_\ell^2}\right| \\
&= \mathsf{E}\left|\mu + W\sqrt{\frac{u_\ell^2/\gamma_\ell^2}{\theta^2 + \theta u_\ell^2/\gamma_\ell^2}}\right| \leq |\mu| + \frac{1}{\sqrt{\theta}}
\end{aligned}$$

where the second equality follows from the fact that

$$\frac{\frac{u_\ell^2}{\gamma_\ell^2}\sqrt{\frac{1}{\theta} + \frac{\gamma_\ell^2}{u_\ell^2}}}{\theta + u_\ell^2/\gamma_\ell^2} = \frac{u_\ell^2/\gamma_\ell^2}{\sqrt{\theta\left(u_\ell^2/\gamma_\ell^2\right)\left(\theta + u_\ell^2/\gamma_\ell^2\right)}} = \sqrt{\frac{u_\ell^2/\gamma_\ell^2}{\theta^2 + \theta u_\ell^2/\gamma_\ell^2}}$$

and the inequality follows from the facts that $\mathsf{E}|W| = \sqrt{2/\pi} \leq 1$ and

$$\sqrt{\frac{u_\ell^2/\gamma_\ell^2}{\theta^2 + \theta u_\ell^2/\gamma_\ell^2}} = \frac{1}{\sqrt{\theta}}\sqrt{\frac{u_\ell^2/\gamma_\ell^2}{\theta + u_\ell^2/\gamma_\ell^2}} \leq \frac{1}{\sqrt{\theta}}$$

■

APPENDIX B
PROOF OF THEOREM 3.3

Proof: Fix some ℓ, and let V be a random variable with the same distribution as E_ℓ^t. For any $s \in \Re$, let $g_\ell(s) = \mathsf{E}[e^{sV}]$.

Note that $g_\ell(0) = 1$. Because of the exponential tails assumption on V, the function $g_\ell(s)$ is finite, and in fact infinitely differentiable, whenever $|s| < \lambda$. Furthermore, its first derivative g' satisfies $g'_\ell(0) = \mathsf{E}[V] = 0$. Finally, its second derivative satisfies

$$g''_\ell(s) = \frac{d^2}{ds^2}\mathsf{E}[e^{sV}] = \mathsf{E}\left[V^2 e^{sV}\right].$$

(This step involves an interchange of differentiation and integration, which is known to be legitimate in this context.)

It is well known that when a random variable $|V|$ is stochastically dominated by another random variable W, we have $\mathsf{E}[f(|V|)] \leq \mathsf{E}[f(W)]$, for any nonnegative nondecreasing function f. In our context, this implies that

$$\begin{aligned} g''_\ell(s) &= \mathsf{E}[V^2 e^{sV}] \leq \mathsf{E}\left[V^2 e^{|s|\cdot|V|}\right] \\ &\leq \mathsf{E}\left[(U+x_0)^2 e^{|s|}(U+x_0)\right]. \end{aligned}$$

The function on the right-hand side above is completely determined by b and x_0. It is finite, continuous, and bounded on the interval $s \in [-b/2, b/2]$ by some constant f_0 which only depends on b and x_0. It then follows that:

$$g_\ell(s) \leq 1 + \frac{f_0^2}{2}\cdot s^2$$

for $-b/2 \leq s \leq b/2$, for all ℓ.

We use the definition of Y_t and the relation $X_j^t = \eta_j + u_j Z$, to express Y_t in the form

$$Y_t = Z + \frac{1}{t}\sum_{\tau=1}^{t}\frac{E_{J_\tau}^\tau}{u_{J_\tau}}.$$

Let $D_\tau = E_{J_\tau}^\tau / u_{J_\tau}$. We will now use the standard Chernoff bound method to characterize the tails of the distribution of the sum $\sum_{\tau=1}^t D_\tau$.

Let $\sigma(Z, H_t)$ be the σ-field generated by Z and H_t, and note that J_t is $\sigma(Z, H_{t-1})$-measurable. Let also $Q_t(s) = \mathsf{E}[e^{sD_t}|\sigma(Z, H_{t-1})]$. We observe that

$$Q_t(s) \leq \max_{\ell=1,\ldots,m}\mathsf{E}\left[e^{sE_\ell^t/u_\ell}\right] \leq 1 + \frac{f_0^2}{2\underline{u}^2}\cdot s^2 = 1 + f_3 s^2 \quad (5)$$

for $-b/2 \leq s \leq b/2$, where the last equality is taken as the definition of f_3.

Now, note that

$$\mathsf{E}\left[\frac{e^{s(D_1+\cdots+D_t)}}{Q_1(s)\cdots Q_t(s)}|\sigma(Z, H_{t-1})\right] = \frac{e^{s(D_1+\cdots+D_{t-1})}}{Q_1(s)\cdots Q_{t-1}(s)}.$$

Since $\sigma(Z, H_{t-2}) \subseteq \sigma(Z, H_{t-1})$, it follows from the tower property that:

$$\begin{aligned} &\mathsf{E}\left[\frac{e^{s(D_1+\cdots+D_t)}}{Q_1(s)\cdots Q_t(s)}|\sigma(Z, H_{t-2})\right] \\ &\quad = \mathsf{E}\left[\frac{e^{s(D_1+\cdots+D_{t-1})}}{Q_1(s)\cdots Q_{t-1}(s)}|\sigma(Z, H_{t-2})\right] \\ &\quad = \frac{e^{s(D_1+\cdots+D_{t-2})}}{Q_1(s)\cdots Q_{t-2}(s)}. \end{aligned}$$

Repeated applications of the above argument show that

$$\mathsf{E}\left[\frac{e^{s(D_1+\cdots+D_t)}}{Q_1(s)\cdots Q_t(s)}|Z\right] = 1.$$

Using the bound from (5), we obtain

$$E\left[e^{s(D_1+\cdots+D_t)}|Z\right] \leq (1+f_3 s^2)^t$$

for $-b/2 \leq s \leq b/2$.

Fix some $t \geq 1$, $a > 0$, and $z \in \Re$. We have, for any $s \in [-b/2, b/2]$

$$\begin{aligned} \Pr(Y_t - z > a|Z = z) &= \Pr(D_1 + \cdots + D_t > ta|Z = z) \\ &\leq e^{-sta}\mathsf{E}\left[e^{s(D_1+\cdots+D_t)}|Z = z\right] \\ &\leq e^{-sta}(1+f_3 s^2)^t \\ &= e^{-sta}e^{t\ln(1+f_3 s^2)} \\ &\leq e^{-sta}e^{f_3 s^2 t}. \end{aligned} \quad (6)$$

Suppose first that a satisfies $a \geq bf_3$. By applying inequality (6) with $s = b/2$, we obtain

$$\begin{aligned} \Pr(Y_t - z > a|Z = z) &\leq e^{-(tba/2)+(tf_3 b^2/4)} \\ &\leq e^{-(tba/2)+(tba/4)} = e^{-tba/4}. \end{aligned}$$

Suppose next that a satisfies $a < bf_3$. By applying inequality (6) with $s = a/(2f_3) < b/2$, we obtain

$$\begin{aligned} \Pr(Y_t - z > a|Z = z) &\leq e^{-(ta^2/2f_3)+\left(tf_3 a^2/4f_3^2\right)} \\ &= e^{-ta^2/4f_3}. \end{aligned}$$

Since for every positive value of a one of the above two bounds applies, we have

$$\Pr(Y_t - z > a|Z = z) \leq e^{-f_1 at} + e^{-f_1 a^2 t}$$

where $f_1 = \min\{b/4, 1/4f_3\}$. The expression $\Pr(Y_t - z < -a|Z = z)$ can be bounded by a symmetrical argument, and the proof of the tail bounds is complete.

The bounds on the moments of Y_t follow by applying the formula $\mathsf{E}[X^2] = 2\int_0^\infty x\Pr(X > x)dx - 2\int_{-\infty}^0 x\Pr(X < x)dx$, which implies that

$$\begin{aligned} \mathsf{E}\left[(Y_t - z)^2|Z = z\right] &\leq 4\int_0^\infty x\left(e^{-f_1 tx} + e^{-f_1 tx^2}\right)dx \\ &= \frac{4}{f_1^2 t^2} + \frac{2}{f_1 t} \leq \left(\frac{4}{f_1^2} + \frac{2}{f_1}\right)\frac{1}{t} \end{aligned}$$

and it follows from the Jensen Inequality that:

$$\begin{aligned} \mathsf{E}\left[|Y_t - z||Z = z\right] &\leq \sqrt{\mathsf{E}\left[(Y_t - z)^2|Z = z\right]} \\ &\leq \sqrt{\frac{4}{f_1^2} + \frac{2}{f_1}}\cdot\frac{1}{\sqrt{t}} \end{aligned}$$

which is the desired result. ∎

APPENDIX C
PROOF OF PROPOSITION 3.4

Proof: We fix some $\epsilon > 0$. Recall that the maximum expected reward function g is defined by $g(z) = \max_\ell\{\eta_\ell + u_\ell z\}$,

where the maximization ranges over all arms in our collection. Consider a problem with an infinite number of arms where the function $g(\cdot)$ is given by

$$g(z) = \begin{cases} -z, & \text{if } z < 0 \\ z + \frac{z^{1+\epsilon}}{1+\epsilon}, & \text{if } 0 \le z \le 1 \\ 2z - \frac{\epsilon}{1+\epsilon}, & \text{if } 1 < z. \end{cases}$$

Note that the function g is convex and continuous, and its derivative is given by

$$g'(z) = \begin{cases} -1, & \text{if } z < 0 \\ 1 + z^{\epsilon}, & \text{if } 0 \le z \le 1 \\ 2, & \text{if } 1 < z. \end{cases}$$

In particular, $\underline{u} = 1$ and $\overline{u} = 2$. We assume that for each $a \in \Re$, the error E_a^t associated with arm $a \in \Re$ is normally distributed with mean zero and variance $(g'(a))^2$; then Assumption 3.1 is satisfied. We will consider the case where $z = 0$ and show that the cumulative regret over T periods is $\Omega(T^{0.5-\epsilon})$.

Consider our estimate Y_t of z at the end of period t, which is normal with zero mean and variance $1/t$. In particular, $\sqrt{t}Y_t$ is a standard normal random variable. If $Y_t < 0$, then the arm chosen in period $t+1$ is the best one, and the instantaneous regret in that period is zero. On the other hand, if $0 \le Y_t \le 1$, then the arm a chosen in period $t+1$ will be for which the line $\eta_a + u_a z$ is the tangent of the function $g(\cdot)$ at Y_t, given by

$$h_{t+1}(x) = g'(Y_t)x - \frac{\epsilon Y_t^{1+\epsilon}}{1+\epsilon}$$

where the choice of the intercept is chosen so that $h_{t+1}(Y_t) = g(Y_t)$. If $Y_t > 1$, the instantaneous regret can only be worse than if $0 \le Y_t \le 1$. This implies that the instantaneous regret incurred in period $t+1$ satisfies

$$\begin{aligned} r(z, Y_t) &\ge \mathbb{1}(0 < Y_t < 1)\{g(z) - h_{t+1}(z)\} \\ &= \mathbb{1}(0 < Y_t < 1)\frac{\epsilon Y_t^{1+\epsilon}}{1+\epsilon} \\ &\ge \mathbb{1}(0 < Y_t < 1/\sqrt{t})\frac{\epsilon Y_t^{1+\epsilon}}{1+\epsilon} \end{aligned}$$

where the equality follows from the fact that $z = 0$. Therefore, the instantaneous regret incurred in period $t+1$ is lower bounded by

$$\begin{aligned} &\mathsf{E}[r(z, Y_t)|Z = z] \\ &\quad \ge \frac{\epsilon}{1+\epsilon}\mathsf{E}\left[\mathbb{1}\left(0 < Y_t < \frac{1}{\sqrt{t}}\right) \cdot Y_t^{1+\epsilon}|Z = z\right] \\ &\quad = \frac{\epsilon}{(1+\epsilon)t^{(1+\epsilon)/2}}\mathsf{E}\left[\mathbb{1}\left(0 < Y_t < \frac{1}{\sqrt{t}}\right) \cdot (\sqrt{t}Y_t)^{1+\epsilon}|Z = z\right] \\ &\quad = \frac{\epsilon}{(1+\epsilon)t^{(1+\epsilon)/2}}\mathsf{E}\left[\mathbb{1}(0 < W < 1) \cdot W^{1+\epsilon}\right] \end{aligned}$$

where W is a standard normal random variable. Therefore, the cumulative regret over T periods can be lower bounded as

$$\begin{aligned} \text{Regret}(z, T, \text{GREEDY}) &\ge \sum_{t=1}^{T-1} \mathsf{E}[r(z, Y_t)|Z = z] \\ &= \frac{\epsilon \mathsf{E}\left[\mathbb{1}(0 < W < 1) \cdot W^{1+\epsilon}\right]}{1+\epsilon} \\ &\quad \times \sum_{t=1}^{T-1} \frac{1}{t^{(1+\epsilon)/2}} \\ &= \Omega\left(T^{(1-\epsilon)/2}\right) \end{aligned}$$

where the last inequality follows, for example, by approximating the sum by an integral. ■

ACKNOWLEDGMENT

The authors would like to thank H. Lopes, G. Samorodnitsky, M. Todd, and R. Zeithammer for insightful discussions on problem formulations and analysis.

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, The Netherlands: 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] M. Rothschild, "A two-armed bandit theory of market pricing," *J. Econ. Theory*, vol. 9, pp. 185–202, 1974.
[5] J. S. Banks and R. K. Sundaram, "Denumerable-armed bandits," *Econometrica*, vol. 60, pp. 1071–1096, 1992.
[6] M. Brezzi and T. L. Lai, "Incomplete learning from endogenous data in dynamic allocation," *Econometrica*, vol. 68, no. 6, pp. 1511–1516, 2000.
[7] T. L. Lai and H. Robbins, "Adaptive design in regression and control," *Proc. Natl. Acad. Sci.*, vol. 75, no. 2, pp. 586–587, 1978.
[8] P. Whittle, "Multi-armed bandits and the Gittins index," *J. Royal Stat. Soc. B*, vol. 42, pp. 143–149, 1980.
[9] R. R. Weber, "On the Gittins index for multiarmed bandits," *Ann. Prob.*, vol. 2, pp. 1024–1033, 1992.
[10] J. N. Tsitsiklis, "A short proof of the Gittins index theorem," *Ann. Appl. Probab.*, vol. 4, no. 1, pp. 194–199, 1994.
[11] D. Bertsimas and J. Niño-Mora, "Conservation laws, extended polymatroids and multi-armed bandit problems," *Math. Oper. Res.*, vol. 21, pp. 257–306, 1996.
[12] E. Frostig and G. Weiss, "Four Proofs of Gittins' Multiarmed Bandit Theorem," Univ. Haifa, Haifa, Israel, 1999.
[13] R. Agrawal, D. Teneketzis, and V. Anantharam, "Asymptotically efficient adaptive allocation schemes for controlled markov chains: Finite parameter space," *IEEE T. Autom. Control*, vol. AC-34, no. 12, pp. 1249–1259, Dec. 1989.
[14] R. Agrawal, "Sample mean based index policies with O(log n) regret for the multi-armed bandit problem," *Adv. Appl. Prob.*, vol. 27, pp. 1054–1078, 1996.
[15] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, pp. 235–256, 2002.
[16] 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.
[17] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, vol. 58, pp. 527–535, 1952.
[18] D. Berry and B. Fristedt, *Bandit Problems: Sequential Allocation of Experiments*. London, U.K.: Chapman and Hall, 1985.
[19] D. Feldman, "Contributions to the 'two-armed bandit' problem," *Ann. Math. Stat.*, vol. 33, pp. 847–856, 1962.

[20] R. Keener, "Further contributions to the 'two-armed bandit' problem," *Ann. Stat.*, vol. 13, no. 1, pp. 418–422, 1985.
[21] E. L. Pressman and I. N. Sonin, *Sequential Control With Incomplete Information*. London, U.K.: Academic Press, 1990.
[22] J. Ginebra and M. K. Clayton, "Response surface bandits," *J. Roy. Stat. Soc. B*, vol. 57, no. 4, pp. 771–784, 1995.
[23] S. Pandey, D. Chakrabarti, and D. Agrawal, "Multi-armed bandit problems with dependent arms," in *Proc. 24th Int. Conf. Mach. Learning*, 2007, pp. 721–728.
[24] T. L. Graves and T. L. Lai, "Asymptotically efficient adaptive choice of control laws in controlled Markov chains," *SIAM J. Control Optim.*, vol. 35, no. 3, pp. 715–743, 1997.
[25] A. Tewari and P. L. Bartlett, "Optimistic linear programming gives logarithmic regret for irreducible MDPs," in *Proc. Adv. Neural Inform. Processing Syst. 20*, 2008, pp. 1505–1512.
[26] R. Kleinberg, "Online linear optimization and adaptive routing," *J. Computer and System Sciences*, vol. 74, no. 1, pp. 97–114, 2008.
[27] H. B. McMahan and A. Blum, "Online geometric optimization in the bandit setting against an adaptive adversary," in *Proc. 17th Annu. Conf. Learning Theory*, 2004, pp. 109–123.
[28] V. Dani and T. Hayes, "Robbing the bandit: Less regret in online geometric optimization against an adaptive adversary," in *Proc. 17th Ann. ACM-SIAM Symp. Discrete Algorithms*, 2006, pp. 937–943.
[29] R. Kleinberg, "Nearly tight bounds for the continuum-armed bandit problem," in *Proc. Adv. Neural Inform. Processing Syst. 17*, 2005, pp. 697–704.
[30] A. Flaxman, A. Kalai, and H. B. McMahan, "Online convex optimization in the bandit setting: Gradient descent without a gradient," in *Proc. 16th Ann. ACM-SIAM Symp. Discrete Algorithms*, 2005, pp. 385–394.
[31] A. Carvalho and M. Puterman, "How Should a Manager Set Prices When the Demand Function is Unknown?" Univ. British Columbia, Vancouver, BC, Canada, 2004.
[32] Y. Aviv and A. Pazgal, "Dynamic Pricing of Short Life-Cycle Products Through Active Learning," Olin School Business, Washington Univ., St. Louis, MO, 2005.
[33] V. F. Farias and B. Van Roy, "Dynamic Pricing With a Prior on Market Response," Stanford Univ., Stanford, CA, 2006.
[34] O. Besbes and A. Zeevi, "Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms," *Oper. Res.*, to be published.
[35] R. Durrett, *Probability: Theory and Examples*. Belmont, CA: Duxbury Press, 1996.
[36] D. P. Foster and R. Vohra, "Regret in the on-line decision problem," *Games Econ. Beh.*, vol. 29, pp. 7–35, 1999.
[37] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, 2002.
[38] M. Shaked and J. G. Shanthikumar, *Stochastic Orders*. New York: Springer, 2007.
[39] S. A. Lippman, "On dynamic programming with unbounded rewards," *Manag. Sci.*, vol. 21, no. 11, pp. 1225–1233, 1975.

Adam J. Mersereau received the B.S.E. degree (with highest honors) in electrical engineering from Princeton University, Princeton, NJ, in 1996, and the Ph.D. degree in operations research from the Massachusetts Institute of Technology, Cambridge, in 2003.

He is currently Assistant Professor with the Kenan-Flagler Business School, University of North Carolina. His research interests include information-sensitive dynamic optimization with applications to operational and marketing systems. Current projects concern multiarmed bandit problems and decomposition techniques for approximating large-scale Markov decision problems.

Paat Rusmevichientong received the B.A. degree in mathematics from University of California, Berkeley, in 1997 and the M.S. and Ph.D. degrees in operations research from Stanford University, Stanford, CA, in 1999 and 2003, respectively.

He is an Assistant Professor in the School of Operations Research and Information Engineering, Cornell University, Ithaca, NY. His research interests include data mining, information technology, and nonparametric algorithms for stochastic optimization problems, with applications to supply chain and revenue management.

John N. Tsitsiklis (F'99) received the B.S. degree in mathematics and the B.S., M.S., and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge, in 1980, 1980, 1981, and 1984, respectively.

He is currently a Clarence J. Lebel Professor with the Department of Electrical Engineering, MIT. He has served as a Codirector of the MIT Operations Research Center from 2002 to 2005, and in the National Council on Research and Technology in Greece (2005 to 2007). His research interests are in systems, optimization, communications, control, and operations research. He has coauthored four books and more than a hundred journal papers in these areas.

Dr. Tsitsiklis received the Outstanding Paper Award from the IEEE Control Systems Society (1986), the M.I.T. Edgerton Faculty Achievement Award (1989), the Bodossakis Foundation Prize (1995), and the INFORMS/CSTS Prize (1997). He is a member of the National Academy of Engineering. Finally, in 2008, he was conferred the title of Doctor honoris causa, from the Université Catholique de Louvain.