MATHEMATICS OF OPERATIONS RESEARCH
Vol. 35, No. 2, May 2010, pp. 395–411
ISSN 0364-765X | EISSN 1526-5471 | 10 | 3502 | 0395

DOI 10.1287/moor.1100.0446
© 2010 INFORMS

Linearly Parameterized Bandits

Paat Rusmevichientong
School of Operations Research and Information Engineering, Cornell University, Ithaca, New York 14853,
paatrus@cornell.edu

John N. Tsitsiklis
Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge,
Massachusetts 02139, jnt@mit.edu

We consider bandit problems involving a large (possibly infinite) collection of arms, in which the expected reward of each arm is a linear function of an r-dimensional random vector $\mathbf{Z} \in \mathbb{R}^r$, where $r \geq 2$. The objective is to minimize the cumulative regret and Bayes risk. When the set of arms corresponds to the unit sphere, we prove that the regret and Bayes risk is of order $\Theta(r\sqrt{T})$, by establishing a lower bound for an arbitrary policy, and showing that a matching upper bound is obtained through a policy that alternates between exploration and exploitation phases. The phase-based policy is also shown to be effective if the set of arms satisfies a strong convexity condition. For the case of a general set of arms, we describe a near-optimal policy whose regret and Bayes risk admit upper bounds of the form $O(r\sqrt{T}\log^{3/2} T)$.

Key words: multi-armed bandit; parametric model; adaptive control
MSC2000 subject classification: Primary: 93E35; secondary: 62L12, 62F35
OR/MS subject classification: Primary: dynamic programming and optimal control, applications; secondary: statistics, estimation
History: Received January 19, 2009; revised January 20, 2010. Published online in *Articles in Advance* April 30, 2010.

1. Introduction. Since its introduction by Thompson [31], the multi-armed bandit problem has served as an important model for decision making under uncertainty. Given a set of arms with unknown reward profiles, the decision maker must choose a sequence of arms to maximize the expected total payoff, where the decision in each period may depend on the previously observed rewards. The multi-armed bandit problem elegantly captures the trade-off between the need to exploit arms with high payoff and the incentive to explore previously untried arms for information gathering purposes.

Much of the previous work on the multi-armed bandit problem assumes that the rewards of the arms are statistically independent (see, for example, Lai and Robbins [23], Lai [22]). This assumption enables us to consider each arm separately, but it leads to policies whose regret scales linearly with the number of arms. Most policies that assume independence require each arm to be tried at least once, and are impractical in settings involving many arms. In such settings, we want a policy whose regret is independent of the number of arms.

When the mean rewards of the arms are assumed to be independent random variables, Lai and Robbins [23] show that the regret under an arbitrary policy must increase linearly with the number of arms. However, the assumption of independence is quite strong in practice. In many applications, the information obtained from pulling one arm can change our understanding of other arms. For instance, in marketing applications, we expect a priori that similar products should have similar sales. By exploiting the correlation among products and (arms), we should be able to obtain a policy whose regret scales more favorably than traditional bandit algorithms that ignore correlation and assume independence.

Mersereau et al. [24] propose a simple model that demonstrates the benefits of exploiting the underlying structure of the rewards. They consider a bandit problem where the expected reward of each arm is a linear function of an unknown scalar, with a known prior distribution. Because the reward of each arm depends on a single random variable, the mean rewards are perfectly correlated. They prove that, under certain assumptions, the cumulative Bayes risk over T periods (defined below) under a greedy policy admits an $O(\log T)$ upper bound, independent of the number of arms.

In this paper, we extend the model of Mersereau et al. [24] to the setting where the expected reward of each arm depends linearly on a *multivariate* random vector $\mathbf{Z} \in \mathbb{R}^r$. We concentrate on the case where $r \geq 2$, which is fundamentally different from the previous model because the mean rewards now depend on more than one random variable, and thus, they are no longer perfectly correlated. The bounds on the regret and Bayes risk and the policies found in Mersereau et al. [24] no longer apply. To give a flavor for the differences, we will show that, in our model, the cumulative Bayes risk under an arbitrary policy is at least $\Omega(r\sqrt{T})$, which is significantly higher than the upper bound of $O(\log T)$ attainable when $r = 1$.

The linearly parameterized bandit is an important model that has been studied by many researchers, including (Ginebra and Clayton [16], Abe and Long [1], Auer [4]). The results in this paper complement and extend the earlier and independent work of Dani et al. [12] in a number of directions. We provide a detailed comparison between our work and the existing literature in §§1.3 and 1.4.

1.1. The model. We have a compact set $\mathcal{U}_r \subset \mathbb{R}^r$ that corresponds to the set of arms, where $r \geq 2$. The reward X_t^u of playing arm $\mathbf{u} \in \mathcal{U}_r$ in period t is given by

$$X_t^u = \mathbf{u}'\mathbf{Z} + W_t^u, \tag{1}$$

where $\mathbf{u}'\mathbf{Z}$ is the inner product between the vector $\mathbf{u} \in \mathcal{U}_r$ and the random vector $\mathbf{Z} \in \mathbb{R}^r$. We assume that the random variables W_t^u are independent of each other and of $\mathbf{Z}$. Moreover, for each $\mathbf{u} \in \mathcal{U}_r$, the random variables $\{W_t^u\colon t \geq 1\}$ are identically distributed, with $\mathsf{E}[W_t^u] = 0$ for all t and $\mathbf{u}$. We allow the error random variables W_t^u to have unbounded support, provided that their moment generating functions satisfy certain conditions (given in Assumption 1). Each vector $\mathbf{u} \in \mathcal{U}_r$ simultaneously represents an arm and determines the expected reward of that arm. So, when it is clear from the context, we will interchangeably refer to a $\mathbf{u} \in \mathcal{U}_r$ as either a vector or an arm.

Let us introduce the following conventions and notation that will be used throughout the paper. We denote vectors and matrices in bold. All vectors are column vectors. For any vector $\mathbf{v} \in \mathbb{R}^r$, its transpose is denoted by $\mathbf{v}'$, and is always a row vector. Let $\mathbf{0}$ denote the zero vector, and for $k = 1, \ldots, r$, let $\mathbf{e}_k = (0, \ldots, 1, \ldots, 0)$ denote the standard unit vector in $\mathbb{R}^r$, with a one in the kth component and a zero elsewhere. Also, let $\mathbf{I}_k$ denote the $k \times k$ identity matrix. We let $\mathbf{A}'$ and $\det(\mathbf{A})$ denote the transpose and determinant of $\mathbf{A}$, respectively. If $\mathbf{A}$ is a symmetric positive semidefinite matrix, then $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ denote the smallest and the largest eigenvalues of $\mathbf{A}$, respectively. We use $\mathbf{A}^{1/2}$ to denote its symmetric nonnegative definite square root, so that $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$. If $\mathbf{A}$ is also positive definite, we let $\mathbf{A}^{-1/2} = (\mathbf{A}^{-1})^{1/2}$. For any vector $\mathbf{v}$, $\|\mathbf{v}\| = \sqrt{\mathbf{v}'\mathbf{v}}$ denotes the standard Euclidean norm, and for any positive definite matrix $\mathbf{A}$, $\|\mathbf{v}\|_{\mathbf{A}} = \sqrt{\mathbf{v}'\mathbf{A}\mathbf{v}}$ denotes a corresponding weighted norm. For any two symmetric positive semidefinite matrices $\mathbf{A}$ and $\mathbf{B}$, we write $\mathbf{A} \leq \mathbf{B}$ if the matrix $\mathbf{B} - \mathbf{A}$ is positive semidefinite. Also, all logarithms $\log(\cdot)$ in the paper denote the natural log, with base e. A random variable is denoted by an uppercase letter and its realized values are denoted in lowercase.

For any $t \geq 1$, let $\mathcal{H}_{t-1}$ denote the set of possible histories until the end of period $t-1$. A policy $\psi = (\psi_1, \psi_2, \ldots)$ is a sequence of functions such that $\psi_t\colon \mathcal{H}_{t-1} \to \mathcal{U}_r$ selects an arm in period t based on the history until the end of period $t-1$. For any policy ψ and $\mathbf{z} \in \mathbb{R}^r$, the T-period cumulative *regret* under ψ given $\mathbf{Z} = \mathbf{z}$, denoted by $\text{Regret}(\mathbf{z}, T, \psi)$, is defined by

$$\text{Regret}(\mathbf{z}, T, \psi) = \sum_{t=1}^{T} \mathsf{E}\left[\max_{\mathbf{v} \in \mathcal{U}_r} \mathbf{v}'\mathbf{z} - \mathbf{U}_t'\mathbf{z} \mid \mathbf{Z} = \mathbf{z}\right],$$

where for any $t \geq 1$, $\mathbf{U}_t \in \mathcal{U}_r$ is the arm chosen under ψ in period t. Because $\mathcal{U}_r$ is compact, $\max_{\mathbf{v} \in \mathcal{U}_r} \mathbf{v}'\mathbf{z}$ is well defined for all $\mathbf{z}$. The T-period cumulative Bayes risk under ψ is defined by

$$\text{Risk}(T, \psi) = \mathsf{E}[\text{Regret}(\mathbf{Z}, T, \psi)],$$

where the expectation is taken with respect to the prior distribution of $\mathbf{Z}$. We aim to develop a policy that minimizes the cumulative regret and Bayes risk. We note that minimizing the T-period cumulative Bayes risk is equivalent to maximizing the expected total reward over T periods.

To facilitate exposition, when we discuss a particular policy, we will drop the superscript and write X_t and W_t to denote $X_t^{U_t}$ and $W_t^{U_t}$, respectively, where $\mathbf{U}_t$ is the arm chosen by the policy in period t. With this convention, the reward obtained in period t under a particular policy is simply $X_t = \mathbf{U}_t'\mathbf{Z} + W_t$.

1.2. Potential applications. Although our paper focuses on a theoretical analysis, we mention briefly potential applications to problems in marketing and revenue management. Suppose we have m arms indexed by $\mathcal{U}_r = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m\} \subset \mathbb{R}^r$. For $k = 1, \ldots, r$, let $\boldsymbol{\phi}_k = (u_{1,k}, u_{2,k}, \ldots, u_{m,k}) \in \mathbb{R}^m$ denote an m-dimensional column vector consisting of the kth coordinates of the different vectors $\mathbf{u}_l$. Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ be the column vector consisting of expected rewards, where μ_l denotes the expected reward of arm $\mathbf{u}_l$. Under our formulation, the vector $\boldsymbol{\mu}$ lies in an r-dimensional subspace spanned by the vectors $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_r$, that is, $\boldsymbol{\mu} = \sum_{k=1}^{r} Z_k \boldsymbol{\phi}_k$, where $\mathbf{Z} = (Z_1, \ldots, Z_r)$. If each arm corresponds to a product to be offered to a customer, we can then interpret the vector $\boldsymbol{\phi}_k$ as a feature vector or basis function, representing a particular characteristic of the products such as price or popularity. We can then interpret the random variables $Z_1, \ldots, Z_r$ as regression coefficients, obtained from approximating the vector of expected rewards using the basis functions $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_r$, or more intuitively, as the weights associated with the different characteristics. Given a prior on the coefficients Z_k, our goal is to choose a sequence of products that gives the maximum expected total reward. This representation suggests that our model might be applicable to problems where we can approximate high-dimensional vectors using a linear combination of a few basis functions, an approach that has been successfully applied to high-dimensional dynamic programming problems (see Bertsekas and Tsitsiklis [8] for an overview).

1.3. Related literature. The multi-armed bandit literature can be divided into two streams, depending on the objective function criteria: maximizing the total discounted reward over an infinite horizon, or minimizing the cumulative regret and Bayes risk over a finite horizon. Our paper focuses exclusively on the second criterion. Much of the work in this area focuses on understanding the rate with which the regret and risk under various policies increase over time. In their pioneering work, Lai and Robbins [23] establish an asymptotic lower bound of $\Omega(m \log T)$ for the T-period cumulative regret for bandit problems with m independent arms whose mean rewards are well-separated, where the difference between the expected reward of the best and second best arms is fixed and bounded away from zero. They further demonstrate a policy whose regret asymptotically matches the lower bound. In contrast, our paper focuses on problems where the number of arms is large (possibly infinite), and where the gap between the maximum expected reward and the expected reward of the second best arm can be arbitrarily small. Lai [22] extends these results to a Bayesian setting, with a prior distribution on the reward characteristics of each arm. He shows that when we have m arms, the T-period cumulative Bayes risk is of order $\Theta(m \log^2 T)$, when the prior distribution has a continuous density function satisfying certain properties (see Theorem 3 in Lai [22]). Subsequent papers along this line include (Agrawal et al. [3], Agrawal [2], Auer et al. [5]).

There has been relatively little research, however, on policies that exploit the dependence among the arms. Thompson [31] allows for correlation among arms in his initial formulation, though he only analyzes a special case involving independent arms. Robbins [28] formulates a continuum-armed bandit regression problem, but does not provide an analysis of the regret or risk. Berry and Fristedt [6] allow for dependence among arms in their formulation in Chapter 2, but mostly focus on the case of independent arms. Feldman [14] and Keener [19] consider two-armed bandit problems with two hidden states, where the rewards of each arm depend on the underlying state that prevails. Pressman and Sonin [27] formulate a general multi-armed bandit problem with an arbitrary number of hidden states, and provide a detailed analysis for the case of two hidden states. Pandey et al. [25] study bandit problems where the dependence of the arm rewards is represented by a hierarchical model.

A somewhat related literature on bandits with dependent arms is the recent work by Wang et al. [32, 33] and Goldenshluger and Zeevi [17, 18] who consider bandit problems with two arms, where the expected reward of each arm depends on an exogenous variable that represents side information. These models, however, differ from ours because they assume that the side information variables are independent and identically distributed over time, and moreover, these variables are *perfectly observed before* we choose which arm to play. In contrast, we assume that the underlying random vector $\mathbf{Z}$ is unknown and fixed over time, to be estimated based on past rewards and decisions.

Our formulation can be viewed as a sequential method for maximizing a linear function based on noisy observations of the function values, and it is thus closely related to the field of stochastic approximation, which was developed by Robbins and Monro [29] and Kiefer and Wolfowitz [20]. We do not provide a comprehensive review of the literature here; interested readers are referred to an excellent survey by Lai [21]. In stochastic approximation, we wish to find an adaptive sequence $\{\mathbf{U}_t \in \mathbb{R}^r \colon t \geq 1\}$ that converges to a maximizer $\mathbf{u}^*$ of a target function, and the focus is on establishing the rate at which the mean squared error $\mathsf{E}[\|\mathbf{U}_t - \mathbf{u}^*\|^2]$ converges to zero (see, for example, Blum [10] and Cicek et al. [11]). In contrast, our cumulative regret and Bayes risk criteria take into account the cost associated with each observation. The different performance measures used in our formulation lead to entirely different policies and performance characteristics.

Our model generalizes the "response surface bandits (p. 771)" introduced by Ginebra and Clayton [16], who assume a normal prior on $\mathbf{Z}$ and provide a simple tunable heuristic, without any analysis on the regret or risk. Abe and Long [1], Auer [4], and Dani et al. [12] all consider a special case of our model where the random vector $\mathbf{Z}$ and the error random variables W_t^u are bounded almost surely, and with the exception of the last paper, focus on the regret criterion. Abe and Long [1] demonstrate a class of bandits where the dimension r is at least $\Omega(\sqrt{T})$, and show that the T-period regret under an arbitrary policy must be at least $\Omega(T^{3/4})$. Auer [4] describes an algorithm based on least squares estimation and confidence bounds and establishes an $O(\sqrt{r}\sqrt{T}\log^{3/2}(T\,|\mathcal{U}_r|))$ upper bound on the regret for the case of finitely many arms. Dani et al. [12] show that the policy of Auer [4] can be extended to problems having an arbitrary compact set of arms, and also make use of a barycentric spanner. They establish an $O(r\sqrt{T}\log^{3/2} T)$ upper bound on the regret and discuss a variation of the policy that is more computationally tractable (at the expense of higher regret). Dani et al. [12] also establish an $\Omega(r\sqrt{T})$ lower bound on the Bayes risk when the set of arms is the Cartesian product of

Table 1. Regret and risk bounds for various values of r and for different collections of arms.

		T-period cumulative regret		T-period cumulative Bayes risk	
Dimension (r)	Set of arms ($\mathcal{U}_r$)	Lower bound	Upper bound	Lower bound	Upper bound
$r=1$	Any compact set (Mersereau et al. [24])	$\Omega(\sqrt{T})$	$O(\sqrt{T})$	$\Omega(\log T)$	$O(\log T)$
$r \geq 2$ (this paper)	Unit sphere (§§2 and 3)	$\Omega(r\sqrt{T})$	$O(r\sqrt{T})$	$\Omega(r\sqrt{T})$	$O(r\sqrt{T})$
	Any compact set (§4)	$\Omega(r\sqrt{T})$	$O(r\sqrt{T}\log^{3/2} T)$	$\Omega(r\sqrt{T})$	$O(r\sqrt{T}\log^{3/2} T)$

circles.[1] However, this leaves a $O(\log^{3/2} T)$ gap from the upper bound, leaving open the question of the exact order of regret and risk.

1.4. Contributions and organizations. One of our contributions is proving that the regret and Bayes risk for a broad class of linearly parameterized bandits is of order $\Theta(r\sqrt{T})$. In §2, we establish an $\Omega(r\sqrt{T})$ lower bound for an arbitrary policy, when the set of arms is the unit sphere in $\mathbb{R}^r$. Then, in §3, we show that a matching $O(r\sqrt{T})$ upper bound can be achieved through a phase-based policy that alternates between exploration and exploitation phases. To the best of our knowledge, this is the first result that establishes matching upper and lower bounds for a class of linearly parameterized bandits. Table 1 summarizes our results and provides a comparison with the results in Mersereau et al. [24] for the case $r=1$. In the ensuing discussion of the bounds, we focus on the main parameters, r and T, with more precise statements given in the theorems.

Although we obtain the same lower bound of $\Omega(r\sqrt{T})$, our example and proof techniques are very different from Dani et al. [12]. We consider the unit sphere, with a multivariate normal prior on $\mathbf{Z}$, and standard normal errors. The analysis in §2 also illuminates the behavior of the least mean squares estimator in this setting, and we believe that it provides an approach that can be used to address more general classes of linear estimation and adaptive control problems.

We also prove that the phase-based policy remains effective (that is, admits an $O(r\sqrt{T})$ upper bound) for a broad class of bandit problems in which the set of arms is strongly convex[2] (defined in §3). To our knowledge, this is the first result that establishes the connection between a geometrical property (strong convexity) of the underlying set of arms and the effectiveness of separating exploration from exploitation. We suspect that strong convexity may have similar implications for other types of bandit and learning problems.

When the set of arms is an arbitrary compact set, the separation of exploration and exploitation may not be effective, and we consider in §4 an active exploration policy based on least squares estimation and confidence regions. We prove that the regret and risk under this policy are bounded above by $O(r\sqrt{T}\log^{3/2} T)$, which is within a logarithmic factor of the lower bound. Our policy is closely related to the one considered in Auer [4] and further analyzed in Dani et al. [12], with differences in a number of respects. First, our model allows the random vector $\mathbf{Z}$ and the errors W_t^u to have unbounded support, which requires a somewhat more complicated analysis. Second, our policy is an *anytime* policy, in the sense that the policy does not depend on the time horizon T of interest. In contrast, the policies of Auer [4] and Dani et al. [12] involve a certain parameter δ whose value must be set in advance as a function of the time horizon T in order to obtain the $O(r\sqrt{T}\log^{3/2} T)$ regret bound.

We finally comment on the case where the set of arms is finite and fixed. We show that the regret and risk under our active exploration policy increase gracefully with time, as $\log T$ and $\log^2 T$, respectively. These results show that our policy is within a constant factor of the asymptotic lower bounds established by Lai and Robbins [23] and Lai [22]. In contrast, for the policies of Auer [4] and Dani et al. [12], the available regret upper bounds grow over time as $\sqrt{T}\log^{3/2} T$ and $\log^3 T$, respectively.

We note that the bounds on the cumulative Bayes risk given in Table 1 hold under certain assumptions on the prior distribution of the random vector $\mathbf{Z}$. For $r=1$, $\mathbf{Z}$ is assumed to be a continuous random variable with a bounded density function (Theorem 3.2 in Mersereau et al. [24]). When the collection of arms is a unit sphere with $r \geq 2$, we require that both $\mathsf{E}[\|\mathbf{Z}\|]$ and $\mathsf{E}[1/\|\mathbf{Z}\|]$ are bounded (see Theorems 2.1 and 3.1, and Lemma 3.2). For general compact sets of arms where our risk bound is not tight, we only require that $\|\mathbf{Z}\|$ has a bounded expectation.

[1] The original lower bound (Theorem 3 on page 360 of Dani et al. [12]) was not entirely correct; a correct version was provided later, in Dani et al. [13].

[2] One can show that the Cartesian product of circles is not strongly convex, and thus, our phase-based policy cannot be applied to give the matching upper bound for the example used in Dani et al. [12].

2. Lower bounds. In this section, we establish $\Omega(r\sqrt{T})$ lower bounds on the regret and risk under an arbitrary policy when the set of arms is the unit sphere. This result is stated in the following theorem.[3]

THEROEM 2.1 (LOWER BOUNDS). *Consider a bandit problem where the set of arms is the unit sphere in $\mathbb{R}^r$, and W_t^u has a standard normal distribution with mean zero and variance one for all t and $\mathbf{u}$. If $\mathbf{Z}$ has a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_r/r$, then for all policies ψ and every $T \geq r^2$,*

$$\text{Risk}(T, \psi) \geq 0.006 r\sqrt{T}.$$

Consequently, for any policy ψ and $T \geq r^2$, there exists $\mathbf{z} \in \mathbb{R}^r$ such that

$$\text{Regret}(\mathbf{z}, T, \psi) \geq 0.006 r\sqrt{T}.$$

It suffices to establish the lower bound on the Bayes risk because the regret bound follows immediately. Throughout this section, we assume that $\mathcal{U}_r = \{\mathbf{u} \in \mathbb{R}^r\colon\ \|\mathbf{u}\| = 1\}$. We fix an arbitrary policy ψ and for any $t \geq 1$, we let $\mathbf{H}_t = (\mathbf{U}_1, X_1, \mathbf{U}_2, X_2, \ldots, \mathbf{U}_t, X_t)$ be the history up to time t. We also let $\hat{\mathbf{Z}}_t$ denote the least mean squares estimator of $\mathbf{Z}$ given the history $\mathbf{H}_t$, that is,

$$\hat{\mathbf{Z}}_t = \mathsf{E}[\mathbf{Z} \mid \mathbf{H}_t].$$

Let $\mathbf{S}_t^1, \ldots, \mathbf{S}_t^{r-1}$ denote a collection of orthogonal unit vectors that are also orthogonal to $\hat{\mathbf{Z}}_t$. Note that $\hat{\mathbf{Z}}_t$ and $\mathbf{S}_t^1, \ldots, \mathbf{S}_t^{r-1}$ are functions of $\mathbf{H}_t$.

Because $\mathcal{U}_r$ is the unit sphere, $\max_{u\in\mathcal{U}_r} \mathbf{u}'\mathbf{z} = (\mathbf{z}'\mathbf{z})/\|\mathbf{z}\| = \|\mathbf{z}\|$, for all $\mathbf{z} \in \mathbb{R}^r$. Thus, the risk in period t is given by $\mathsf{E}[\|\mathbf{Z}\| - \mathbf{U}_t'\mathbf{Z}]$. The following lemma establishes a lower bound on the cumulative risk in terms of the estimator error variance and the total amount of exploration along the directions $\mathbf{S}_T^1, \ldots, \mathbf{S}_T^{r-1}$.

LEMMA 2.2 (RISK DECOMPOSITION). *For any $T \geq 1$,*

$$\text{Risk}(T, \psi) \geq \frac{1}{2}\sum_{k=1}^{r-1} \mathsf{E}\left[\|\mathbf{Z}\| \sum_{t=1}^{T} (\mathbf{U}_t'\mathbf{S}_T^k)^2 + \frac{T}{\|\mathbf{Z}\|}\{(\mathbf{Z} - \hat{\mathbf{Z}}_T)'\mathbf{S}_T^k\}^2\right].$$

PROOF. Using the fact that for any two unit vectors $\mathbf{w}$ and $\mathbf{v}$, $1 - \mathbf{w}'\mathbf{v} = \|\mathbf{w} - \mathbf{v}\|^2/2$, the instantaneous regret in period t satisfies

$$\|\mathbf{Z}\| - \mathbf{U}_t'\mathbf{Z} = \|\mathbf{Z}\|\left(1 - \mathbf{U}_t'\frac{\mathbf{Z}}{\|\mathbf{Z}\|}\right) = \frac{\|\mathbf{Z}\|}{2}\left\|\mathbf{U}_t - \frac{\mathbf{Z}}{\|\mathbf{Z}\|}\right\|^2 \geq \frac{\|\mathbf{Z}\|}{2}\sum_{k=1}^{r-1}\left\{\left(\mathbf{U}_t - \frac{\mathbf{Z}}{\|\mathbf{Z}\|}\right)'\mathbf{S}_T^k\right\}^2,$$

where the inequality follows from the fact that $\mathbf{S}_T^1, \ldots, \mathbf{S}_T^{r-1}$ are orthogonal unit vectors. Therefore, the cumulative conditional risk satisfies

$$\begin{aligned}
2\sum_{t=1}^{T}\mathsf{E}[\|\mathbf{Z}\| - \mathbf{U}_t'\mathbf{Z} \mid \mathbf{H}_T] &\geq \sum_{t=1}^{T}\mathsf{E}\left[\|\mathbf{Z}\|\sum_{k=1}^{r-1}\left\{\left(\mathbf{U}_t - \frac{\mathbf{Z}}{\|\mathbf{Z}\|}\right)'\mathbf{S}_T^k\right\}^2 \,\middle|\, \mathbf{H}_T\right] \\
&= \sum_{t=1}^{T}\sum_{k=1}^{r-1}\mathsf{E}\left[\|\mathbf{Z}\|\left\{\left(\mathbf{U}_t - \frac{\mathbf{Z}}{\|\mathbf{Z}\|}\right)'\mathbf{S}_T^k\right\}^2 \,\middle|\, \mathbf{H}_T\right] \\
&= \sum_{t=1}^{T}\sum_{k=1}^{r-1}\mathsf{E}\left[\|\mathbf{Z}\|(\mathbf{U}_t'\mathbf{S}_T^k)^2 - 2(\mathbf{U}_t'\mathbf{S}_T^k)(\mathbf{Z}'\mathbf{S}_T^k) + \frac{(\mathbf{Z}'\mathbf{S}_T^k)^2}{\|\mathbf{Z}\|} \,\middle|\, \mathbf{H}_T\right],
\end{aligned}$$

with probability one. From the definition of $\mathbf{S}_T^k$, we have $\hat{\mathbf{Z}}_T'\mathbf{S}_T^k = 0$ for $k = 1, \ldots, r-1$. Therefore, for $t \leq T$,

$$\mathsf{E}[(\mathbf{U}_t'\mathbf{S}_T^k)(\mathbf{Z}'\mathbf{S}_T^k) \mid \mathbf{H}_T] = (\mathbf{U}_t'\mathbf{S}_T^k)\mathsf{E}[\mathbf{Z}' \mid \mathbf{H}_T]\mathbf{S}_T^k = (\mathbf{U}_t'\mathbf{S}_T^k)\hat{\mathbf{Z}}_T'\mathbf{S}_T^k = 0,$$

which eliminates the middle term in the summand above. Furthermore, we see that $\mathbf{Z}'\mathbf{S}_T^k = (\mathbf{Z} - \hat{\mathbf{Z}}_T)'\mathbf{S}_T^k$ for all k. Thus, with probability one,

$$\sum_{t=1}^{T}\mathsf{E}[\|\mathbf{Z}\| - \mathbf{U}_t'\mathbf{Z} \mid \mathbf{H}_T] \geq \frac{1}{2}\sum_{k=1}^{r-1}\mathsf{E}\left[\|\mathbf{Z}\|\sum_{t=1}^{T}(\mathbf{U}_t'\mathbf{S}_T^k)^2 + \frac{T}{\|\mathbf{Z}\|}\{(\mathbf{Z} - \hat{\mathbf{Z}}_T)'\mathbf{S}_T^k\}^2 \,\middle|\, \mathbf{H}_T\right],$$

and the desired result follows by taking the expectation of both sides. □

[3] The result of Theorem 2.1 easily extends to the case where the covariance matrix is $\mathbf{I}_r$, rather than $\mathbf{I}_r/r$. The proof is essentially the same.

Because $\mathbf{S}_T^k$ is orthogonal to $\hat{\mathbf{Z}}_T$, we can interpret $\sum_{t=1}^T (\mathbf{U}_t'\mathbf{S}_T^k)^2$ and $\{(\mathbf{Z}-\hat{\mathbf{Z}}_T)'\mathbf{S}_T^k\}^2$ as the total amount of exploration over T periods and the squared estimation error, respectively, in the direction $\mathbf{S}_T^k$. Thus, Lemma 2.2 tells us that the cumulative risk is bounded below by the sum of the squared estimation error and the total amount of exploration in the past T periods. This result suggests an approach for establishing a lower bound on the risk. If the amount of exploration $\sum_{t=1}^T (\mathbf{U}_t'\mathbf{S}_T^k)^2$ is large, then the risk will be large. On the other hand, if the amount of exploration is small, we expect significant estimation errors, which in turn imply large risk. This intuition is made precise in Lemma 2.3, which relates the squared estimation error and the amount of exploration.

Lemma 2.3 (Little Exploration Implies Large Estimation Errors). *For any k and $T \geq 1$,*

$$\mathsf{E}[\{(\mathbf{Z}-\hat{\mathbf{Z}}_T)'\mathbf{S}_T^k\}^2 \mid \mathbf{H}_T] \geq \frac{1}{r+\sum_{t=1}^T (\mathbf{U}_t'\mathbf{S}_T^k)^2},$$

with probability one.

Proof. Let $\mathbf{Q}_T = \hat{\mathbf{Z}}_T/\|\hat{\mathbf{Z}}_T\|$. For any t, we have that $\mathbf{U}_t = \sum_{k=1}^{r-1}(\mathbf{U}_t'\mathbf{S}_T^k)\mathbf{S}_T^k + (\mathbf{U}_t'\mathbf{Q}_T)\mathbf{Q}_T$. Let

$$\mathbf{V} = [\mathbf{S}_T^1 \mathbf{S}_T^2 \cdots \mathbf{S}_T^{r-1} \mathbf{Q}_T]$$

be an $r \times r$ orthonormal matrix whose columns are the vectors $\mathbf{S}_T^1, \ldots, \mathbf{S}_T^{r-1}$, and $\mathbf{Q}_T$, respectively. Then, it is easy to verify that

$$\sum_{t=1}^T \mathbf{U}_t\mathbf{U}_t' = \mathbf{VAV}',$$

where $\mathbf{A} = \begin{pmatrix} \Sigma & \mathbf{c} \\ \mathbf{c}' & a \end{pmatrix}$, is an $r \times r$ matrix, with $a = \mathbf{Q}_T'(\sum_{t=1}^T \mathbf{U}_t\mathbf{U}_t')\mathbf{Q}_T$, $\mathbf{c}$ is an $(r-1)$-dimensional column vector, and where Σ is an $(r-1)\times(r-1)$ matrix with $\boldsymbol{\Sigma}_{kl} = (\mathbf{S}_T^k)'(\sum_{t=1}^T \mathbf{U}_t\mathbf{U}_t')\mathbf{S}_T^l = \sum_{t=1}^T (\mathbf{U}_t'\mathbf{S}_T^k)(\mathbf{U}_t'\mathbf{S}_T^l)$ for k, $l = 1, \ldots, r-1$.

Because $\mathbf{Z}$ has a multivariate normal prior distribution with covariance matrix $\mathbf{I}_r/r$, it is a standard result (use, for example, Corollary E.3.5 in Appendix E in Bertsekas [7]) that

$$\mathsf{E}[(\mathbf{Z}-\hat{\mathbf{Z}}_T)(\mathbf{Z}-\hat{\mathbf{Z}}_T)' \mid \mathbf{H}_T] = \left(r\mathbf{I}_r + \sum_{t=1}^T \mathbf{U}_t\mathbf{U}_t'\right)^{-1} = \mathbf{V}(r\mathbf{I}_r+\mathbf{A})^{-1}\mathbf{V}'.$$

Because $\mathbf{S}_T^k$ is a function of $\mathbf{H}_T$ and $\mathbf{V}'\mathbf{S}_T^k = \mathbf{e}_k$, we have, for $k \leq r-1$, that

$$\begin{aligned}\mathsf{E}[\{(\mathbf{Z}-\hat{\mathbf{Z}}_T)'\mathbf{S}_T^k\}^2 \mid \mathbf{H}_T] &= (\mathbf{V}'\mathbf{S}_T^k)'(r\mathbf{I}_r+\mathbf{A})^{-1}(\mathbf{V}'\mathbf{S}_T^k) = [(r\mathbf{I}_r+\mathbf{A})^{-1}]_{kk}\\ &\geq \frac{1}{(r\mathbf{I}_r+\mathbf{A})_{kk}} = \frac{1}{r+\sum_{t=1}^T(\mathbf{U}_t'\mathbf{S}_T^k)^2},\end{aligned}$$

where the inequality follows from Fiedler's inequality (see, for example, Theorem 2.1 in Fiedler and Pták [15]), and the final equality follows from the definition of $\mathbf{A}$. □

The next lemma gives a lower bound on the probability that $\mathbf{Z}$ is bounded away from the origin. The proof follows from simple calculations involving normal densities, and the details are given in Appendix A.1.

Lemma 2.4. *For any $\theta \leq 1/2$ and $\beta > 0$,* $\Pr\{\theta \leq \|\mathbf{Z}\| \leq \beta\} \geq 1 - 4\theta^2 - 1/\beta^2$.

The last lemma establishes a lower bound on the sum of the total amount of exploration and the squared estimation error, which is also the minimum cumulative Bayes risk along the direction $\mathbf{S}_T^k$ by Lemma 2.2.

Lemma 2.5 (Minimum Directional Risk). *For $k = 1, \ldots, r-1$, and $T \geq r^2$,*

$$\mathsf{E}\left[\|\mathbf{Z}\|\sum_{t=1}^T(\mathbf{U}_t'\mathbf{S}_T^k)^2 + \frac{T}{\|\mathbf{Z}\|}\{(\mathbf{Z}-\hat{\mathbf{Z}}_T)'\mathbf{S}_T^k\}^2\right] \geq 0.027\sqrt{T}.$$

We note that if $\|\mathbf{Z}\|$ were a constant, rather than a random variable, Lemma 2.5 would follow immediately. Hence, most of the work in the proof below involves constraining $\|\mathbf{Z}\|$ to a certain range $[\theta, \beta]$.

Proof. Consider an arbitrary k, and let $\Xi = \sum_{t=1}^{T} (\mathbf{U}_t' \mathbf{S}_T^k)^2$, $\Gamma = \{(\mathbf{Z} - \hat{\mathbf{Z}}_T)' \mathbf{S}_T^k\}^2$. Our proof will make use of positive constants θ, β, and η, whose values will be chosen later. Note that

$$\begin{aligned}\mathsf{E}\left[\|\mathbf{Z}\|\,\Xi + \frac{T\Gamma}{\|\mathbf{Z}\|} \,\middle|\, \mathbf{H}_T\right] &\geq \mathsf{E}\left[\left(\|\mathbf{Z}\|\,\Xi + \frac{T\Gamma}{\|\mathbf{Z}\|}\right)\mathbb{1}_{\{\theta\leq\|\mathbf{Z}\|\leq\beta\}}\mathbb{1}_{\{\Xi\geq\sqrt{T}\}} \,\middle|\, \mathbf{H}_T\right]\\ &\quad + \mathsf{E}\left[\left(\|\mathbf{Z}\|\,\Xi + \frac{T\Gamma}{\|\mathbf{Z}\|}\right)\mathbb{1}_{\{\theta\leq\|\mathbf{Z}\|\leq\beta\}}\mathbb{1}_{\{\Xi<\sqrt{T}\}} \,\middle|\, \mathbf{H}_T\right]\\ &\geq \theta\sqrt{T}\mathbb{1}_{\{\Xi\geq\sqrt{T}\}}\mathsf{E}[\mathbb{1}_{\{\theta\leq\|\mathbf{Z}\|\leq\beta\}} \mid \mathbf{H}_T] + \frac{T}{\beta}\mathbb{1}_{\{\Xi<\sqrt{T}\}}\mathsf{E}[\Gamma\mathbb{1}_{\{\theta\leq\|\mathbf{Z}\|\leq\beta\}} \mid \mathbf{H}_T],\end{aligned}$$

where we use the fact that Ξ is a function of $\mathbf{H}_T$ in the final inequality. We will now lower bound the last term on the right-hand side of the above inequality. Let $\Theta = \mathsf{E}[\Gamma \mid \mathbf{H}_T]$. Because Θ is a function of $\mathbf{H}_T$,

$$\begin{aligned}\frac{T}{\beta}\mathbb{1}_{\{\Xi<\sqrt{T}\}}\mathsf{E}[\Gamma\mathbb{1}_{\{\theta\leq\|\mathbf{Z}\|\leq\beta\}} \mid \mathbf{H}_T] &\geq \frac{T}{\beta}\mathbb{1}_{\{\Xi<\sqrt{T}\}}\mathsf{E}[\Gamma\mathbb{1}_{\{\theta\leq\|\mathbf{Z}\|\leq\beta\}}\mathbb{1}_{\{\Gamma\geq\eta\Theta\}} \mid \mathbf{H}_T]\\ &\geq \frac{\eta T}{\beta}\Theta\mathbb{1}_{\{\Xi<\sqrt{T}\}}\mathsf{E}[\mathbb{1}_{\{\theta\leq\|\mathbf{Z}\|\leq\beta\}}\mathbb{1}_{\{\Gamma\geq\eta\Theta\}} \mid \mathbf{H}_T]\\ &\geq \frac{\eta\sqrt{T}}{2\beta}\mathbb{1}_{\{\Xi<\sqrt{T}\}}\mathsf{E}[\mathbb{1}_{\{\theta\leq\|\mathbf{Z}\|\leq\beta\}}\mathbb{1}_{\{\Gamma\geq\eta\Theta\}} \mid \mathbf{H}_T],\end{aligned}$$

where the last inequality follows from Lemma 2.3 which implies that, with probability one,

$$\frac{\eta T}{\beta}\Theta\mathbb{1}_{\{\Xi<\sqrt{T}\}} \geq \frac{\eta T}{\beta}\cdot\frac{1}{r+\Xi}\mathbb{1}_{\{\Xi<\sqrt{T}\}} \geq \frac{\eta T}{\beta}\cdot\frac{1}{r+\sqrt{T}}\mathbb{1}_{\{\Xi<\sqrt{T}\}} \geq \frac{\eta\sqrt{T}}{2\beta}\mathbb{1}_{\{\Xi<\sqrt{T}\}},$$

and where the last inequality follows from the fact that $T \geq r^2$, and thus, $1/(r+\sqrt{T}) \geq 1/(2\sqrt{T})$.

Putting everything together, we obtain

$$\begin{aligned}\mathsf{E}\left[\|\mathbf{Z}\|\,\Xi + \frac{T\Gamma}{\|\mathbf{Z}\|} \,\middle|\, \mathbf{H}_T\right] &\geq \theta\sqrt{T}\mathbb{1}_{\{\Xi\geq\sqrt{T}\}}\mathsf{E}[\mathbb{1}_{\{\theta\leq\|\mathbf{Z}\|\leq\beta\}} \mid \mathbf{H}_T] + \frac{\eta\sqrt{T}}{2\beta}\mathbb{1}_{\{\Xi<\sqrt{T}\}}\mathsf{E}[\mathbb{1}_{\{\theta\leq\|\mathbf{Z}\|\leq\beta\}}\mathbb{1}_{\{\Gamma\geq\eta\Theta\}} \mid \mathbf{H}_T],\\ &\geq \min\left\{\theta, \frac{\eta}{2\beta}\right\}\sqrt{T}\mathsf{E}[\mathbb{1}_{\{\theta\leq\|\mathbf{Z}\|\leq\beta\}}\mathbb{1}_{\{\Gamma\geq\eta\Theta\}} \mid \mathbf{H}_T],\end{aligned}$$

with probability one. By the Bonferroni inequality, we have that

$$\begin{aligned}\mathsf{E}[\mathbb{1}_{\{\theta\leq\|\mathbf{Z}\|\leq\beta\}}\mathbb{1}_{\{\Gamma\geq\eta\Theta\}} \mid \mathbf{H}_T] &= \Pr\{\theta \leq \|\mathbf{Z}\| \leq \beta \text{ and } \Gamma \geq \eta\Theta \mid \mathbf{H}_T\}\\ &\geq \Pr\{\theta \leq \|\mathbf{Z}\| \leq \beta \mid \mathbf{H}_T\} + \Pr\{\Gamma \geq \eta\Theta \mid \mathbf{H}_T\} - 1,\end{aligned}$$

with probability one. Conditioned on $\mathbf{H}_T$, $(\mathbf{Z} - \hat{\mathbf{Z}}_T)'\mathbf{S}_T^k$ is normally distributed with mean zero and variance

$$\mathsf{E}[\{(\mathbf{Z} - \hat{\mathbf{Z}}_T)'\mathbf{S}_T^k\}^2 \mid \mathbf{H}_T] = \mathsf{E}[\Gamma \mid \mathbf{H}_T] = \Theta.$$

Let $\Phi(\cdot)$ be the cumulative distribution function of the standard normal random variable, that is, $\Phi(x) = (1/\sqrt{2\pi})\int_{-\infty}^{x} e^{-u^2/2}\,du$. Then,

$$\Pr\{\Gamma \geq \eta\Theta \mid \mathbf{H}_T\} = \Pr\{|(\mathbf{Z} - \hat{\mathbf{Z}}_T)'\mathbf{S}_T^k| \geq \sqrt{\eta}\sqrt{\Theta} \mid \mathbf{H}_T\} = 2(1 - \Phi(\sqrt{\eta})),$$

from which it follows that, with probability one,

$$\mathsf{E}[\mathbb{1}_{\{\theta\leq\|\mathbf{Z}\|\leq\beta\}}\mathbb{1}_{\{\Gamma\geq\eta\Theta\}} \mid \mathbf{H}_T] \geq \Pr\{\theta \leq \|\mathbf{Z}\| \leq \beta \mid \mathbf{H}_T\} + 2(1 - \Phi(\sqrt{\eta})) - 1.$$

Therefore,

$$\mathsf{E}\left[\|\mathbf{Z}\|\,\Xi + \frac{T\Gamma}{\|\mathbf{Z}\|} \,\middle|\, \mathbf{H}_T\right] \geq \min\left\{\theta, \frac{\eta}{2\beta}\right\}[\Pr\{\theta \leq \|\mathbf{Z}\| \leq \beta \mid \mathbf{H}_T\} + 2(1 - \Phi(\sqrt{\eta})) - 1]\sqrt{T},$$

with probability one, which implies that

$$\mathsf{E}\left[\|\mathbf{Z}\|\,\Xi+\frac{T\Gamma}{\|\mathbf{Z}\|}\right]\geq\min\left\{\theta,\,\frac{\eta}{2\beta}\right\}[\Pr\{\theta\leq\|\mathbf{Z}\|\leq\beta\}+2(1-\Phi(\sqrt{\eta}))-1]\sqrt{T},$$
$$\geq\min\left\{\theta,\,\frac{\eta}{2\beta}\right\}\left[2(1-\Phi(\sqrt{\eta}))-\frac{1}{\beta^2}-4\theta^2\right]\sqrt{T},$$

where the last inequality follows from Lemma 2.4. Set $\theta=0.09$, $\beta=3$, and $\eta=0.5$, to obtain $\mathsf{E}[\|\mathbf{Z}\|\,\Xi+T\Gamma/\|\mathbf{Z}\|]\geq 0.027\sqrt{T}$, which is the desired result. □

Finally, here is the proof of Theorem 2.1.

Proof. It follows from Lemmas 2.2 and 2.5 that

$$\mathrm{Risk}(T,\psi)\geq\frac{1}{2}\sum_{k=1}^{r-1}\mathsf{E}\left[\|\mathbf{Z}\|\sum_{t=1}^{T}(\mathbf{U}_t'\mathbf{S}_T^k)^2+\frac{T}{\|\mathbf{Z}\|}\{(\mathbf{Z}-\hat{\mathbf{Z}}_T)'\mathbf{S}_T^k\}^2\right]$$
$$\geq\frac{r-1}{2}\cdot 0.027\sqrt{T}\geq\frac{r}{4}\cdot 0.027\sqrt{T}\geq 0.006r\sqrt{T},$$

where we have used the fact $r\geq 2$, which implies that $r-1\geq r/2$. □

3. Matching upper bounds.

We have established $\Omega(r\sqrt{T})$ lower bounds when the set of arms $\mathcal{U}_r$ is the unit sphere. We now prove that a policy that alternates between exploration and exploitation phases yields matching upper bounds on the regret and risk, and is therefore optimal for this problem. Surprisingly, we will see that the phase-based policy is effective for a large class of bandit problems, involving a strongly convex set of arms. We introduce the following assumption on the tails of the error random variables W_t^u and on the set of arms $\mathcal{U}_r$, which will remain in effect throughout the rest of paper.

Assumption 1.

(a) *There exists a positive constant σ_0 such that for any $r\geq 2$, $\mathbf{u}\in\mathcal{U}_r$, $t\geq 1$, and $x\in\mathbb{R}$, we have $\mathsf{E}[e^{xW_t^u}]\leq e^{x^2\sigma_0^2/2}$.*

(b) *There exist positive constants $\bar{u}$ and λ_0 such that for any $r\geq 2$,*

$$\max_{\mathbf{u}\in\mathcal{U}_r}\|\mathbf{u}\|\leq\bar{u},$$

and the set of arms $\mathcal{U}_r\subset\mathbb{R}^r$ has r linearly independent elements $\mathbf{b}_1,\ldots,\mathbf{b}_r$ such that $\lambda_{\min}(\sum_{k=1}^r\mathbf{b}_k\mathbf{b}_k')\geq\lambda_0$.

Under Assumption 1(a), the tails of the distribution of the errors W_t^u decay at least as fast as for a normal random variable with variance σ_0^2. The first part of Assumption 1(b) ensures that the expected reward of the arms remain bounded as the dimension r increases, and the arms $\mathbf{b}_1,\ldots,\mathbf{b}_r$ given in the second part of Assumption 1(b) will be used during the exploration phase of our policy.

Our policy—which we refer to as the Phased Exploration and Greedy Exploitation (PEGE)—operates in cycles, and in each cycle, we alternate between exploration and exploitation phases. During the exploration phase of cycle c, we play the r linearly independent arms from Assumption 1(b). Using the rewards observed during the exploration phases in the past c cycles, we compute an *ordinary least squares* (OLS) estimate $\hat{\mathbf{Z}}(c)$. In the exploitation phase of cycle c, we use $\hat{\mathbf{Z}}(c)$ as a proxy for $\mathbf{Z}$ and compute a *greedy* decision $\mathbf{G}(c)\in\mathcal{U}_r$ defined by

$$\mathbf{G}(c)=\arg\max_{\mathbf{v}\in\mathcal{U}_r}\mathbf{v}'\hat{\mathbf{Z}}(c),\tag{2}$$

where we break ties arbitrarily. We then play the arm $\mathbf{G}(c)$ for an additional c periods to complete cycle c. Here is a formal description of the policy.

Phased Exploration and Greedy Exploitation (PEGE)

Description: For each cycle $c\geq 1$, complete the following two phases.

(i) **Exploration (r periods):** For $k=1,2,\ldots,r$, play arm $\mathbf{b}_k\in\mathcal{U}_r$ given in Assumption 1(b), and observe the reward $X^{b_k}(c)$. Compute the OLS estimate $\hat{\mathbf{Z}}(c)\in\mathbb{R}^r$, given by

$$\hat{\mathbf{Z}}(c)=\frac{1}{c}\left(\sum_{k=1}^r\mathbf{b}_k\mathbf{b}_k'\right)^{-1}\sum_{s=1}^c\sum_{k=1}^r\mathbf{b}_kX^{b_k}(s)=\mathbf{Z}+\frac{1}{c}\left(\sum_{k=1}^r\mathbf{b}_k\mathbf{b}_k'\right)^{-1}\sum_{s=1}^c\sum_{k=1}^r\mathbf{b}_kW^{b_k}(s),$$

where for any k, $X^{b_k}(s)$, and $W^{b_k}(s)$ denote the observed reward and the error random variable associated with playing arm $\mathbf{b}_k$ in cycle s. Note that the last equality follows from Equation (1) defining our model.

(ii) **Exploitation (c periods):** Play the greedy arm $\mathbf{G}(c)=\arg\max_{\mathbf{v}\in\mathcal{U}_r}\mathbf{v}'\hat{\mathbf{Z}}(c)$ for c periods.

Because $\mathcal{U}_r$ is compact, for each $\mathbf{z} \in \mathbb{R}^r$, there is an optimal arm that gives the maximum expected reward. When this best arm varies smoothly with $\mathbf{z}$, we will show that the T-period regret and risk under the PEGE policy is bounded above by $O(r\sqrt{T})$. More precisely, we say that a set of arms $\mathcal{U}_r$ satisfies the *smooth best arm response with parameter J* (SBAR(J), for short) condition if for any nonzero vector $\mathbf{z} \in \mathbb{R}^r \backslash \{\mathbf{0}\}$, there is a unique best arm $\mathbf{u}^*(\mathbf{z}) \in \mathcal{U}_r$ that gives the maximum expected reward, and for any two unit vectors $\mathbf{z} \in \mathbb{R}^r$ an $\mathbf{y} \in \mathbb{R}^r$ with $\|\mathbf{z}\| = \|\mathbf{y}\| = 1$, we have

$$\|\mathbf{u}^*(\mathbf{z}) - \mathbf{u}^*(\mathbf{y})\| \le J\,\|\mathbf{z} - \mathbf{y}\|.$$

Even though the SBAR condition appears to be an implicit one, it admits a simple interpretation. According to Corollary 4 of Polovinkin [26], a compact set $\mathcal{U}_r$ satisfies condition SBAR(J) if and only if it is strongly convex with parameter J, in the sense that the set $\mathcal{U}_r$ can be represented as the intersection of closed balls of radius J. Intuitively, the SBAR condition requires the boundary of $\mathcal{U}_r$ to have a curvature that is bounded below by a positive constant. For some examples, the unit ball satisfies the SBAR(1) condition. Furthermore, according to Theorem 3 of Polovinkin [26], an ellipsoid of the form $\{\mathbf{u} \in R^r \colon \mathbf{u}'\mathbf{Q}^{-1}\mathbf{u} \le 1\}$, where $\mathbf{Q}$ is a symmetric positive definite matrix, satisfies the condition SBAR $(\lambda_{\max}(\mathbf{Q})/\sqrt{\lambda_{\min}(\mathbf{Q})})$.

The main result of this section is stated in the following theorem. The proof is given in §3.1.

Theorem 3.1 (Regret and Risk Under the Greedy Policy). *Suppose that Assumption 1 holds and that the sets $\mathcal{U}_r$ satisfy the SBAR(J) condition. Then, there exists a positive constant a_1 that depends only on σ_0, $\bar{u}$, λ_0, and J, such that for any $\mathbf{z} \in \mathbb{R}^r \backslash \{\mathbf{0}\}$ and $T \ge r$,*

$$\text{Regret}(\mathbf{z}, T, \text{PEGE}) \le a_1 \left(\|\mathbf{z}\| + \frac{1}{\|\mathbf{z}\|} \right) r\sqrt{T}.$$

Suppose in addition, that there exists a constant $M > 0$ such that for every $r \ge 2$ we have $\mathsf{E}[\|\mathbf{Z}\|] \le M$ and $\mathsf{E}[1/\|\mathbf{Z}\|] \le M$. Then, there exists a positive constant a_2 that depends only on σ_0, $\bar{u}$, λ_0, J, and M, such that for any $T \ge r$,

$$\text{Risk}(T, \text{PEGE}) \le a_2 r\sqrt{T}.$$

Dependence on $\|\mathbf{z}\|$ in the regret bound: By Assumption 1(b), for any $\mathbf{z} \in \mathbb{R}^r$, the instantaneous regret under arm $\mathbf{v} \in \mathcal{U}$ is bounded by $\max_{u \in \mathcal{U}} \mathbf{z}'(\mathbf{u} - \mathbf{v}) \le 2\bar{u}\,\|\mathbf{z}\|$. Thus, $2\bar{u}\,\|\mathbf{z}\|\,T$ provides a trivial upper bound on the T-period cumulative regret under the PEGE policy. Combining this with Theorem 3.1, we have that

$$\text{Regret}(\mathbf{z}, T, \text{PEGE}) \le \max\{a_1, 2\bar{u}\} \cdot \min\left\{ \left(\|\mathbf{z}\| + \frac{1}{\|\mathbf{z}\|} \right) r\sqrt{T},\ \|\mathbf{z}\|\,T \right\}.$$

The above result shows that the performance of our policy does *not* deteriorate as the norm of $\mathbf{z}$ approaches zero.

Intuitively, the requirement $\mathsf{E}[\|\mathbf{Z}\|] \le M$ in Theorem 3.1 implies that, as r increases, the maximum expected reward (over all arms) remains bounded. Moreover, the assumption on the boundedness of $\mathsf{E}[1/\|\mathbf{Z}\|]$ means that $\mathbf{Z}$ does not have too much mass near the origin. The following lemma provides conditions under which this assumption holds, and shows that the case of the multivariate normal distribution used in Theorem 2.1 is also covered. The proof is given in Appendix A.2.

Lemma 3.2 (Small Mass Near the Origin).

(a) *Suppose that there exist constants M_0 and $\rho \in (0, 1]$ such that for any $r \ge 2$, the random variable $\|\mathbf{Z}\|$ has a density function $g\colon \mathbb{R}_+ \to \mathbb{R}_+$ such that $g(x) \le M_0 x^\rho$ for all $x \in [0, \rho]$. Then, $\mathsf{E}[1/\|\mathbf{Z}\|] \le M$, where M depends only on M_0 and ρ.*

(b) *Suppose that for any $r \ge 2$, the random vector $\mathbf{Z}$ has a multivariate normal distribution with mean $\mathbf{0} \in \mathbb{R}^r$ and covariance matrix $\mathbf{I}_r/r$. Then, $\mathsf{E}[\|\mathbf{Z}\|] \le 1$ and $\mathsf{E}[1/\|\mathbf{Z}\|] \le \sqrt{\pi}$.*

The following corollary shows that the example in §2 admits tight matching upper bounds on the regret and risk.

Corollary 3.3 (Matching Upper Bounds). *Consider a bandit problem where the set of arms is the unit sphere in $\mathbb{R}^r$, and where W_t^u has a standard normal distribution with mean zero and variance one for all t and $\mathbf{u}$. Then, there exists an absolute constant a_3 such that for any $\mathbf{z} \in \mathbb{R}^r \backslash \{\mathbf{0}\}$ and $T \ge r$,*

$$\text{Regret}(\mathbf{z}, T, \text{PEGE}) \le a_3 \left(\|\mathbf{z}\| + \frac{1}{\|\mathbf{z}\|} \right) r\sqrt{T}.$$

Moreover, if $\mathbf{Z}$ *has a multivariate normal distribution with mean* $\mathbf{0}$ *and covariance matrix* $\mathbf{I}_r/r$, *then for all* $T \geq r$,

$$\mathrm{Risk}(T, \mathrm{PEGE}) \leq a_3 r\sqrt{T}.$$

Proof. Because the set of arms is the unit sphere and the errors are standard normal, Assumption 1 is satisfied with $\sigma_0 = \bar{u} = \lambda_0 = 1$. Moreover, as already discussed, the unit sphere satisfies the SBAR(1) condition. Finally, By Lemma 3.2, the random vector $\mathbf{Z}$ satisfies the hypotheses of Theorem 3.1. The regret and risk bounds then follow immediately. □

3.1. Proof of Theorem 3.1.

The proof of Theorem 3.1 relies on the following upper bound on the square of the norm difference between $\hat{\mathbf{Z}}(c)$ and $\mathbf{Z}$.

Lemma 3.4 (Bound on Squared Norm Difference). *Under Assumption 1, there exists a positive constant* h_1 *that depends only on* σ_0, $\bar{u}$, *and* λ_0 *such that for any* $\mathbf{z} \in \mathbb{R}^r$ *and* $c \geq 1$,

$$\mathsf{E}[\|\hat{\mathbf{Z}}(c) - \mathbf{z}\|^2 \mid \mathbf{Z} = \mathbf{z}] \leq \frac{h_1 r}{c}.$$

Proof. Recall from the definition of the PEGE policy that the estimate $\hat{\mathbf{Z}}(c)$ at the end of the exploration phase of cycle c is given by

$$\hat{\mathbf{Z}}(c) = \mathbf{Z} + \frac{1}{c}\left(\sum_{k=1}^{r} \mathbf{b}_k \mathbf{b}_k'\right)^{-1} \sum_{s=1}^{c}\sum_{k=1}^{r} \mathbf{b}_k W^{b_k}(s) = \mathbf{Z} + \frac{1}{c}\sum_{s=1}^{c} \mathbf{B}\mathbf{V}(s),$$

where $\mathbf{B} = (\sum_{k=1}^r \mathbf{b}_k \mathbf{b}_k')^{-1}$ and $\mathbf{V}(s) = \sum_{k=1}^r \mathbf{b}_k W^{b_k}(s)$. Note that the mean-zero random variables $W^{b_k}(s)$ are independent of each other and their variance is bounded by some constant γ_0 that depends only on σ_0. Then, it follows from Assumption 1 that

$$\begin{aligned}\mathsf{E}[\|\hat{\mathbf{Z}}(c) - \mathbf{z}\|^2 \mid \mathbf{Z} = \mathbf{z}] &= \frac{1}{c^2}\sum_{s=1}^{c} \mathsf{E}[\mathbf{V}(s)'\mathbf{B}^2\mathbf{V}(s)] = \frac{1}{c^2}\sum_{s=1}^{c}\sum_{k=1}^{r} \mathsf{E}[(W^{b_k}(s))^2]\mathbf{b}_k'\mathbf{B}^2\mathbf{b}_k \\ &\leq \frac{\gamma_0}{c}\sum_{k=1}^{r} \mathbf{b}_k'\mathbf{B}^2\mathbf{b}_k \leq \frac{\gamma_0}{c}\sum_{k=1}^{r} \lambda_{\max}(\mathbf{B}^2)\,\|\mathbf{b}_k\|^2 \leq \frac{\gamma_0 \bar{u}^2 r}{\lambda_0^2 c},\end{aligned}$$

which is the desired result. □

The next lemma gives an upper bound on the difference between two normalized vectors in terms of the difference of the original vectors.

Lemma 3.5 (Difference Between Normalized Vectors). *For any* $\mathbf{z}, \mathbf{w} \in \mathbb{R}^r$, *not both equal to zero,*

$$\left\| \frac{\mathbf{w}}{\|\mathbf{w}\|} - \frac{\mathbf{z}}{\|\mathbf{z}\|} \right\| \leq \frac{2\,\|\mathbf{w} - \mathbf{z}\|}{\max\{\|\mathbf{z}\|, \|\mathbf{w}\|\}},$$

where we define $\mathbf{0}/\|\mathbf{0}\|$ *to be some fixed unit vector.*

Proof. The inequality is easily seen to hold if either $\mathbf{w} = \mathbf{0}$ or $\mathbf{z} = \mathbf{0}$. So, assume that both $\mathbf{w}$ and $\mathbf{z}$ are nonzero. Using the triangle inequality and the fact that $|\,\|\mathbf{w}\| - \|\mathbf{z}\|\,| \leq \|\mathbf{w} - \mathbf{z}\|$, we have that

$$\left\| \frac{\mathbf{w}}{\|\mathbf{w}\|} - \frac{\mathbf{z}}{\|\mathbf{z}\|} \right\| \leq \left\| \frac{\mathbf{w}}{\|\mathbf{w}\|} - \frac{\mathbf{z}}{\|\mathbf{w}\|} \right\| + \left\| \frac{\mathbf{z}}{\|\mathbf{w}\|} - \frac{\mathbf{z}}{\|\mathbf{z}\|} \right\| = \frac{\|\mathbf{w} - \mathbf{z}\|}{\|\mathbf{w}\|} + \|\mathbf{z}\| \left| \frac{1}{\|\mathbf{w}\|} - \frac{1}{\|\mathbf{z}\|} \right| \leq \frac{2\,\|\mathbf{w} - \mathbf{z}\|}{\|\mathbf{w}\|}.$$

By symmetry, we also have $\|\mathbf{w}/\|\mathbf{w}\| - \mathbf{z}/\|\mathbf{z}\|\| \leq 2\,\|\mathbf{w} - \mathbf{z}\|/\|\mathbf{z}\|$, which gives the desired result. □

The following lemma gives an upper bound on the expected instantaneous regret under the greedy decision $\mathbf{G}(c)$ during the exploitation phase of cycle c.

Lemma 3.6 (Regret Under the Greedy Decision). *Suppose that Assumption 1 holds and the sets* $\mathcal{U}_r$ *satisfy the SBAR*(J) *condition. Then, there exists a positive constant* h_2 *that depends only on* σ_0, $\bar{u}$, λ_0, *and* J, *such that for any* $\mathbf{z} \in \mathbb{R}^r$ *and* $c \geq 1$,

$$\mathsf{E}\left[\max_{\mathbf{u} \in \mathcal{U}^r} \mathbf{z}'(\mathbf{u} - \mathbf{G}(c)) \,\middle|\, \mathbf{Z} = \mathbf{z}\right] \leq \frac{r h_2}{c\,\|\mathbf{z}\|}.$$

Proof. The result is trivially true when $\mathbf{z}=\mathbf{0}$. So, let us fix some $\mathbf{z}\in\mathbb{R}^r\backslash\{\mathbf{0}\}$. By comparing the greedy decision $\mathbf{G}(c)$ with the best arm $\mathbf{u}^*(\mathbf{z})$, we see that the instantaneous regret satisfies

$$\begin{aligned}\mathbf{z}'(\mathbf{u}^*(\mathbf{z})-\mathbf{G}(c)) &= (\mathbf{z}-\hat{\mathbf{Z}}(c))'\mathbf{u}^*(\mathbf{z})+(\mathbf{u}^*(\mathbf{z})-\mathbf{G}(c))'\hat{\mathbf{Z}}(c)+(\hat{\mathbf{Z}}(c)-\mathbf{z})'\mathbf{G}(c)\\ &\le (\mathbf{z}-\hat{\mathbf{Z}}(c))'\mathbf{u}^*(\mathbf{z})+(\hat{\mathbf{Z}}(c)-\mathbf{z})'\mathbf{G}(c)\\ &= (\hat{\mathbf{Z}}(c)-\mathbf{z})'(\mathbf{G}(c)-\mathbf{u}^*(\mathbf{z}))=(\hat{\mathbf{Z}}(c)-\mathbf{z})'(\mathbf{u}^*(\hat{\mathbf{Z}}(c))-\mathbf{u}^*(\mathbf{z})),\end{aligned}$$

where the inequality follows from the definition of the greedy decision in Equation (2), and the final equality follows from the fact that $\mathbf{G}(c)=\mathbf{u}^*(\hat{\mathbf{Z}}(c))$. As a convention, we define $\mathbf{0}/\|\mathbf{0}\|$ to some fixed unit vector and set $\mathbf{u}^*(\mathbf{0})=\mathbf{u}^*(\mathbf{0}/\|\mathbf{0}\|)$.

It then follows from the Cauchy-Schwarz inequality that, with probability one,

$$\begin{aligned}\mathbf{z}'(\mathbf{u}^*(\mathbf{z})-\mathbf{G}(c)) &\le \|\hat{\mathbf{Z}}(c)-\mathbf{z}\|\|\mathbf{u}^*(\hat{\mathbf{Z}}(c))-\mathbf{u}^*(\mathbf{z})\|\\ &= \|\hat{\mathbf{Z}}(c)-\mathbf{z}\|\left\|\mathbf{u}^*\left(\frac{\hat{\mathbf{Z}}(c)}{\|\hat{\mathbf{Z}}(c)\|}\right)-\mathbf{u}^*\left(\frac{\mathbf{z}}{\|\mathbf{z}\|}\right)\right\|\\ &\le J\|\hat{\mathbf{Z}}(c)-\mathbf{z}\|\left\|\frac{\hat{\mathbf{Z}}(c)}{\|\hat{\mathbf{Z}}(c)\|}-\frac{\mathbf{z}}{\|\mathbf{z}\|}\right\|\le\frac{2J\|\hat{\mathbf{Z}}(c)-\mathbf{z}\|^2}{\|\mathbf{z}\|},\end{aligned}$$

where the equality follows from the fact that $\mathbf{u}^*(\mathbf{z})=\mathbf{u}^*(\lambda\mathbf{z})$ for all $\lambda>0$. The second inequality follows from condition SBAR(J), and the final inequality follows from Lemma 3.5. The desired result follows by taking conditional expectations, given $\mathbf{Z}=\mathbf{z}$, and applying Lemma 3.4. □

We can now complete the proof of Theorem 3.1, by adding the regret over the differnt times and cycles. By Assumption 1 and the Cauchy-Schwarz inequality, the instantaneous regret from playing any arm $\mathbf{u}\in\mathcal{U}_r$ is bounded above by $\max_{v\in\mathcal{U}_r}\mathbf{z}'(\mathbf{v}-\mathbf{u})\le 2\bar{u}\|\mathbf{z}\|$. Consider an arbitrary cycle c. Then, the total regret incurred during the exploration phase (with r periods) in this cycle is bounded above by $2\bar{u}r\|\mathbf{z}\|$. During the exploitation phase of cycle c, we always play the greedy arm $\mathbf{G}(c)$. The expected instantaneous regret in each period during the exploitation phase is bounded above by $rh_2/c\,\|\mathbf{z}\|$. So, the total regret during cycle c is bounded above by $2\bar{u}r\|\mathbf{z}\|+h_2r/\|\mathbf{z}\|$. Summing over K cycles, we obtain

$$\text{Regret}\left(\mathbf{z}, rK+\sum_{c=1}^{K}c, \text{PEGE}\right)\le h_3r\|\mathbf{z}\|K+h_4\sum_{c=1}^{K}\frac{r}{\|\mathbf{z}\|},$$

for some positive constants h_3 and h_4 that depend only on σ_0, $\bar{u}$, λ_0, and J.

Consider an arbitrary time period $T\ge r$ and $\mathbf{z}\in\mathbb{R}^r$. Let $K_0=\lceil\sqrt{2T}\rceil$. Note that the total time periods after K_0 cycles is at least T because $rK_0+\sum_{c=1}^{K_0}c\ge\sum_{c=1}^{K_0}c=K_0(K_0+1)/2\ge K_0^2/2\ge T$. Because the cumulative regret is nondecreasing over time, it follows that

$$\begin{aligned}\text{Regret}(\mathbf{z},T,\text{PEGE}) &\le \text{Regret}\left(\mathbf{z}, rK_0+\sum_{c=1}^{K_0}c, \text{PEGE}\right)\\ &\le h_3r\|\mathbf{z}\|K_0+h_4\frac{rK_0}{\|\mathbf{z}\|}\le 3\max\{h_3,\,h_4\}\left(\|\mathbf{z}\|+\frac{1}{\|\mathbf{z}\|}\right)r\sqrt{T},\end{aligned}$$

where the final inequality follows because $K_0=\lceil\sqrt{2T}\rceil\le 3\sqrt{T}$. The risk bound follows by taking expectations and using the assumption on the boundedness of $\mathsf{E}[\|\mathbf{Z}\|]$ and $\mathsf{E}[1/\|\mathbf{Z}\|]$.

4. A policy for general bandits.

We have shown that when a bandit has a smooth best arm response, the PEGE policy achieves optimal $O(r\sqrt{T})$ regret and Bayes risk. The general idea is that when the estimation error is small, the instantaneous regret of the greedy decision based on our estimate $\hat{\mathbf{Z}}(c)$ can be of the same order as $\|\mathbf{Z}-\hat{\mathbf{Z}}(c)\|$. However, under the smoothness assumption, this upper bound on the instantaneous regret is improved to $O(\|\mathbf{Z}-\hat{\mathbf{Z}}(c)\|^2)$, as shown in the proof of Lemma 3.6, and this enables us to separate exploration from exploitation.

However, if the number of arms is finite or if the collection of arms is an arbitrary compact set, then the PEGE policy may not be effective. This is because a small estimation error may have a disproportionately large effect on the arm chosen by a greedy policy, leading to a large instantaneous regret. In this section, we discuss a policy—which we refer to as the Uncertainty Ellipsoid (UE) policy—that can be applied to any bandit problem, at the price of slightly higher regret and Bayes risk. In contrast to the PEGE policy, the UE policy combines active exploration and exploitation in every period.

As discussed in the introduction, the UE policy is closely related to the algorithms described in Auer [4] and Dani et al. [12], but also has the anytime property (the policy does not require prior knowledge of the time horizon T), and also allows the random vector $\mathbf{Z}$ and the errors W_t^u to be unbounded. For the sake of completeness, we give a detailed description of our policy and state the regret and risk bounds that we obtain. The reader can find the proofs of these bounds in Appendix B in Rusmevichientong and Tsitsiklis [30].

To facilitate exposition, we introduce a constant that will appear in the description of the policy, namely,

$$\kappa_0 = 2\sqrt{1+\log\left(1+\frac{36\bar{u}^2}{\lambda_0}\right)}, \tag{3}$$

where the parameters $\bar{u}$ and λ_0 are given in Assumption 1. The UE policy maintains, at each time period t, the following two pieces of information.

(i) The ordinary least squares (OLS) estimate defined as follows: if $\mathbf{U}_1, \ldots, \mathbf{U}_t$ are the arms chosen during the first t periods, then the OLS estimate $\hat{\mathbf{Z}}_t$ is given by[4]

$$\mathbf{C}_t = \left(\sum_{s=1}^{t}\mathbf{U}_s\mathbf{U}_s'\right)^{-1}, \quad \mathbf{M}_t = \sum_{s=1}^{t}\mathbf{U}_s W_s, \qquad \text{and} \qquad \hat{\mathbf{Z}}_t = \mathbf{C}_t\sum_{s=1}^{t}\mathbf{U}_s X_s = \mathbf{Z}+\mathbf{C}_t\mathbf{M}_t. \tag{4}$$

In contrast to the PEGE policy, whose estimates relied only on the rewards observed in the exploration phases, the estimate $\hat{\mathbf{Z}}_t$ incorporates *all* available information up to time t. We initialize the policy by playing r linearly independent arms, so that $\mathbf{C}_t$ is positive definite for $t \geq r$.

(ii) An *uncertainty ellipsoid* $\mathscr{E}_t \subseteq \mathbb{R}^r$ associated with the estimate $\hat{\mathbf{Z}}_t$, defined by

$$\mathscr{E}_t = \{\mathbf{w}\in\mathbb{R}^r\colon \mathbf{w}'\mathbf{C}_t^{-1}\mathbf{w} \leq (\alpha\sqrt{\log t}\sqrt{\min\{r\log t,\, |\mathscr{U}_r|\}})^2\} \qquad \text{and} \qquad \alpha = 4\sigma_0\kappa_0^2, \tag{5}$$

where the parameters σ_0 and κ_0 are given in Assumption 1(a) and Equation (3). The uncertainty ellipsoid $\mathscr{E}_t$ represents the set of likely errors associated with the estimate $\hat{\mathbf{Z}}_t$. We define the *uncertainty radius* R_t^u associated with each arm $\mathbf{u}$ as follows:

$$R_t^u = \max_{\mathbf{v}\in\mathscr{E}_t}\mathbf{v}'\mathbf{u} = \alpha\sqrt{\log t}\sqrt{\min\{r\log t,\, |\mathscr{U}_r|\}}\,\|\mathbf{u}\|_{\mathbf{C}_t}. \tag{6}$$

A formal description of the policy is given below.

Uncertainty Ellipsoid (UE)

Initialization: During the first r periods, play the r linearly independent arms $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_r$ given in Assumption 1(b). Determine the OLS estimate $\hat{\mathbf{Z}}_r$, the uncertainty ellipsoid $\mathscr{E}_r$, and the uncertainty radius associated with each arm.

Description: For $t \geq r+1$, do the following:

(i) Let $\mathbf{U}_t \in \mathscr{U}_r$ be an arm that gives the maximum estimated reward over the ellipsoid $\hat{\mathbf{Z}}_{t-1}+\mathscr{E}_{t-1}$, that is,

$$\mathbf{U}_t = \arg\max_{\mathbf{v}\in\mathscr{U}_r}\left\{\mathbf{v}'\hat{\mathbf{Z}}_{t-1} + \max_{\mathbf{w}\in\mathscr{E}_{t-1}}\mathbf{w}'\mathbf{v}\right\} = \arg\max_{\mathbf{v}\in\mathscr{U}_r}\{\mathbf{v}'\hat{\mathbf{Z}}_{t-1}+R_{t-1}^{v}\}, \tag{7}$$

where the uncertainty radius R_{t-1}^v is defined in Equation (6); ties are broken arbitrarily.

(ii) Play arm $\mathbf{U}_t$ and observe the resulting reward X_t.

(iii) Update the OLS estimate $\hat{\mathbf{Z}}_t$, the uncertainty ellipsoid $\mathscr{E}_t$, and the uncertainty radius R_t^u of each arm $\mathbf{u}$, using the formulas in Equations (4)–(6).

By choosing an arm that maximizes the estimated reward over the ellipsoid $\hat{\mathbf{Z}}_t+\mathscr{E}_t$, our policy involves simultaneous exploitation (via the term $\mathbf{v}'\hat{\mathbf{Z}}_t$) and exploration (via the term $R_t^v = \max_{\mathbf{w}\in\mathscr{E}_t}\mathbf{w}'\mathbf{v}$) in every period. The ellipsoid $\mathscr{E}_t$ reflects the uncertainty in our OLS estimate $\hat{\mathbf{Z}}_t$. It generalizes the classical upper confidence index introduced by Lai and Robbins [23], to account for correlations among the arm rewards. In the special case of r independent arms where $\mathscr{U}_r = \{\mathbf{e}_1, \ldots, \mathbf{e}_r\}$, it is easy to verify that for each arm $\mathbf{e}_l$, the expression

[4] Let us note that we are abusing notation here. Throughout this section $\hat{\mathbf{Z}}_t$ stands for the OLS estimate, which is different from the least mean squares estimator $\mathsf{E}[\mathbf{Z}\mid\mathbf{H}_t]$ introduced in §2.

$\mathbf{e}_l'\hat{\mathbf{Z}}_t + R_t^{e_l}$ coincides (up to a scaling constant) with the upper confidence bound used by Auer et al. [5]. Our definition of the uncertainty radius involves an extra factor of $\sqrt{\min\{r\log t, |\mathcal{U}_r|\}}$, in order to handle the case where the arms are not standard unit vectors, and the rewards are correlated.

The main results of this section are given in the following two theorems. The first theorem establishes upper bounds on the regret and risk when the set of arms is an arbitrary compact set. This result shows that the UE policy is nearly optimal, admitting upper bounds that are within a logarithmic factor of the $\Omega(r\sqrt{T})$ lower bounds given in Theorem 2.1. Although the proof of this theorem makes use of somewhat different (and novel) large deviation inequalities for adaptive least squares estimators, the argument shares similarities with the proofs given in Dani et al. [12], and we omit the details. The reader can find a complete proof in Appendix B.2 in Rusmevichientong and Tsitsiklis [30].

Theorem 4.1 (Bounds for General Compact Sets of Arms). *Under Assumption 1, there exist positive constants a_4 and a_5 that depend only on the parameters σ_0, $\bar{u}$, and λ_0, such that for all $T \ge r+1$ and $\mathbf{z} \in \mathbb{R}^r$,*

$$\text{Regret}(\mathbf{z}, T, \text{UE}) \le a_4 r\,\|\mathbf{z}\| + a_5 r\sqrt{T}\log^{3/2} T,$$

and

$$\text{Risk}(T, \text{UE}) \le a_4 r \mathsf{E}[\|\mathbf{Z}\|] + a_5 r\sqrt{T}\log^{3/2} T.$$

For any arm $\mathbf{u} \in \mathcal{U}_r$ and $\mathbf{z} \in \mathbb{R}^r$, let $\Delta^u(\mathbf{z})$ denote the difference between the maximum expected reward and the expected reward of arm $\mathbf{u}$ when $\mathbf{Z} = \mathbf{z}$, that is,

$$\Delta^u(\mathbf{z}) = \max_{\mathbf{v}\in\mathcal{U}_r} \mathbf{v}'\mathbf{z} - \mathbf{u}'\mathbf{z}.$$

When the number of arms is finite, it turns out that we can obtain bounds on regret and risk that scale more gracefully over time, growing as $\log T$ and $\log^2 T$, respectively. This result is stated in Theorem 4.2, which shows that, for a fixed set of arms, the UE policy is asymptotically optimal as a function time, within a constant factor of the lower bounds established by Lai and Robbins [23] and Lai [22].

Theorem 4.2 (Bounds for Finitely Many Arms). *Under Assumption 1, there exist positive constants a_6 and a_7 that depend only on the parameters σ_0, $\bar{u}$, and λ_0 such that for all $T \ge r+1$ and $\mathbf{z} \in \mathbb{R}^r$,*

$$\text{Regret}(\mathbf{z}, T, \text{UE}) \le a_6\,|\mathcal{U}_r|\,\|\mathbf{z}\| + a_7\,|\mathcal{U}_r| \sum_{\mathbf{u}\in\mathcal{U}_r} \min\left\{\frac{\log T}{\Delta^u(\mathbf{z})},\, T\Delta^u(\mathbf{z})\right\}.$$

Moreover, suppose that there exists a positive constant M_0 such that, for all arms $\mathbf{u}$, the distribution of the random variable $\Delta^u(\mathbf{Z})$ is described by a point mass at 0, and a density function that is bounded above by M_0 on $\mathbb{R}_+$. Then, there exist positive constants a_8 and a_9 that depend only on the parameters σ_0, $\bar{u}$, λ_0, and M_0, such that for all $T \ge r+1$,

$$\text{Risk}(T, \text{UE}) \le a_8\,|\mathcal{U}_r|\,\mathsf{E}[\|\mathbf{Z}\|] + a_9\,|\mathcal{U}_r|^2 \log^2 T.$$

Proof. For any arm $\mathbf{u} \in \mathcal{U}_r$ and $\mathbf{z} \in \mathbb{R}^r$, let the random variable $N^u(\mathbf{z}, T)$ denote the total number of times that the arm $\mathbf{u}$ is chosen during periods 1 through T, given that $\mathbf{Z} = \mathbf{z}$. Using an argument similar to the one in Auer et al. [5], we can show that

$$\mathsf{E}[N^u(\mathbf{z}, T) \mid \mathbf{Z} = \mathbf{z}] \le 6 + \frac{4\alpha^2\,|\mathcal{U}_r|\log T}{(\Delta^u(\mathbf{z}))^2}.$$

The reader can find a proof of this result in Appendix B.3 in Rusmevichientong and Tsitsiklis [30].

The regret bound in Theorem 4.2 then follows immediately from the above upper bound and the fact that $N^u(\mathbf{z}, T) \le T$ with probability one, because

$$\begin{aligned}\text{Regret}(\mathbf{z}, T, \text{UE}) &= \sum_{u\in\mathcal{U}_r} \Delta^u(\mathbf{z})\mathsf{E}[N^u(\mathbf{z}, T) \mid \mathbf{Z} = \mathbf{z}] \le \sum_{u\in\mathcal{U}_r} \Delta^u(\mathbf{z}) \min\left\{6 + \frac{4\alpha^2\,|\mathcal{U}_r|\log T}{(\Delta^u(\mathbf{z}))^2},\, T\right\}\\ &\le 6\sum_{u\in\mathcal{U}_r} \Delta^u(\mathbf{z}) + \max\{4\alpha^2, 1\}\,|\mathcal{U}_r| \sum_{u\in\mathcal{U}_r} \min\left\{\frac{\log T}{\Delta^u(\mathbf{z})},\, T\Delta^u(\mathbf{z})\right\},\end{aligned}$$

and the desired result follows from the fact that $\Delta^u(\mathbf{z}) = \max_{v\in\mathcal{U}_r}(\mathbf{v} - \mathbf{u})'\mathbf{z} \le 2\bar{u}\,\|\mathbf{z}\|$, by the Cauchy-Schwarz inequality.

We will now establish an upper bound on the Bayes risk. From the regret bound, it suffices to show that for any $\mathbf{u} \in \mathcal{U}_r$,

$$\mathsf{E}\left[\min\left\{\frac{\log T}{\Delta^u(\mathbf{Z})},\, T\Delta^u(\mathbf{Z})\right\}\right] \le (M_0+1)\log T + M_0 \log^2 T.$$

Let $q^u(\cdot)$ denote the density function associated with the random variable $\Delta^u(\mathbf{Z})$. Then,

$$\begin{aligned}\mathsf{E}\left[\min\left\{\frac{\log T}{\Delta^u(\mathbf{Z})}, T\Delta^u(\mathbf{Z})\right\}\right] &= \int_0^{\sqrt{(\log T)/T}} \min\left\{\frac{\log T}{x},\, Tx\right\} q^u(x)\,dx \\ &\quad + \int_{\sqrt{(\log T)/T}}^1 \min\left\{\frac{\log T}{x},\, Tx\right\} q^u(x)\,dx + \int_1^\infty \min\left\{\frac{\log T}{x},\, Tx\right\} q^u(x)\,dx.\end{aligned}$$

We will now proceed to bound each of the three terms on the right-hand side of the above equality. Having assumed that $q^u(\cdot) \le M_0$, the first term satisfies

$$\int_0^{\sqrt{(\log T)/T}} \min\left\{\frac{\log T}{x},\, Tx\right\} q^u(x)\,dx \le M_0 \int_0^{\sqrt{(\log T)/T}} Tx\,dx = M_0 T \frac{x^2}{2}\bigg|_0^{\sqrt{(\log T)/T}} \le M_0 \log T.$$

For the second term, note that

$$\begin{aligned}\int_{\sqrt{(\log T)/T}}^1 \min\left\{\frac{\log T}{x},\, Tx\right\} q^u(x)\,dx &\le M_0 \int_{\sqrt{(\log T)/T}}^1 \frac{\log T}{x}\,dx = M_0 \log T \cdot \left(\log x \,|_{\sqrt{(\log T)/T}}^1\right) \\ &= M_0(\log T)\cdot \frac{\log T - \log\log T}{2} \le M_0 \log^2 T,\end{aligned}$$

where the last inequality follows from the fact that $\log T - \log\log T \le 2\log T$ for all $T \ge 2$. To evaluate the last term, note that $(\log T)/x \le \log T$ for all $x \ge 1$, and thus, $\int_1^\infty \min\{(\log T)/x,\, Tx\} q^u(x)\,dx \le \log T \int_1^\infty q^u(x) \le \log T$. Putting everything together, we have that $\mathsf{E}[\min\{(\log T)/(\Delta^u(\mathbf{Z})),\, T\Delta^u(\mathbf{Z})\}] \le (M_0+1)\log T + M_0 \log^2 T$, which is the desired result. □

We conclude this section by giving an example of a random vector $\mathbf{Z}$ that satisfies the condition in Theorem 4.2. A similar example also appears in Example 2 of Lai [22].

Example 4.3 (IID Random Variables). Suppose $\mathcal{U}_r = \{\mathbf{e}_1, \ldots, \mathbf{e}_r\}$ and $\mathbf{Z} = (Z_1, \ldots, Z_r)$, where the random variables Z_k are independent and identically distributed with a common cumulative distribution function F and a density function $f\colon \mathbb{R} \to \mathbb{R}$ that is bounded above by M. Then, for each k, the random variable $\Delta^{e_k}(\mathbf{Z})$ is given by $\Delta^{e_k}(\mathbf{Z}) = (\max_{j=1,\ldots,r} Z_j) - Z_k = \max\{0, \max_{j\ne k}\{Z_j - Z_k\}\}$. It is easy to verify that $\Delta^{e_k}(\mathbf{Z})$ has a point mass at zero and a continuous density function $q_k(\cdot)$ on $\mathbb{R}_+$ given by: for any $x > 0$,

$$q_k(x) = (r-1)\int \{F(z_k + x)\}^{r-2} f(z_k + x) f(z_k)\,dz_k \le (r-1)M.$$

4.1. Regret bounds for polyhedral sets of arms. In this section, we focus on the regret profiles when the set of arms $\mathcal{U}_r$ is a polyhedral set. Let $\mathcal{E}(\mathcal{U}_r)$ denote the set of extreme points of $\mathcal{U}_r$. From a standard result in linear programming, for all $\mathbf{z} \in \mathbb{R}^r$,

$$\max_{u\in\mathcal{U}_r} \mathbf{u}'\mathbf{z} = \max_{u\in\mathcal{E}(\mathcal{U}_r)} \mathbf{u}'\mathbf{z}.$$

Because a polyhedral set has a finite number of extreme points ($|\mathcal{E}(\mathcal{U}_r)| < \infty$), the parameterized bandit problem can be reduced to the standard multi-armed bandit problem, where each arm corresponds to an extreme point of $\mathcal{U}_r$. We can thus apply the algorithm of Lai and Robbins [23] and obtain the following upper bound on the T-period cumulative regret for polyhedra

$$\text{Regret}(\mathbf{z}, T, \text{Lai's Algorithm}) = O\left(\frac{|\mathcal{E}(\mathcal{U}_r)|\cdot \log T}{\min\{\Delta^u(\mathbf{z})\colon \Delta^u(\mathbf{z}) > 0\}}\right), \tag{8}$$

where the denominator corresponds to the difference between the expected reward of the optimal and the second best extreme points. The algorithm of Lai and Robbins [23] is effective only when the polyhedral set $\mathcal{U}_r$ has a small number of extreme points, as shown by the following examples.

Example 4.4 (Simplex). Suppose $\mathcal{U}_r = \{\mathbf{u} \in \mathbb{R}^r\colon \sum_{i=1}^r |u_i| \le 1\}$ is an r-dimensional unit simplex. Then, $\mathcal{U}_r$ has $2r$ extreme points, and Equation (8) gives an $O(r\log T)$ upper bound on the regret.

EXAMPLE 4.5 (LINEAR CONSTRAINTS). Suppose that $\mathcal{U}_r = \{\mathbf{u} \in \mathbb{R}^r\colon \mathbf{A}\mathbf{u} \le \mathbf{b} \text{ and } \mathbf{u} \ge \mathbf{0}\}$, where $\mathbf{A}$ is a $p \times r$ matrix with $p \le r$. It follows from the standard linear programming theory that every extreme point is a basic feasible solution, which has at most p nonzero coordinates (see, for example, Bertsimas and Tsitsiklis [9]). Thus, the number of extreme points is bounded above by $\binom{r+p}{p} = O((2r)^p)$, and Equation (8) gives an $O((2r)^p \log T)$ upper bound on the regret.

In general, the number of extreme points of a polyhedron can be very large, rendering the bandit algorithm of Lai and Robbins [23] ineffective; consider, for example, the r-dimensional cube $\mathcal{U}_r = \{\mathbf{u} \in \mathbb{R}^r\colon |u_i| \le 1 \text{ for all } i\}$, which has 2^r extreme points. Moreover, we cannot apply the results and algorithms from §3 to the convex hull of $\mathcal{U}_r$. This is because the convex hull of a polyhedron is *not* strongly convex (it cannot be written as an intersection of Euclidean balls), and thus, it does not satisfy the required SBAR$(\cdot)$ condition in Theorem 3.1. The UE policy in the previous section gives $O(r\sqrt{T}\log^{3/2} T)$ regret and risk upper bounds. However, finding an algorithm specifically for polyhedral sets that yields an $O(r\sqrt{T})$ regret upper bound (without an additional logarithmic factor) remains an open question.

5. Conclusion.

We analyzed a class of multi-armed bandit problems where the expected reward of each arm depends linearly on an unobserved random vector $\mathbf{Z} \in \mathbb{R}^r$, with $r \ge 2$. Our model allows for correlations among the rewards of different arms. When we have a smooth best arm response, we showed that a policy that alternates between exploration and exploitation is optimal. For a general bandit, we proposed a near-optimal policy that performs active exploration in every period. For finitely many arms, our policy achieves asymptotically optimal regret and risk as a function of time, but scales with the square of the number of arms. Improving the dependence on the number of arms remains an open question. It would also be interesting to study more general correlation structures. Our formulation assumes that the vector of expected rewards lies in an r-dimensional subspace spanned by a known set of basis functions that describe the characteristics of the arms. Extending our work to a setting where the basis functions are unknown has the potential to broaden the applicability of our model.

Appendix A. Properties of normal vectors.

In this section, we prove that if $\mathbf{Z}$ has a multivariate normal distribution with mean $\mathbf{0} \in \mathbb{R}^r$ and covariance matrix $\mathbf{I}_r/r$, then $\mathbf{Z}$ has the properties described in Lemmas 2.4 and 3.2.

A.1. Proof of Lemma 2.4.

We want to establish a lower bound on $\Pr\{\theta \le \|\mathbf{Z}\| \le \beta\}$. Let $\mathbf{Y} = (Y_1, \ldots, Y_r)$ denote the standard multivariate normal random vector with mean $\mathbf{0}$ and identity covariance matrix $\mathbf{I}_r$. By our hypothesis, $\mathbf{Z}$ has the same distribution as $\mathbf{Y}/\sqrt{r}$, which implies that

$$\Pr\{\theta \le \|\mathbf{Z}\| \le \beta\} = \Pr\{\theta\sqrt{r} \le \|\mathbf{Y}\| \le \beta\sqrt{r}\} = 1 - \Pr\{\|\mathbf{Y}\|^2 < \theta^2 r\} - \Pr\{\|\mathbf{Y}\|^2 > \beta^2 r\}.$$

By definition, $\|\mathbf{Y}\|^2 = Y_1^2 + \cdots + Y_r^2$ has a chi-square distribution with r degrees of freedom. By the Markov inequality, $\Pr\{\|\mathbf{Y}\|^2 > \beta^2 r\} \le \mathsf{E}[\|\mathbf{Y}\|^2]/(\beta^2 r) = 1/\beta^2$. We will now establish an upper bound on $\Pr\{\|\mathbf{Y}\|^2 < \theta^2 r\}$. Note that, for any $\lambda > 0$,

$$\Pr\{\|\mathbf{Y}\|^2 < \theta^2 r\} = \Pr\{e^{-\lambda\sum_{k=1}^r Y_k^2} > e^{-\lambda\theta^2 r}\} \le e^{\lambda\theta^2 r}\cdot \mathsf{E}\left[\prod_{k=1}^r e^{-\lambda Y_k^2}\right] = \left(\frac{e^{\lambda\theta^2}}{\sqrt{1+2\lambda}}\right)^r,$$

where last equality follows from the fact that $Y_1, \ldots, Y_r$ are independent standard normal random variables and thus, $\mathsf{E}[e^{-\lambda Y_k^2}] = 1/\sqrt{1+2\lambda}$ for $\lambda > 0$. Set $\lambda = 1/\theta^2$, and use the facts $\theta \le 1/2 \le \sqrt{2}/e$ and $r \ge 2$, to obtain

$$\Pr\{\|\mathbf{Y}\|^2 < \theta^2 r\} \le \left(\frac{e\theta}{\sqrt{2+\theta^2}}\right)^r \le \left(\frac{e\theta}{\sqrt{2}}\right)^r \le \left(\frac{e\theta}{\sqrt{2}}\right)^2 = \frac{e^2\theta^2}{2} \le 4\theta^2,$$

which implies that $\Pr\{\theta \le \|\mathbf{Z}\| \le \beta\} \ge 1 - 1/\beta^2 - 4\theta^2$, which is the desired result.

A.2. Proof of Lemma 3.2.

For part (a) of the lemma, we have

$$\mathsf{E}[1/\|\mathbf{Z}\|] = \int_0^\infty \frac{1}{x} g(x)\,dx \le M_0 \int_0^\rho x^{\rho-1}\,dx + \frac{1}{\rho}\int_\rho^\infty g(x)\,dx \le M_0\frac{\rho^\rho}{\rho} + \frac{1}{\rho}.$$

For the proof of part (b), let $\mathbf{Y} = (Y_1, \ldots, Y_r)$ be a standard multivariate normal random vector with mean $\mathbf{0}$ and identity covariance matrix, $\mathbf{I}_r$. Then, $\mathbf{Z}$ has the same distribution as $\mathbf{Y}/\sqrt{r}$. Note that $\|\mathbf{Y}\|^2$ has a chi-square distribution with r degrees of freedom. Thus,

$$\mathsf{E}[\|\mathbf{Z}\|] = \frac{1}{\sqrt{r}}\mathsf{E}[\|\mathbf{Y}\|] \le \frac{1}{\sqrt{r}}\sqrt{\mathsf{E}[\|\mathbf{Y}\|^2]} = \frac{1}{\sqrt{r}}\sqrt{r} = 1.$$

We will now establish an upper bound on $\mathsf{E}[1/\|\mathbf{Z}\|] = \sqrt{r}\mathsf{E}[1/\|\mathbf{Y}\|]$. For $r = 2$, because $\|\mathbf{Y}\|$ has a chi distribution with two degrees of freedom, we have that

$$\mathsf{E}[1/\|\mathbf{Z}\|] = \sqrt{2}\int_0^\infty \frac{1}{x}\cdot xe^{-x^2/2}\,dx = \sqrt{2}\int_0^\infty e^{-x^2/2}\,dx = \sqrt{\pi}.$$

Consider the case where $r \geq 3$. Then,

$$\mathsf{E}[1/\|\mathbf{Z}\|] = \sqrt{r}\mathsf{E}[1/\|\mathbf{Y}\|] \leq \sqrt{r}\sqrt{\mathsf{E}[1/\|\mathbf{Y}\|^2]}.$$

Using the formula for the density of the chi-square distribution, we have

$$\begin{aligned}\mathsf{E}[1/\|\mathbf{Y}\|^2] &= \int_0^\infty \frac{1}{x}\cdot\frac{1}{2^{r/2}\Gamma(r/2)}x^{(r/2)-1}e^{-x/2}\,dx\\ &= \frac{2^{(r/2)-1}}{2^{r/2}}\cdot\frac{\Gamma((r/2)-1)}{\Gamma(r/2)}\cdot\int_0^\infty\frac{1}{2^{(r-2)/2}\Gamma((r-2)/2)}x^{((r-2)/2)-1}e^{-x/2}\,dx\\ &= \frac{1}{2((r/2)-1)} = \frac{1}{r-2} \leq \frac{3}{r},\end{aligned}$$

where the third equality follows from the fact that $\Gamma(r/2) = ((r/2)-1)\cdot\Gamma((r/2)-1)$ for $r \geq 3$ and the integrand is the density function of the chi-square distribution with $r-2$ degrees of freedom and evaluates to one. The last inequality follows because $r \geq 3$. Thus, we have $\mathsf{E}[1/\|\mathbf{Z}\|] \leq \sqrt{3} \leq \sqrt{\pi}$, which is the desired result.

Acknowledgments. The authors would like to thank Adrian Lewis and Mike Todd for helpful and stimulating discussions on the structure of positive definite matrices and the eigenvalues associated with least squares estimators, Gena Samorodnitsky for sharing his deep insights on the application of large deviation theory to this problem, Adam Mersereau for his contributions to the problem formulation, and Assaf Zeevi for helpful suggestions and discussions during the first author's visit to Columbia Graduate School of Business. We also want to thank the associate editor and the referee for their helpful comments and suggestions on the paper. This research is supported in part by the National Science Foundation through grants DMS-0732196, ECCS-0701623, CMMI-0856063, and CMMI-0855928.

References

[1] Abe, N., P. M. Long. 1999. Associative reinforcement learning using linear probabilistic concepts. *Proc. 16th Internat. Conf. Machine Learn.*, Morgan Kaufman, San Francisco, 3–11.

[2] Agrawal, R. 1995. Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Adv. Appl. Probab.* **27**(4) 1054–1078.

[3] Agrawal, R., D. Teneketzis, V. Anantharam. 1989. Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: Finite parameter space. *IEEE Trans. Automatic Control* **34**(3) 258–267.

[4] Auer, P. 2002. Using confidence bounds for exploitation-exploration trade-offs. *J. Machine Learn. Res.* **3**(3) 397–422.

[5] Auer, P., N. Cesa-Bianchi, P. Fischer. 2002. Finite-time analysis of the multi-armed bandit problem. *Machine Learn.* **47**(2) 235–256.

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

[7] Bertsekas, D. 1995. *Dynamic Programming and Optimal Controls*, Vol. 1. Athena Scientific, Belmont, MA.

[8] Bertsekas, D., J. N. Tsitsiklis. 1996. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.

[9] Bertsimas, D., J. N. Tsitsiklis. 1997. *Introduction to Linear Optimization*. Athena Scientific, Belmont, MA.

[10] Blum, J. R. 1954. Multidimensional stochastic approximation methods. *Ann. Math. Statist.* **25**(4) 737–744.

[11] Cicek, D., M. Broadie, A. Zeevi. 2009. General bounds and finite-time performance improvement for the Kiefer-Wolfowitz stochastic approximation algorithm. Working paper, Columbia Graduate School of Business, New York.

[12] Dani, V., T. P. Hayes, S. M. Kakade. 2008a. Stochastic linear optimization under bandit feedback. *Proc. 21st Annual Conf. Learn. Theory* (*COLT 2008*), *Helsinki, Finland*, 355–366.

[13] Dani, V., T. P. Hayes, S. M. Kakade. 2008b. Stochastic linear optimization under bandit feedback. Working paper, University of Chicago, Chicago. http://ttic.uchicago.edu/~sham/papers/ml/bandit_linear_long.pdf.

[14] Feldman, D. 1962. Contributions to the "two-armed bandit" problem. *Ann. Math. Statist.* **33**(3) 847–856.

[15] Fiedler, M., V. Pták. 1997. A new positive definite geometric mean of two positive definite matrices. *Linear Algebra Its Appl.* **251**(1) 1–20.

[16] Ginebra, J., M. K. Clayton. 1995. Response surface bandits. *J. Roy. Statist. Soc. Ser. B* (*Methodological*) **57**(4) 771–784.

[17] Goldenshluger, A., A. Zeevi. 2008. Performance limitations in bandit problems with side observations. Working paper, Columbia Graduate School of Business, Columbia University Graduate School of Business, New York.

[18] Goldenshluger, A., A. Zeevi. 2009. Woodroofe's one-armed bandit problem revisited. *Ann. Appl. Probab.* **19**(4) 1603–1633.

[19] Keener, R. 1985. Further contributions to the "two-armed bandit" problem. *Ann. Statist.* **13**(1) 418–422.

[20] Kiefer, J., J. Wolfowitz. 1952. Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.* **23**(3) 462–466.

[21] Lai, T. 2003. Stochastic approximation (invited paper). *Ann. Statist.* **31**(2) 391–406.

[22] Lai, T. L. 1987. Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Statist.* **15**(3) 1091–1114.

[23] Lai, T. L., H. Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* **6**(1) 4–22.

[24] Mersereau, A. J., P. Rusmevichientong, J. N. Tsitsiklis. 2009. A structured multi-armed bandit problem and the greedy policy. *IEEE Trans. Automatic Control* **54**(12) 2787–2802.

[25] Pandey, S., D. Chakrabarti, D. Agrawal. 2007. Multi-armed bandit problems with dependent arms. *Proc. 24th Internat. Conf. Machine Learn., Corvallis, OR*, 721–728.

[26] Polovinkin, E. S. 1996. Strongly convex analysis. *Sbornik: Math.* **187**(2) 259–286.

[27] Pressman, E. L., I. N. Sonin. 1990. *Sequential Control with Incomplete Information*. Academic Press, London.

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

[29] Robbins, H., S. Monro. 1951. A stochastic approximation method. *Ann. Math. Statist.* **22**(3) 400–407.

[30] Rusmevichientong, P., J. N. Tsitsiklis. 2010. Linearly parameterized bandits (extended version). http://arxiv.org/abs/0812.3465.

[31] Thompson, W. R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25**(3) 285–294.

[32] Wang, C.-C., S. R. Kulkarni, H. V. Poor. 2005a. Bandit problems with side observations. *IEEE Trans. Automatic Control* **50**(3) 338–355.

[33] Wang, C.-C., S. R. Kulkarni, H. V. Poor. 2005b. Arbitrary side observations in bandit problems. *Adv. Appl. Math.* **34**(4) 903–938.