Stochastic Linear Optimization Under Bandit Feedback

Varsha Dani
*University of Chicago*

Thomas P. Hayes
*Toyota Technological Institute*

Sham M. Kakade
*University of Pennsylvania*

Follow this and additional works at: [http://repository.upenn.edu/statistics_papers](http://repository.upenn.edu/statistics_papers)

Part of the [Computer Sciences Commons](http://repository.upenn.edu/computer_science_commons), and the [Statistics and Probability Commons](http://repository.upenn.edu/statistics_probability_commons)

Recommended Citation


This paper is posted at ScholarlyCommons. [http://repository.upenn.edu/statistics_papers/101](http://repository.upenn.edu/statistics_papers/101)

For more information, please contact repository@pobox.upenn.edu.
Stochastic Linear Optimization Under Bandit Feedback

Abstract
In the classical stochastic $k$-armed bandit problem, in each of a sequence of $T$ rounds, a decision maker chooses one of $k$ arms and incurs a cost chosen from an unknown distribution associated with that arm. The goal is to minimize regret, defined as the difference between the cost incurred by the algorithm and the optimal cost. In the linear optimization version of this problem (first considered by Auer [2002]), we view the arms as vectors in $\mathbb{R}^n$, and require that the costs be linear functions of the chosen vector. As before, it is assumed that the cost functions are sampled independently from an unknown distribution. In this setting, the goal is to find algorithms whose running time and regret behave well as functions of the number of rounds $T$ and the dimensionality $n$ (rather than the number of arms, $k$, which may be exponential in $n$ or even infinite).

We give a nearly complete characterization of this problem in terms of both upper and lower bounds for the regret. In certain special cases (such as when the decision region is a polytope), the regret is $\text{polylog}(T)$. In general though, the optimal regret is $\Theta^*(\sqrt{T})$ — our lower bounds rule out the possibility of obtaining $\text{polylog}(T)$ rates in general.

We present two variants of an algorithm based on the idea of “upper confidence bounds.” The first, due to Auer [2002], but not fully analyzed, obtains regret whose dependence on $n$ and $T$ are both essentially optimal, but which may be computationally intractable when the decision set is a polytope. The second version can be efficiently implemented when the decision set is a polytope (given as an intersection of half-spaces), but gives up a factor of $\sqrt{n}$ in the regret bound. Our results also extend to the setting where the set of allowed decisions may change over time.

Disciplines
Computer Sciences | Statistics and Probability

This conference paper is available at ScholarlyCommons: http://repository.upenn.edu/statistics_papers/101
Stochastic Linear Optimization under Bandit Feedback

Varsha Dani* and Thomas P. Hayes† and Sham M. Kakade‡

Abstract

In the classical stochastic \( k \)-armed bandit problem, in each of a sequence of \( T \) rounds, a decision maker chooses one of \( k \) arms and incurs a cost chosen from an unknown distribution associated with that arm. The goal is to minimize regret, defined as the difference between the cost incurred by the algorithm and the optimal cost.

In the linear optimization version of this problem (first considered by [Auer, 2002]), we view the arms as vectors in \( \mathbb{R}^n \), and require that the costs be linear functions of the chosen vector. As before, it is assumed that the cost functions are sampled independently from an unknown distribution. In this setting, the goal is to find algorithms whose running time and regret behave well as functions of the number of rounds \( T \) and the dimensionality \( n \) (rather than the number of arms, \( k \), which may be exponential in \( n \) or even infinite).

We give a nearly complete characterization of this problem in terms of both upper and lower bounds for the regret. In certain special cases (such as when the decision region is a polytope), the regret is polylog\((T)\). In general though, the optimal regret is \( \Theta^*(\sqrt{T}) \) — our lower bounds rule out the possibility of obtaining polylog\((T)\) rates in general.

We present two variants of an algorithm based on the idea of “upper confidence bounds.” The first, due to [Auer, 2002], but not fully analyzed, obtains regret whose dependence on \( n \) and \( T \) are both essentially optimal, but which may be computationally intractable when the decision set is a polytope. The second version can be efficiently implemented when the decision set is a polytope (given as an intersection of half-spaces), but gives up a factor of \( \sqrt{n} \) in the regret bound.

Our results also extend to the setting where the set of allowed decisions may change over time.

1 Introduction

The seminal work of [Robbins, 1952] introduced a formalism for studying the sequential design of experiments, which is now referred to as the multi-armed bandit problem. In this foundational paradigm, at each time step a decision maker chooses one of \( K \) decisions or “arms” (e.g., treatments, job schedules, manufacturing processes, etc) and receives some feedback loss only for the chosen decision. In the most unadorned model, it is assumed that the cost for each decision is independently sampled from some fixed underlying (and unknown) distribution (that is different for each decision). The goal of the decision maker is to minimize the average loss over some time horizon. This basic model of decision making under uncertainty already typifies the conflict between minimizing the immediate loss and gathering information that will be useful in the long-run. This sequential design problem — often referred to as the stochastic multi-armed bandit problem — and a long line of successor bandit problems have been extensively studied in the statistics community (see, e.g., [Berry and Fristedt, 1985]), with close attention paid to obtaining sharp convergence rates.

While this paradigm offers a formalism to a host of natural decision problems (e.g. clinical treatment, manufacturing processes, job scheduling), a vital issue to address for applicability to modern problems is how to tackle a set of feasible decisions that is often large (or infinite). For example, the classical bandit problem of clinical treatments (often considered in statistics) — where each decision is a choice of one of \( K \) treatments — is often better modelled by choosing from some (potentially infinite) set of mixed treatments subject to some budget constraint (where there is a cost per unit amount of each of drug). In manufacturing problems, often the goal is to maximize revenue subject to choosing among some large set of decisions that satisfy certain manufacturing constraints (where the revenue from each decision may be unknown). A modern variant of this problem that is receiving increasing attention is the routing problem where the goal is to send packets from \( A \) to \( B \) and the cost of each route is unknown (see, e.g., [Awerbuch and Kleinberg, 2004]).

We study a natural extension of the stochastic multi-armed bandit problem to linear optimization — a problem first considered in [Auer, 2002]. Here, we assume the decision space is an arbitrary subset \( D \subset \mathbb{R}^n \) and that there is fixed distribution \( \pi \) over cost functions. At each round, the learner chooses a decision \( x \in D \), then a cost function \( f(\cdot) : D \rightarrow [0,1] \) is
sampled from \( \pi \). Only the loss \( f(x) \) is revealed to the learner (and not the function \( f(\cdot) \)). We assume that the expected loss is a fixed linear function, i.e., that \( \mathbb{E}[f(x)] = \mu \cdot x \), where the expectation is with respect to \( f \) sampled from \( \pi \) (technically, we make a slightly weaker assumption, precisely stated in the next section). The goal is to minimize the total loss over \( T \) steps. As is standard, success is measured by the regret — the difference between the performance of the learner and that of the optimal algorithm which has knowledge of \( \pi \). Note that the optimal algorithm here simply chooses the best decision with respect to the linear mean vector \( \mu \).

Perhaps the most important and natural example in this paradigm is the (stochastic) online linear programming problem. Here, \( D \) is specified by linear inequality constraints. If the mean \( \mu \) were known, then this is simply a linear programming problem. Instead, at each round, the learner only observes noisy feedback of the chosen decision, with respect to the underlying linear cost function.

### 1.1 Summary of Our Results and Related Work

\[ \text{Auer} \ [2002] \] provides the first analysis of this problem. This paper builds and improves upon the work of \[ \text{Auer} \ [2002] \] in a number of ways. A related model was considered by \[ \text{Abe and Long} \ [1999] \], where the decision sets are allowed to vary as a function of the time. Our results can be extended to this more general model, which we discuss in Section 4.7.

While \[ \text{Auer} \ [2002] \] provides an elegant deterministic algorithm, based on upper confidence bounds of \( \mu \), an analysis of the performance of this algorithm was not provided, due to rather subtle independence issues (though it was conjectured that this simple algorithm was sufficient). Instead, a more complicated master algorithm was analyzed — this master algorithm called the simpler upper confidence algorithm as a subroutine. In this work, we directly analyze the simpler upper confidence algorithm. Unfortunately, implementing this algorithm in certain cases (when \( D \) is large or infinite) may be inefficient. However, we also provide a modification to this algorithm (that uses a different confidence region based on the \( L_1 \)-norm), which may be implemented efficiently for the case when \( D \) is an (infinite) convex set, given certain oracle optimization access to \( D \).

The analysis of \[ \text{Auer} \ [2002] \] achieves a regret bound of \( O^*(\ellog |D|)^{3/2}\poly(n)\sqrt{T} \) where \( n \) is dimension of the decision space, \( T \) is the time horizon, and \( |D| \) is the number of feasible decisions. For the simpler upper confidence algorithm, we show that it enjoys a bound of \( O^*(n\sqrt{T}) \), which does not depend on the cardinality of the decision region, \( |D| \). While this algorithm may be inefficient in some cases, we also provide an efficient algorithm (that uses a slightly different confidence region), which achieves a slightly worse bound of \( O^*(n^{3/2}/\sqrt{T}) \). Using the result in \[ \text{Auer} \ [2002] \], one can also derive a bound of the form \( O(\poly(n)\sqrt{T}) \) for infinite decision sets by appealing to a naive (inefficient) covering argument (where the algorithm is run on an appropriately fine cover of \( D \)). However, this argument results in a less sharp bound in terms of \( n \), though a better reduction to \( \text{Auer} \ [2002] \) may be possible.

For the case of finite decision sets, such as the \( K \)-arm bandit case, a regret that is only logarithmic in the time horizon is achievable. In particular, in a different line of work, \[ \text{Auer et al.} \ [2002] \] showed that the optimal regret for the \( K \)-arm bandit case was characterized as \( \frac{\Delta}{4} \log T \), where \( \Delta \) is the “gap” between the performance of the best arm and the second best arm. This result is stated in terms of the problem dependent constant \( \Delta \), so one can view it as the asymptotic regret for a given problem. In fact, historically, there is long line of work in the \( K \)-arm bandit literature (e.g., \[ \text{Lai and Robbins,} \ [1985], \text{ Agrawal,} \ [1995] \]) concerned with obtaining optimal rates for a fixed problem, which are often logarithmic in \( T \) when stated in terms of some problem dependent constant.

Hence, in our setting, in the case where \( |D| \) is finite, we know that a log rate in the time is achievable by a direct reduction to the \( K \)-arm bandit case (though this naive reduction results in an exponentially worse dependence in terms of \( |D| \)). Our work shows that a regret of \( \frac{\Delta}{2} \polylog(T) \) can be achieved, where \( \Delta \) is a generalized definition of the gap that is appropriate for a potentially infinite \( D \). Hence, a polylogarithmic rate in \( T \) is achievable with a constant that is only polynomial in \( n \) and has no dependence on the size of the (potentially infinite) decision region. Here, \( \Delta \) can be thought of as the gap between the values of the best and second best extremal points of the decision set (which we define precisely later). For example, if \( D \) is a polytope, then \( \Delta \) is the gap in value between the first and second best corner decisions. For the case where \( D \) is finite, \( \Delta \) is exactly the same as in the \( K \)-arm case. However, for some natural decision regions, such as a sphere, \( \Delta = 0 \) so this (problem dependent) bound is not applicable. Note that \( \Delta \) is never 0 for the \( K \)-arm case (unless there is effectively one arm), so a logarithmic rate in \( T \) is always possible in the \( K \)-arm case.

Note that this set of results still raises the question of whether there is an algorithm achieving polylogarithmic regret (as a function of \( T \)) for the case when \( \Delta = 0 \), which could be characterized in terms of some different, more appropriate problem dependent constant. Our final contribution answers this question in the negative. We provide a lower bound showing that the regret of any algorithm on a particular problem (which we construct with \( \Delta = 0 \)) is \( \Omega(n/\sqrt{T}) \). In addition to showing that a polylogarithmic rate is not achievable in general, it also shows our upper bound is tight in terms of \( n \) and \( T \). Note this result is in stark contrast to the \( K \)-arm case where the optimal asymptotic regret for any given problem is always logarithmic in \( T \).

We should also note that the lower bound in this paper is significantly stronger than the bound provided in \[ \text{Dani et al.} \ [2008] \], which is also \( \Omega(n/\sqrt{T}) \). In this latter lower bound, the decision problem the algorithm faces is chosen as a function of the time \( T \). In particular, the construction in \[ \text{Dani et al.} \ [2008] \] used a decision region which was a hypercube

\( D \) to be exponential in \( n \). Second, \[ \text{Auer} \ [2002] \] assumes that \( D \) is a subset of the sphere, which leads to an additional \( \sqrt{n} \) factor. To see this, note the comments in the beginning of Section 5 essentially show that a general decision region can be thought of as living in a hypercube (due to the barycentric spanner property), so the additional \( \sqrt{n} \) factor comes from rescaling the cube into a sphere.
though fixed.) Under these assumptions, the noise sequence, $x_t$ was compact was actually unnecessary, in light of our further assumptions. Note, however, that our assumptions are also met under the addition of any time-dependent unbiased random noise function.

In this paper we address the bandit version of the geometric optimization problem, where the decision maker’s feedback on each round is only the actual cost $c_i(x_t)$ received on that round, not the entire cost function $c_i(\cdot)$.

If $x_1, \ldots, x_T$ are the decisions made in the game, then define the cumulative regret by

$$R_T = \sum_{t=1}^{T} (\mu^\dagger x_t - \mu^\dagger x^*)$$

where $x^* \in D$ is an optimal decision for $\mu$, i.e.,

$$x^* \in \arg\min_{x \in D} \mu^\dagger x$$

which exists since $D$ is compact. Observe that if the mean $\mu$ were known, then the optimal strategy would be to play $x^*$ every round. Since the expected loss for each decision $x$ equals $\mu^\dagger x$, the cumulative regret is just the difference between the expected loss of the optimal algorithm and the expected loss for the actual decisions $x_t$. Since the sequence of decisions $x_1, \ldots, x_T$ may depend on the particular sequence of random noise encountered, $R_T$ is a random variable. Our goal in designing an algorithm is to keep $R_T$ as small as possible.

It is also important for us to make use of a barycentric spanner for $D$ as defined in Awerbuch and Kleinberg [2004]. A barycentric spanner for $D$ is a set of vectors $b_1, \ldots, b_n$, all contained in $D$, such that every vector in $D$ can be expressed as a linear combination of the spanner with coefficients in $[-1, 1]$. Awerbuch and Kleinberg [2004] showed that such a set exists for compact sets $D$. We assume we have access to such a spanner of the decision region, though an approximate spanner would suffice for our purposes (Awerbuch and Kleinberg [2004] provide an efficient algorithm for computing an approximate spanner).

Let $A$ be a positive definite $n \times n$ matrix, and let $\nu \in \mathbb{R}^n$. We will use the following notation for the $1$- and $2$-norms based on $A$.

$$\|\nu\|_{2, A} := \|A^{1/2}\nu\|_2 = \sqrt{\nu^\dagger A \nu}.$$  

$$\|\nu\|_{1, A} := \|A^{1/2}\nu\|_1 = \sum_{i=1}^{n} |A^{1/2}\nu|_i.$$  

Here $A^{1/2}$ is the unique positive definite $n \times n$ matrix whose square is $A$.

### 3 Main Results

#### 3.1 Algorithms

We now present our main algorithms, ConfidenceBall$_2$ and ConfidenceBall$_1$. The subscripts on the names refer to the type of norm used in the algorithm; apart from scaling the radius differently, which we do only for convenience, this is the sole difference between the algorithm statements. As we shall discuss later, we are able to prove better regret guarantees for ConfidenceBall$_2$, matching the lower bound, up to log factors.

Both algorithms can be efficiently implemented in the simplistic case when the decision set is a small finite set.
However, in the important special case when the decision set is a polytope presented as the intersection of halfspaces, ConfidenceBall1 can be implemented in polynomial time, while ConfidenceBall2 is NP-hard to implement, at least for some decision sets. More generally, ConfidenceBall1 can be implemented efficiently given oracle access to an algorithm which can find a decision in \( \arg\min_{x \in D} \nu \cdot x \) (where \( \nu \) is the input). We discuss these issues further in Subsection 3.2.

ConfidenceBall2

Algorithm 3.1 is due to Auer [2002], who called it the LinRel algorithm. We have generalized the statement slightly so that it can be applied in settings where \( D \) is not necessarily stored in enumerated form, and indeed, may not even be finite. We have renamed the algorithm ConfidenceBall2 to emphasize its key feature of maintaining an \( \ell_2 \) ball, \( B_T \), which contains \( \mu \) with high probability.

The algorithm is motivated as follows. Suppose decisions \( x_1, \ldots, x_{t-1} \) have been made, incurring corresponding losses \( \ell_1, \ldots, \ell_{t-1} \). Then a reasonable estimate \( \hat{\mu} \) to the true mean cost vector \( \mu \) can be constructed by minimizing the square loss:

\[
\hat{\mu} := \arg\min_{\nu} \mathcal{L}(\nu), \quad \text{where} \quad \mathcal{L}(\nu) := \sum_{\tau < t} (\nu^\dagger x_{\tau} - \ell_{\tau})^2.
\]

Defining \( A := \sum_{\tau < t} x_{\tau} x_{\tau}^\dagger \), we have that the least squares estimator is

\[
\hat{\mu} = A^{-1} \sum_{\tau < t} \ell_{\tau} x_{\tau}.
\]

A natural confidence region for \( \mu \) is the set of \( \nu \) for which \( \mathcal{L}(\nu) \) exceeds \( \mathcal{L}(\hat{\mu}) \) by at most some amount \( \beta \), i.e. the set

\[
\{ \nu : \mathcal{L}(\nu) - \mathcal{L}(\hat{\mu}) \leq \beta \}
\]

It is straightforward to see that:

\[
\mathcal{L}(\nu) - \mathcal{L}(\hat{\mu}) = (\nu - \hat{\mu})^t A (\nu - \hat{\mu})
\]

Thus the confidence region proposed above has the shape of an ellipsoid centered on \( \hat{\mu} \), where the axes are defined through \( A \). This set is commonly referred to as the set of vectors \( \nu \) with bounded Mahalanobis distance with respect to mean \( \hat{\mu} \) and covariance matrix \( A^{-1} \).

A difficulty with the above reasoning is that we have implicitly assumed that \( A \) is invertible, which is clearly false for \( t < n \). Under a slight alteration, define the estimator \( \hat{\mu}_t \) at time \( t \) by

\[
\hat{\mu}_t = A_t^{-1} \sum_{\tau < t} \ell_{\tau} x_{\tau},
\]

where \( A_t \) is now defined as

\[
A_t = \sum_{i=1}^n b_i b_i^\dagger + \sum_{\tau < t} x_{\tau} x_{\tau}^\dagger.
\]

where \( b_1, \ldots, b_n \) is the barycentric spanner (see Preliminaries for the definition). It is easily seen that \( A_t \) is positive definite (and hence invertible), since the spanner is linearly independent. Intuitively, the first term in \( A_t \) (the sum of outerproducts of the spanner vectors) is a natural initialization of the confidence region, as it imposes uncertainty along the directions in which \( D \) varies most (namely the spanner directions). Our proofs effectively show that an approximate spanner would suffice instead. Note that \( \hat{\mu}_t \) is the least squares estimator for the sampled data if we pretend that decisions \( b_1, \ldots, b_n \) were selected on fictitious rounds \( t = -n + 1, \ldots, t = 0 \) and all incurred loss 0.

Now define the confidence region at time \( t \) to be the ellipsoid

\[
B_t^2 := \{ \nu : \|\nu - \hat{\mu}_t\|_{2,A_t} \leq \sqrt{\beta} \}
\]

In the proofs, we show that, with our choice of \( \beta_t \), \( \mu \) always remains inside this ellipsoid for all times \( t \), with high probability.

The decision at the next round is then the greedy optimistic decision:

\[
x_t = \arg\min_{x \in D} \sum_{\nu \in B_t^2} \nu^t x.
\]

Again, this exists since \( D \) is compact.

It should be remarked that although the linear function \( x \mapsto \mu \cdot x \) is a feasible cost function, and \( \mu_t \) is an approximation to \( \mu \), the function \( x \mapsto \mu_t \cdot x \) may be far from being a
feasible (i.e. $[-1, 1]$-valued) cost function on $D$ — however, it is bounded in $[-n, n]$.

**ConfidenceBall**

**Problem Dependent Upper Bound**

Recall that $\delta < \frac{B_1}{\Delta}$.

Theorem 1

$\Pr(\forall T, R_T \leq \frac{8n^2 \beta_T \ln(T)}{\Delta}) \geq 1 - \delta$.

- **ConfidenceBall_1**: If ConfidenceBall_1 is implemented so that it only chooses extremal points $x_i \in D$ (which is always possible) then, for all sufficiently large $T$, the cumulative regret $R_T$ of ConfidenceBall_1($D, \delta$) is with high probability at most $O(\frac{n^2}{\Delta} \log^3 T)$. More precisely,

$$\Pr(\forall T, R_T \leq \frac{8n \beta_T \ln(T)}{\Delta}) \geq 1 - \delta,$$

- **ConfidenceBall_2**: For all sufficiently large $T$, the cumulative regret $R_T$ of ConfidenceBall_2($D, \delta$) is with high probability at most $O(\frac{n^2}{\Delta} \log^3 T)$. More precisely,

$$\Pr(\forall T, R_T \leq \frac{8n^2 \beta_T \ln(T)}{\Delta}) \geq 1 - \delta.$$
is not appropriate here. See Introduction for further discussion.

The decision region is constructed as follows. Assume $n$ is even. Let $D_n = (S^1)^{n/2}$ be the Cartesian product of $n/2$ circles. That is, $D_n = \{ (x_1, \ldots, x_n) : x_1^2 + x_2^2 = x_3^2 + x_4^2 = \cdots = x_{n-1}^2 + x_n^2 = 1 \}$.

Observe that $D_n$ is a subset of the intersection of the cube $[-1, 1]^n$ with the sphere of radius $\sqrt{n}/2$ centered at the origin.

Our cost functions take values in $\{-1, +1\}$, and for every $x \in D_n$, the expected cost is $\mu \cdot x$, where $\mu \in D_n$.

Since each cost function is only evaluated at one point, any two distributions over $\{-1, +1\}$-valued cost functions with the same value of $\mu$ are equivalent for the purposes of our model.

4 Concentration of Martingales

In our analysis, we use the following Bernstein-type concentration inequality for martingale differences, due to Friedman (1975) (see also McDiarmid (1998, Theorem 3.13)).

Theorem 4 (Friedman) Suppose $X_1, \ldots, X_T$ is a martingale difference sequence, and $b$ is an uniform upper bound on the steps $X_t$. Let $V$ denote the sum of conditional variances,

$$ V = \sum_{i=1}^{n} \Var(X_i | X_1, \ldots, X_{i-1}). $$

Then, for every $a, \nu > 0$,

$$ \Prob\left( \sum_{i=1}^{n} X_i \geq a \text{ and } V \leq v \right) \leq \exp\left( \frac{-a^2}{2v + 2ab/3} \right). $$

5 Upper Bound Analysis

Throughout the proof, without loss of generality, assume that the barycentric spanner is the standard basis $\vec{e}_1, \ldots, \vec{e}_n$ (this just amounts to a choice of a coordinate system, where we identify the spanner with the standard basis). Hence, the decision set $D$ is a subset of the cube $[-1, 1]^n$. In particular, this implies $||x|| \leq \sqrt{n}$ for all $x \in D$. This is really only a notational convenience; the problem is stated in terms of decisions in an abstract vector space, and expected costs in its dual, with no implicit standard basis.

In establishing the upper bounds there are two main theorems from which the upper bounds follow. The first is in showing that the confidence region is appropriate. Let $E$ be the event that for every time $t \leq T$, the true mean $\mu$ lies in the confidence region, $B^2_t$ or $B^1_t$. The following shows that event $E$ occurs with high probability. More precisely,

Theorem 5 (Confidence) Let $\delta > 0$.

- For ConfidenceBall$_2$,
  $$ \Prob(\forall t, \mu \in B^2_t) \geq 1 - \delta. $$

- For ConfidenceBall$_1$,
  $$ \Prob(\forall t, \mu \in B^1_t) \geq 1 - \delta. $$

Section 5.2 is devoted to establishing this confidence bound. In essence, the proof seeks to understand the growth of the quantity $\langle \mu_t - \mu \rangle \langle A_t \rangle$, which involves a rather technical construction of a martingale (using the matrix inversion
Let $\sum_{t=1}^{T} r_t^2 \leq 8nT\beta_T \ln T$. Applying the Cauchy-Schwarz inequality, we have, with probability at least $1 - \delta$

$$R_T = \sum_{t=1}^{T} r_t$$

$$\leq \left( \frac{T \sum_{t=1}^{T} r_t^2}{T} \right)^{1/2}$$

$$\leq \sqrt{8nT\beta_T \ln T}$$

which completes the proof.

We now provide the proofs of these two theorems.

5.1 Proof of Theorem 6

In this section, we prove Theorem 6 which says that the sum of the squares of the instantaneous regrets of the algorithm is small, assuming the evolving confidence balls always contain the true mean $\mu$. A key insight is that on any round $t$ in which $\mu \in B_t^2$, the instantaneous regret is at most the “width” of the ellipsoid in the direction of the chosen decision. Moreover, the algorithm’s choice of decisions forces the ellipsoids to shrink at a rate that ensures that the sum of the squares of the widths is small. We now formalize this.

Lemma 7 Let $x \in D$. Then

- For ConfidenceBall2, if $x \in B_t^2$ and $x \in D$. Then
  $$\|\nu - \hat{\mu}_t\|^2 \leq \sqrt{\beta_t x^T A_t^{-1} x}$$

- For ConfidenceBall1, if $\nu \in B_t^1$ and $x \in D$. Then
  $$\|\nu - \hat{\mu}_t\|^2 \leq \sqrt{\beta_t x^T A_t^{-1} x}$$

Proof: Unless explicitly stated, all norms refer to the $\ell_2$ norm. For ConfidenceBall2,

$$\|\nu - \hat{\mu}_t\|^2 = \|\nu - \hat{\mu}_t\| A_t^{-1/2} A_t^{-1/2} x$$

$$\leq \|A_t^{-1/2} (\nu - \hat{\mu}_t)\| \|A_t^{-1/2} x\|$$

.... by Cauchy-Schwarz

$$= \|A_t^{-1/2} (\nu - \hat{\mu}_t)\| \sqrt{x^T A_t^{-1} x}$$

$$\leq \sqrt{\beta_t x^T A_t^{-1} x}$$

where the last inequality holds since $\nu \in B_t^2$.

For ConfidenceBall1,

$$\|\nu - \hat{\mu}_t\|^2 \leq \|A_t^{1/2} (\nu - \hat{\mu}_t)\|_1 \|A_t^{-1/2} x\|_\infty$$

.... by Holder’s Inequality

$$\leq \|A_t^{1/2} (\nu - \hat{\mu}_t)\|_1 \|A_t^{-1/2} x\|_2$$

$$\leq \sqrt{n \beta_t x^T A_t^{-1} x}$$

where the last inequality holds since $\nu \in B_t^1$. ■

Define

$$w_t := \sqrt{x_t^T A_t^{-1} x_t}$$

which we interpret as the “normalized width” at time $t$ in the direction of the chosen decision. The true width, $2\sqrt{\beta_t w_t}$, turns out to be an upper bound for the instantaneous regret.
Lemma 8 Fix \( t \).

- For ConfidenceBall\(_2\), if \( \mu \in B^2_t \), then
  \[ r_t \leq 2 \min(\sqrt{\beta_t}w_t, 1) \]
- For ConfidenceBall\(_1\), if \( \mu \in B^1_t \), then
  \[ r_t \leq 2 \min(\sqrt{n\beta_t}w_t, 1) \]

Proof: Let \( \tilde{\mu} \in B^2_t \) denote the vector which minimizes the dot product \( \tilde{\mu}^\top x_t \). By choice of \( x_t \), we have
  \[ \tilde{\mu}^\top x_t = \min_{\mu \in B^2_t} \mu^\top x \leq \mu^\top x, \]
  where the inequality used the hypothesis \( \mu \in B^2_t \). Hence,
  \[ r_t = \mu^\top x_t - \tilde{\mu}^\top x_t \]
  \[ \leq (\mu - \tilde{\mu})^\top x_t \]
  \[ = (\mu - \tilde{\mu})^\top x_t + (\tilde{\mu} - \mu)^\top x_t \]
  \[ \leq 2\sqrt{\beta_t}w_t \]
  where the last step follows from Lemma 7 since \( \mu \) and \( \tilde{\mu} \) are in \( B^2_t \). Since \( \ell_t \in [-1, 1] \), \( r_t \) is always at most 2 and the result follows. The proof for ConfidenceBall\(_1\) is analogous. \( \blacksquare \)

Next we show that the sum of the squares of the widths does not grow too fast.

Lemma 9 We have for all \( t \)
  \[ \sum_{\tau=1}^{t} \min(w_{\tau}^2, 1) \leq 2n \ln t. \]

The following two facts prove useful to this end.

Lemma 10 For every \( t \leq T \),
  \[ \det A_{t+1} = \prod_{\tau=1}^{t}(1 + w_{\tau}^2). \]

Proof: By the definition of \( A_{t+1} \), we have

\[
\begin{align*}
\det A_{t+1} &= \det(A_t + x_t x_t^\top) \\
&= \det(A_t^{1/2}(I + A_t^{-1/2}x_t x_t^\top A_t^{-1/2})A_t^{1/2}) \\
&= \det(A_t) \det(I + A_t^{-1/2}x_t A_t^{-1/2}x_t^\top) \\
&= \det(A_t) \det(I + v_tv_t^\top),
\end{align*}
\]

where \( v_t := A_t^{-1/2}x_t \). Now observe that \( v_t^\top v_t = w_t^2 \) and

\[ (I + v_tv_t^\top)v_t = v_t + v_t(v_t^\top v_t) = (1 + w_t^2)v_t. \]

Hence \( (1 + w_t^2) \) is an eigenvalue of \( I + v_tv_t^\top \). Since \( v_tv_t^\top \) is a rank one matrix, all the other eigenvalues of \( I + v_tv_t^\top \) equal 1. It follows that \( \det(I + v_tv_t^\top) = (1 + w_t^2) \), and so

\[ \det A_{t+1} = (1 + w_t^2) \det A_t. \]

Recalling that \( A_1 \) is the identity matrix, the result follows by induction. \( \blacksquare \)

Lemma 11 For all \( t \), \( \det A_t \leq t^n \).

Proof: The rank one matrix \( x_t x_t^\top \) has \( x_t^\top x_t = \|x_t\|^2 \) as its unique non-zero eigenvalue. Also, since we have identified the spanner with the standard basis, we have \( \sum_{i=1}^{n} b_i^2 = 1 \).

Since the trace is a linear operator, it follows that

\[
\begin{align*}
\text{trace } A_t &= \text{trace } \left( I + \sum_{\tau \leq t} x_{\tau} x_{\tau}^\top \right) \\
&= n + \sum_{\tau < t} \text{trace}(x_{\tau} x_{\tau}^\top) \\
&= n + \sum_{\tau < t} \|x_{\tau}\|^2 \\
&\leq nt.
\end{align*}
\]

Now, recall that \( \text{trace } A_t \) equals the sum of the eigenvalues of \( A_t \). On the other hand, \( \det(A_t) \) equals the product of the eigenvalues. Since \( A_t \) is positive definite, its eigenvalues are all positive. Subject to these constraints, \( \det(A_t) \) is maximized when all the eigenvalues are equal; the desired bound follows. \( \blacksquare \)

Proof: [Proof of Lemma 9]

Using the fact that for \( 0 \leq y \leq 1 \), \( \ln(1 + y) \geq y/2 \), we have

\[
\begin{align*}
\sum_{\tau=1}^{t} \min(w_{\tau}^2, 1) &\leq \sum_{\tau=1}^{t} 2 \ln(1 + w_{\tau}^2) \\
&= 2 \ln(\det A_{t+1}) \\
&\leq 2n \ln t
\end{align*}
\]

by Lemmas 10 and 11. \( \blacksquare \)

Finally, we are ready to prove that if \( \mu \) always stays within the evolving confidence region, then our regret is under control.

Proof: [Proof of Theorem 6] Assume that \( \mu \in B^2_t \) for all \( t \).

Then

\[
\begin{align*}
\sum_{t=1}^{T} r_t^2 &\leq \sum_{t=1}^{T} 4\beta_t \min(w_{t}^2, 1) & \text{by Lemma 8} \\
&\leq 4\beta_T \sum_{t=1}^{T} \min(w_{t}^2, 1) & \text{since } 1 < \beta_1 < \cdots < \beta_T \\
&\leq 8\beta_T n \ln T & \text{by Lemma 9}.
\end{align*}
\]

The proof for ConfidenceBall\(_1\) is analogous. \( \blacksquare \)

5.2 Proof of Theorem 5

In this section, we prove Theorem 5, which states that with high probability, for all \( t \), the true mean \( \mu \) lies in the confidence ball \( B_t \).

Recall that

\[ \eta_t := c_t(x_t) - \mu^\top x_t = \ell_t - \mathbb{E}(\ell_t \mid \mathcal{H}_t) \]

where \( \mathcal{H}_t \) denotes the complete history of the game on rounds \( 1, \ldots, t-1 \), that is, the \( \sigma \)-algebra generated by \( \ell_1, \ldots, \ell_{t-1} \).

For either algorithm, we will analyze the quantity:

\[ Z_t := (\tilde{\mu}_t - \mu)^\top A_t(\tilde{\mu}_t - \mu) \]
which measures the error of $\hat{\mu}_t$ as an approximation to the true mean, $\mu$, under the norm induced by $A_t$.

We will show that, with probability greater than $1 - \delta$, $Z_t \leq \beta_t$ for all $t$ for either algorithm. For ConfidenceBall$_2$, this directly implies that $\mu \in B^2_t$. For ConfidenceBall$_1$, note that

$$||A_t^{1/2}(\hat{\mu}_t - \mu)||_1 \leq \sqrt{\eta}||A_t^{1/2}(\hat{\mu}_t - \mu)||_2 = \sqrt{n}Z_t$$

so if $Z_t \leq \beta_t$ then $\mu \in B^1_t$.

The next lemma bounds the growth of $Z_t$.

**Lemma 12** For all $t$,

$$Z_t \leq n + 2\sum_{\tau=1}^{t-1} \eta_\tau x_\tau^\dagger(\hat{\mu}_\tau - \mu) + \sum_{\tau=1}^{t-1} \eta_\tau^2 \frac{w_\tau^2}{1 + w_\tau^2}.$$

**Proof:** For notational convenience, define:

$$Y_t = A_t^1(\hat{\mu}_t - \mu)$$

We have the following relations:

$$Z_t = Y_t^\dagger A_t^{-1} Y_t$$

$$Y_t = \sum_{\tau < t} \eta_\tau x_\tau - \mu$$

$$Y_{t+1} = Y_t + \eta_t x_t$$

which are immediate from the definitions of $A_t$, $\hat{\mu}_t$, and $\eta_t$.

Now examining the growth of $Z_t$, we have:

$$Z_{t+1} = Y_{t+1}^\dagger A_{t+1}^{-1} Y_{t+1}$$

$$= (Y_t + \eta_t x_t)^\dagger A_t^{-1} (Y_t + \eta_t x_t)$$

$$= Y_t^\dagger A_t^{-1} Y_t + 2\eta_t x_t^\dagger A_t^{-1} Y_t + \eta_t^2 x_t^\dagger A_{t-1} x_t$$

(1)

Applying the matrix inversion lemma to $A_{t+1}^{-1}$, we note that:

$$A_{t+1}^{-1} = (A_t + x_t x_t^\dagger)^{-1}$$

$$= A_t^{-1} - A_t^{-1} x_t x_t^\dagger A_t^{-1}$$

$$= A_t^{-1} - \frac{A_t^{-1} x_t x_t^\dagger A_t^{-1}}{1 + x_t^\dagger A_t^{-1} x_t}$$

We can use this to bound the three terms of (1) as follows.

For the first term,

$$Y_t^\dagger A_t^{-1} Y_t \leq Z_t.$$

For the second term,

$$2\eta_t x_t^\dagger A_{t+1}^{-1} Y_t = 2\eta_t x_t^\dagger A_{t+1}^{-1} Y_t - 2\eta_t x_t^\dagger A_t^{-1} x_t x_t^\dagger A_t^{-1} Y_t$$

$$= 2\eta_t x_t^\dagger A_t^{-1} Y_t - 2\eta_t w_t^2 x_t^\dagger A_t^{-1} Y_t$$

$$= 2\eta_t x_t^\dagger A_t^{-1} Y_t$$

$$= 2\eta_t x_t^\dagger (\hat{\mu}_t - \mu)$$

For the third term,

$$\eta_t^2 x_t^\dagger A_{t+1}^{-1} x_t = \eta_t^2 w_t^2 - \eta_t^2 \frac{w_t^4}{1 + w_t^2}$$

$$= \eta_t^2 \frac{w_t^2}{1 + w_t^2}$$

Putting these together, we have shown

$$Z_{t+1} \leq Z_t + 2\eta_t x_t^\dagger (\hat{\mu}_t - \mu) + \eta_t^2 \frac{w_t^2}{1 + w_t^2}.$$

By induction, it follows that

$$Z_t \leq Z_1 + 2\sum_{\tau=1}^{t-1} \eta_\tau x_\tau^\dagger (\hat{\mu}_\tau - \mu) + \sum_{\tau=1}^{t-1} \eta_\tau^2 \frac{w_\tau^2}{1 + w_\tau^2}.$$

Finally, we check that $Z_1 \leq n$. To see this, recall from the algorithm that $A_1 = I$ and $\hat{\mu}_1 = 0$. Also, since $e_1, \ldots, e_n \in D$, by assumption, $\mu \cdot c_j \in [-1, 1]$.

$$Z_1 = (\hat{\mu}_1 - \mu)^\dagger A_1 (\hat{\mu}_1 - \mu)$$

$$= ||\mu||^2$$

$$= \sum_{j=1}^n (\mu^\dagger e_j)^2$$

$$\leq n.$$

This completes the proof. ■

We now define a useful martingale difference sequence. First, it is convenient to define an “escape event” $E_t$ as:

$$E_t = \{Z_t \leq \beta_t \text{ for all } \tau \leq t\} = \{\mu \in B_\tau \text{ for all } \tau \leq t\}$$

where $\{\} = \{\}$ is the indicator function.

**Lemma 13** Define a random variable $M_t$ by

$$M_t = 2\eta_t E_t^\dagger x_t^\dagger (\hat{\mu}_t - \mu) \frac{1}{1 + w_t^2}.$$

Then $M_t$ is a martingale difference sequence with respect to the sequence of game histories $\mathcal{H}_t$.

**Proof:** To see that $M_t$ is a martingale difference sequence, note that:

$$\mathbb{E}(M_t | \mathcal{H}_t) = 2\eta_t x_t^\dagger (\hat{\mu}_t - \mu) \frac{1}{1 + w_t^2} \mathbb{E}(\eta_t | \mathcal{H}_t)$$

$$= 0$$

since the history $\mathcal{H}_t$ fully determines $x_1, \ldots, x_t, \hat{\mu}_1, \ldots, \hat{\mu}_t, Z_1, \ldots, Z_t,$ and $E_1, \ldots, E_t$, and since the noise functions $\eta_t$ are a martingale difference sequence with respect to $\mathcal{H}_t$. ■

We show that with high probability, the associated martingale, $\sum_{\tau=1}^t M_\tau$, never grows too large.

**Lemma 14** Given $\delta < 1$,

$$\mathbb{P}(\forall t, \sum_{\tau=1}^{t-1} M_\tau \leq \beta_t/2) \geq 1 - \delta,$$
We defer the proof to Section 5.2.1. Equipped with this lemma, we can prove Theorem 5.

**Proof:**[Proof of Theorem 5] It suffices to show that the high-probability event described in Lemma 14 is contained in the support of $E_t$ for every $t$. We prove the latter by induction on $t$.

By Lemma 12 and the definition of $\beta_t$, we know that $Z_t \leq n < \beta_t$. Hence $E_t$ is always 1 (equivalently, $\mu$ is always in $B_t$).

Now suppose the high-probability event of Lemma 14 holds, so in particular,

$$\sum_{\tau=1}^{t-1} M_\tau \leq \beta_t/2.$$  

By inductive hypothesis, $E_\tau = 1$ for $\tau \leq t - 1$. Hence by Lemma 12 we have

$$Z_t \leq n + 2\sum_{\tau=1}^{t-1} \eta_\tau \frac{x_{\tau}^4(\hat{\mu}_\tau - \mu)}{1 + w_\tau^2} + \sum_{\tau=1}^{t-1} \eta_\tau^2 \frac{w_\tau^2}{1 + w_\tau^2}$$

$$= n + \sum_{\tau=1}^{t-1} M_\tau + \sum_{\tau=1}^{t-1} \eta_\tau^2 \frac{w_\tau^2}{1 + w_\tau^2}$$

$$\leq n + \beta_t/2 + \sum_{\tau=1}^{t-1} \eta_\tau^2 \frac{w_\tau^2}{1 + w_\tau^2}$$

$$\leq n + \beta_t/2 + \sum_{\tau=1}^{t-1} \min(w_\tau^2, 1) \quad \text{since } |\eta_\tau| \leq 1$$

$$\leq n + \beta_t/2 + 2n \ln t \quad \text{by Lemma 9}$$

Thus we have shown $E_t = 1$, completing the induction. ■

### 5.2.1 Concentration

All that remains to complete the proof now is to show that our martingale $\sum_{t} M_t$ has good concentration properties. As we show, the step sizes $|M_t|$ are uniformly bounded so that an application of the Hoeffding-Azuma inequality would bound the probability that $\sum_{t} M_t$ grows too large. Unfortunately, the bound thus obtained translates into a regret bound of $T^{3/4}$, which is not good enough for our purpose.

Instead we use Theorem 3 which allows us to to bound the step sizes in terms of random variables, as long as the conditional variances remain under control.

**Proof:**[Proof of Lemma 14] Let us first obtain upper bounds on the step sizes of our martingale.

$$|M_t| = 2|\eta_t| E_t \frac{|x_{\tau}^4(\hat{\mu}_t - \mu)|}{1 + w_t^2}$$

$$\leq 2|\eta_t| E_t \frac{\sqrt{\beta_t x_{\tau}^4 A_{t-1}^{-1} x_t}}{1 + w_t^2}$$

$$= 2|\eta_t| E_t \frac{w_t \sqrt{\beta_t}}{1 + w_t^2}$$

$$\leq 2|\eta_t| E_t \sqrt{\beta_t} \min(w_t, 1/2) \quad (2)$$

where the first inequality follows trivially when $E_t = 0$, and by Lemma 12 when $E_t = 1$. Additionally this gives a family of uniform upper bounds:

$$|M_t| \leq \sqrt{\beta_t} \quad \text{for all } \tau \leq t$$

since $|\eta_\tau| \leq 1$ and (by choice) $\beta_\tau$ is a non-decreasing sequence.

Next we bound the sum of the conditional variances of our martingale. Note that $(\min(w_t, 1/2))^2 = \min(w_t^2, 1/4)$

$$V_t := \sum_{\tau=1}^{t} \text{Var}(M_t | M_1 \ldots M_{\tau-1})$$

$$\leq \sum_{\tau=1}^{t} 4|\eta_\tau|^2 E_t \beta_\tau \min(w_\tau^2, 1/4) \quad \text{by (2)}$$

$$\leq 4(\max_{\tau \leq t} \beta_\tau) \sum_{\tau=1}^{t} E_t \min(w_\tau^2, 1) \quad \text{since } |\eta_\tau| \leq 1$$

$$\leq 4\beta_t \sum_{\tau=1}^{t} E_t \min(w_\tau^2, 1)$$

$$\leq 8\beta_t \ln (\max_{\tau \leq t} E_\tau = 1) \quad \text{by Lemma 9}$$

$$\leq 8\beta_t \ln t$$

Since we have established that the sum of conditional variances, $V_t$, is always bounded by $8\beta_t \ln t$, we can apply Theorem 4 with parameters $a = \beta_t/2$, $b = \sqrt{\beta_t}$ and $\nu = 8n\beta_t \ln t$, to get

$$\text{Prob} \left( \sum_{\tau=1}^{t-1} M_\tau \geq \beta_t/2 \right)$$

$$= \text{Prob} \left( \sum_{\tau=1}^{t-1} M_\tau \geq \beta_t/2 \text{ and } V_t \leq 8n\beta_t \ln t \right)$$

$$\leq \exp \left( \frac{-\beta_t}{2(8n\beta_t \ln t) + \frac{8}{3}(\beta_t/2)(\sqrt{\beta_t})} \right)$$

$$\leq \exp \left( \frac{-\beta_t}{64n \ln t + \frac{8}{3} \sqrt{\beta_t}} \right)$$

$$\leq \max \left( \exp \left( -\frac{-\beta_t}{128n \ln t} \right), \exp \left( -\frac{3\sqrt{\beta_t}}{8} \right) \right)$$

$$\leq \frac{\delta}{\tau^2}$$

where the last inequality follows from the definition of $\beta_t$.

Finally, we apply a union bound to get

$$\text{Prob} \left( \sum_{\tau=1}^{t} M_\tau \geq \beta_t/2 \text{ for some } t \right)$$

$$\leq \sum_{t=1}^{T} \text{Prob} \left( \sum_{\tau=1}^{t-1} M_\tau \geq \beta_t/2 \right)$$

$$\leq \sum_{t=2}^{T} \frac{\delta}{\tau^2}$$

$$\leq \delta \left( \frac{\pi^2}{6} - 1 \right)$$

$$\leq \delta$$

completing the proof of Lemma 14. ■
6 Lower Bound Analysis

Recall \( n \) is even and that \( D_n = (S^1)^{n/2} \) is the Cartesian product of \( n/2 \) circles, i.e., \( D_n = \{(x_1, \ldots, x_n) : x_1^2 + x_2^2 = \ldots = x_{n-1}^2 + x_n^2 = 1\} \). Under the cost vector \( \mu \in D_n/n \), one can see that the probability of observing a particular loss \( \ell \) in the next round equals

\[
\Pr(\ell_t = \ell \mid x_t) = \frac{1 + \ell \mu \cdot x_t}{2},
\]

since \( \ell_t \) has support \([-1, 1]\) and mean \( \mu \cdot x_t \). Note the probability that \( \ell_t = 1 \) is bounded between \( 1/4 \) and \( 3/4 \). Also, the optimal decision is \( x^* = -\nu_1 \in D_n \), which achieves an expected loss of \( \mu \cdot x^* = -1/2 \).

Fix an index \( 1 \leq i \leq n/2 \), and consider \( R_T^{(i)} \), the contribution to the total regret from the choice of \( x^{(i)} = (x_{2i-1}, x_{2i}) \), \( i \), the component from the \( i \)-th circle. More precisely,

\[
R_T^{(i)} = \sum_{t=1}^T \left( \mu^{(i)} \cdot x_t^{(i)} + \frac{1}{n} \right) = \sum_{t=1}^T r_T^{(i)},
\]

where \( \mu^{(i)} = (\mu_{2i-1}, \mu_{2i}) \) and \( r_T^{(i)} \) is the instantaneous regret. Clearly, we have \( R_T = \sum_{i=1}^{n/2} R_T^{(i)} \).

Thus, the proof of Theorem \( \text{B} \) reduces to proving a lower bound \( \mathbb{E}(R_T^{(i)}) = \Omega(\sqrt{T}) \). Let us condition on the event that the \( i \)-th component of \( \mu \), namely \( (\mu_{2i-1}, \mu_{2i}) \), is one of two vectors, \( \nu_1, \nu_2 \in S^1/n \), such that \( \|\nu_1 - \nu_2\| = \varepsilon \) (where \( \varepsilon \leq 2/n \)). We further condition on the exact values of the other \( n/2 - 1 \) components of \( \mu \), thus leaving only two possibilities for \( \mu \), which we call \( \mu_1 \) and \( \mu_2 \). We denote \( \varepsilon = \|\nu_1 - \nu_2\| = \|\mu_1 - \mu_2\| \). Note that \( \mu \) is uniform over \( \{\mu_1, \mu_2\} \) in this event. Ultimately, we will choose \( \varepsilon \) as a function \( T \) and \( n \).

To make this conditioning precise, consider the following two-stage sampling procedure for \( \mu \), where we first sample \( \mu_1 \) and \( \mu_2 \) and then sample \( \nu_1 \) uniformly at random over \( \{\mu_1, \mu_2\} \) — we do this sampling such that the marginal distribution over \( \nu \) is still uniform over \( D_n/n \). More precisely, first sample \( \nu_1 \) and \( \nu_2 \) such that \( \|\nu_1 - \nu_2\| = \varepsilon \) and such that the marginal distributions \( \text{Prob}(\nu_1) \) and \( \text{Prob}(\nu_2) \) are both uniform over \( S^1/n \). Independently, sample the remaining components \( \mu^{(i)} \), where \( j \neq i \), uniformly at random over \( S^1/n \). This provides us with \( \mu_1 \) and \( \mu_2 \) (which only differ in the \( i \)-th component). Finally, we sample \( \nu_1 \) uniformly at random over \( \{\mu_1, \mu_2\} \). By the tower property of conditional expectations (Fubini’s theorem),

\[
\mathbb{E}(r_T^{(i)}) = \mathbb{E}_{\mu} \mathbb{E}(r_T^{(i)} | \mathcal{H}_t, \{\mu_1, \mu_2\})
\]

where the inner expectation is over \( \mu \) which lies in \( \{\mu_1, \mu_2\} \).

Our next Lemma, roughly speaking, proves that decisions \( x_t \) that provide more information about the value of \( \mu \), necessarily produce more instantaneous regret. Moreover, the result holds even if the algorithm were allowed to magically know the values of \( \mu_{1}, \mu_{2} \). This places the algorithm designer on the horns of a dilemma: give up on distinguishing the cases \( \mu = \mu_1 \) and \( \mu = \mu_2 \), or play high-regret decisions to try to acquire the information. As we shall see in the proof of Theorem \( \text{B} \) either way, we end up with \( \mathbb{E}(R_T^{(i)}) = \Omega(\sqrt{T}) \): in the former case, with certainty, and in the latter case, with high probability.

Let

\[
b_t := \Pr(\mu = \mu_1 \mid \mathcal{H}_t) - \Pr(\mu = \mu_2 \mid \mathcal{H}_t)
\]

be the bias towards \( \mu_1 \) at time \( t \). Note that \( b_0 = 0 \), and that the sequence \( (b_t) \) is a martingale with respect to \( \{\mathcal{H}_t\} \). Our Lemma gives a lower bound on regret in terms of the martingale differences \( b_{t+1} - b_t \). We have not attempted to optimize the constants.

**Lemma 15** For \( \varepsilon \) as defined above, for all \( t \geq 1 \), for any sequence of decisions \( x_1, \ldots, x_t \) and outcomes \( \ell_1, \ldots, \ell_{t-1}, \ell_t \), the regret from round \( t \) satisfies

\[
\mathbb{E}(r_t^{(i)} \mid \mathcal{H}_t, \{\mu_1, \mu_2\}) \geq \left( \frac{7n\varepsilon^2}{256} + \frac{|b_{t+1} - b_t|^2}{18n^2} \right) 1\{|b_t| \leq 1/2\}.
\]

**Proof:** Let \( p_t = \Pr(\mu = \mu_1 \mid \mathcal{H}_t) \). Then \( b_t = 2p_t - 1 \). By Bayes’ rule and equation \( \text{B} \) we have that

\[
p_{t+1} = \frac{\Pr(\ell_t \mid \mu = \mu_1)p_t}{\Pr(\ell_t \mid \mathcal{H}_t)} = \frac{p_t(1 + \ell_t(\mu_1 \cdot x_t))}{p_t(1 + \ell_t(\mu_1 \cdot x_t)) + (1 - p_t)(1 + \ell_t(\mu_2 \cdot x_t))} = \frac{p_t + \ell_t(\mu_1 \cdot x_t)}{1 + p_t\ell_t(\mu_1 \cdot x_t) + (1 - p_t)\ell_t(\mu_2 \cdot x_t)}
\]

Hence

\[
b_{t+1} - b_t = 2(p_{t+1} - p_t) = 2\frac{p_t(1 + \ell_t(\mu_1 \cdot x_t)) - p_t - p_t^2\ell_t(\mu_1 \cdot x_t) - p_t(1 - p_t)\ell_t(\mu_2 \cdot x_t)}{1 + p_t\ell_t(\mu_1 \cdot x_t) + (1 - p_t)\ell_t(\mu_2 \cdot x_t)} = \frac{2p_t(1 - p_t)\ell_t(\mu_1 - \mu_2) \cdot x_t}{1 + p_t\ell_t(\mu_1 \cdot x_t) + (1 - p_t)\ell_t(\mu_2 \cdot x_t)}.
\]

Since \( |\ell_t| \cdot |x_t| \leq \|\mu_1\| \cdot \|x_t\| = \sqrt{\frac{1}{T} \sqrt{T}} = 1/2 \), the denominator of the above expression is at least \( 1/2 \). Since \( p_t(1 - p_t) \leq 1/4 \), it follows that

\[
|b_{t+1} - b_t| \leq |\ell_t| \cdot |(\mu_1 - \mu_2) \cdot x_t| = |\ell_t| \cdot |(\nu_1 - \nu_2) \cdot x_t^{(i)}| = \varepsilon |\alpha|
\]

where \( \alpha \) may be defined as follows.

Let \( v_1 \) be the unit vector in the direction of \( \nu_1 - \nu_2 \), and let \( v_2 \) be the unit vector in the direction of \( \nu_1 + \nu_2 \). Note that \( v_1, v_2 \) is an orthonormal basis for the plane, since \( \|\nu_1\| = \|\nu_2\| \). Decompose \( x_t^{(i)} = \alpha v_1 + \beta v_2 \), and \( \mathbb{E}(\mu^{(i)} \mid \mathcal{H}_t, \{\mu_1, \mu_2\}) = \alpha v_1 + \delta v_2 \). By definition of \( b_t \),

\[
\mathbb{E}(\mu^{(i)} \mid \mathcal{H}_t, \{\mu_1, \mu_2\}) = \frac{\nu_1 + \nu_2}{2} + \frac{b_t(\nu_1 - \nu_2)}{2}
\]

so we have \( \gamma = \varepsilon b_t/2 \) and \( \delta = \frac{\sqrt{\varepsilon^2 - \gamma^2}}{2} \).

Assume the game history is such that \( |b_t| \leq 1/2 \). Otherwise, since the regret is non-negative, the claim trivially.
follows. Now we calculate
\[\mathbb{E}(\eta_t^{(i)} | \mathcal{H}_t, \{\mu_1, \mu_2\}) \]
\[= \frac{1}{n} + x_t^{(i)} \cdot \mathbb{E}(\mu_t^{(i)} | \mathcal{H}_t, \{\mu_1, \mu_2\}) \]
\[= \frac{1}{n} + \alpha \gamma + \beta \delta \]
\[= \frac{1}{n} + \alpha \varepsilon b_t/2 + \beta \sqrt{\frac{1}{n^2} - \frac{\varepsilon^2}{4}} \]
\[\geq \frac{1}{n} + \alpha \varepsilon b_t/2 + \left(\frac{\alpha^2}{2} - 1\right) \sqrt{\frac{1}{n^2} - \frac{\varepsilon^2}{4}} \quad (5) \]
\[\geq \frac{1}{n} + \alpha \varepsilon b_t/2 + \left(\frac{\alpha^2}{2} - 1\right) \left(\frac{1}{n} - \frac{n \varepsilon^2}{8}\right) \quad (6) \]
\[= \frac{\alpha^2 b_t}{2n} + \frac{n \varepsilon^2}{8} - \frac{\alpha^2 n \varepsilon^2}{16} \]
\[\geq -\frac{\alpha \varepsilon}{4} + \frac{\alpha^2}{2n} + \frac{n \varepsilon^2}{16} \quad (7) \]
\[= \frac{\alpha^2}{18n} + \frac{7 n \varepsilon^2}{256} + \left(\frac{2}{3\sqrt{n}} - \frac{3\sqrt{n} \varepsilon}{16}\right) \]
\[\geq \frac{\alpha^2}{18n} + \frac{7 n \varepsilon^2}{256} \geq \frac{|b_t|}{2} \quad (8) \]

Here \(5\) follows because \(\alpha^2 + \beta^2 = 1\) implies that \(1 + \beta = \alpha^2/(1 - \beta) \geq \alpha^2/2\), with equality iff \(\beta = -1\). Inequality \(6\) follows since \(\alpha^2 - 1 \leq 0\) and \(\sqrt{T - x} \leq 1 - x/2\). Inequality \(7\) follows because \(|b_t| \leq 1/2\) and \(|\alpha| \leq 1\). Inequality \(8\) follows from \(4\), completing the proof.

We are now ready to prove Theorem 3.

**Proof:** Choose \(\varepsilon = T^{-1/4} n^{-1/2}\). Note this is possible since the only constraint on \(\varepsilon\) is \(\varepsilon \leq \frac{2}{n}\), which is satisfied as long as \(T \geq n^2/16\), which is true by hypothesis. Now, observe that, by the tower property of conditional expectations (Fubini’s theorem) and linearity of expectation,
\[
\mathbb{E}(R_t^{(i)})
\]
\[= \sum_{i=1}^{T} \mathbb{E}(\eta_t^{(i)} | \mathcal{H}_t, \{\mu_1, \mu_2\}) \]
\[\geq \sum_{i=1}^{T} \left(\frac{7 n \varepsilon^2}{256} + \frac{|b_t|}{18 n \varepsilon^2} \right) \mathbf{1}\{|b_t| \leq 1/2\} \]

... by Lemma \(15\)
\[= \sum_{i=1}^{T} \left(\frac{7}{256 \sqrt{T}} + \frac{|b_t|}{18 \sqrt{T}} \right) \mathbf{1}\{|b_t| \leq 1/2\} \]
\[\geq \frac{7 \sqrt{T}}{256} \text{Prob}(\text{for all } t, |b_t| \leq 1/2) \]
\[+ \frac{\sqrt{T}}{18} \sum_{t=1}^{T} \mathbb{E}(\{b_t = 1/2\}) \]

Thus, if \(\text{Prob}(\text{for all } t, |b_t| \leq 1/2) \geq 1/2 - 1/e\), then we are done by the first term on the right-hand side. Otherwise, with probability at least \(1/2 + 1/e\), there exists \(t \leq T\) such that \(|b_t| \geq 1/2\). Define
\[V := \sum_{t=1}^{T} \mathbf{1}\{\forall \tau \leq t, |b_{\tau+1} - b_\tau| \leq 1/2\} \mathbb{E}(b_{t+1} - b_t | \mathcal{H}_t) \]

By Freedman’s Bernstein-type inequality for martingales (Theorem \(4\) applied to the martingale \(b_{\tau+1} - b_\tau\), where \(\sigma = \min\{\tau : |b_\tau| \geq 1/2\}\), which by \(4\) satisfies \(|b_{t+1} - b_t| \leq \varepsilon\), we have,
\[\text{Prob} \left( (\exists t \leq T) |b_t| \geq 1/2 \text{ and } V \leq \frac{\sqrt{T}}{16} \right) \]
\[\leq 2 \exp \left( -\frac{1/4}{1/16 + \varepsilon/3} \right) \leq \frac{1}{e^2} \]

since by choice of \(\varepsilon\) and hypothesis on \(T\) we have
\[\varepsilon \leq T^{-1/4} \leq \frac{1}{4} < \frac{3}{16} - \frac{\ln(2)}{16} \]

It follows that
\[\text{Prob} \left( V > \frac{1}{32} \right) \geq 1/2. \]

In particular,
\[\sum_{t=1}^{T} \mathbb{E}(|b_{t+1} - b_t|^2 \mathbf{1}\{|b_t| \leq 1/2\}) \geq \mathbb{E} V \geq \frac{1}{64}. \]

Combining the above two cases, we have established that
\[\mathbb{E}(R_t^{(i)}) \geq \sqrt{T} \min \left\{ \frac{7}{256} \left(\frac{1}{2} - \frac{1}{e}\right), \frac{7 \sqrt{T}}{18} \right\} = \frac{\sqrt{T}}{1152}. \]

Since \(R_T = \sum_{t=1}^{n/2} R_t^{(i)}\), it follows by linearity of expectation that
\[\mathbb{E}(R_T) = \sum_{t=1}^{n/2} \mathbb{E}(R_t^{(i)}) \geq \frac{n \sqrt{T}}{2 \cdot 1152} \]

which completes the proof.

### 7 Extension: time-varying decision sets

Our techniques also apply to the setting when only a subset of the full decision set \(D\) is available in each round. Suppose, at time \(t\) only a subset of decisions \(D_t \subset D\) are available. In this case, the correct notion of regret is to compare each chosen decision \(x_t\), not with the global optimum \(x^*\), but with the best choice from the available subset \(D_t\). Thus
\[R_T = \sum_{t=1}^{T} (\mu^*_t x_t - \mu^*_t x^*_t) \]
where \(x^*_t \in D_t\) is an optimal decision for \(\mu\), i.e.,
\[x^*_t \in \arg\min_{x \in D_t} \mu^x \]

The only change that needs to be made to our algorithm is that now \(x_t\) is chosen from \(D_t\) instead of \(D\).
With these changes in definitions, all of our numbered Theorems and Lemmas still hold, with $D$ replaced by $D_t$ and $x^*$ replaced by $x^*_t$ where they appear. (This is trivial in the case of the lower bounds.) The changes to the proofs are minimal.

We note that a very similar model was considered by \cite{AbeLong99}, who proved a lower bound of $\Omega(T^{3/4})$ in their setting. However, this does not contradict our results, because their lower bound requires the dimension $n$ to be a function of $T$.

References


