Stochastic Linear Optimization under Bandit Feedback

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

Abstract

In the classical stochastic k-armed bandit problem, in each of a sequence of rounds, a decision maker chooses one of k arms and incurs a cost chosen from an unknown distribution associated with that arm. In the linear optimization analog of this problem, rather than finitely many arms, the decision set is a compact subset of $\mathbb{R}^n$ and the cost of each decision is just the evaluation of a randomly chosen linear cost function at that point. As before, it is assumed that the cost functions are sampled independently from an unknown but time-invariant distribution. The goal is to minimize the total cost incurred over some number of rounds T, and success is measured by low regret. Auer [2003] was the first to study this problem and provided an algorithm with a regret bound that is $O(\mathrm{poly}(n, \log |D|)\sqrt{T})$, where $|D|$ is the cardinality of the decision set (which he assumed to be finite).

We present a near complete characterization of this problem in terms of both upper and lower bounds. We consider a deterministic algorithm based on upper confidence bounds, which was described by Auer and conjectured to have small regret. (Auer analyzed a more complicated master algorithm, which called this simpler algorithm as a subroutine.) In certain natural cases, such as when D is finite or when D is a polytope, we show that this algorithm achieves an expected regret of only $O(n^2 \log^3 T)$, which has no dependence on the size of D. This polylogarithmic dependence on the time is analogous to the K-arm bandit setting, where logarithmic rates are optimal. Here, however, the rates also depend on a problem dependent constant that is sometimes characterized as the "gap" in performance of the best two arms (which is always effectively nonzero). In our setting, this polylogarithmic rate also depends on a problem dependent constant, which we characterize by a gap in performance of "extremal points" of the decision region. Our polylogarithmic upper bound is only applicable when this gap is nonzero. For general decision regions, where this gap is zero (such as for spheres), we provide a different regret bound that is $O^*(n\sqrt{T})$, and also a nearly matching lower bound, showing that this rate is optimal in terms of both n and T, up to polylogarithmic factors. Importantly, this lower bound shows that a polylogarithmic rate as function of the time is *not* achievable for certain infinite decision regions (where the gap is 0), which is in stark contrast to the K-arm bandit setting where logarithmic rates are *always* achievable.

*Department of Computer Science, University of Chicago, `varsha@cs.uchicago.edu`
†Toyota Technological Institute at Chicago, {`hayest,sham`}`@tti-c.org`

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 [2003]. Here, we assume the decision space is an arbitrary subset $D \subset \mathbb{R}^n$ and that there is fixed distribution π over cost functions. At each round, the learner chooses a decision $x \in D$, then a cost function $f(\cdot) : D \to [0,1]$ is sampled from π. 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 π (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 π. Note that the optimal algorithm here simply chooses the best decision with respect to the linear mean vector μ.

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 μ 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 Results and Comparison to Previous Work

Auer [2003] provides the first analysis of this problem. This paper builds and improves upon the work of Auer [2003] in a number of important ways, which we now summarize.

First, while Auer [2003] provides a natural, deterministic algorithm, based on upper confidence bounds of μ, 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 significantly 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. This simpler algorithm is directly applicable to infinite decision regions (Auer [2003] only considered the finite case). Furthermore, this algorithm may be implemented efficiently for the case when D is convex and when given certain oracle optimization access to D. A key technical tool in our analysis of this simpler algorithm is a (not particularly well known) concentration result by Freedman (Theorem 6.1 in the Appendix), which can be viewed as a Bernstein-type bound for martingales.

Second, Auer [2003] achieves a regret of $O^*((\log|D|)^{3/2}\text{poly}(n)\sqrt{T})$ where n is dimension of the decision space, T is the time horizon, and $|D|$ is the number of feasible decisions. Note 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 earlier work by Auer et al. [2002],the optimal regret for the K-arm bandit case was characterized as $\frac{K}{\Delta}\log T$, where Δ is the "gap" between the performance of the best arm and the second best arm. Note that this result is stated in terms of the problem dependent constant Δ, 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. [Lai and Robbins., 1985, 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 the case where $|D|$ is finite (such as the case considered in Auer [2003]), 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 $|D|$). This work shows that a regret of $\frac{n^2}{\Delta}\,\text{polylog}(T)$ can be achieved, where Δ 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, Δ 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 Δ is the gap in value between the first and second best corner decisions. Also, for the case where D is finite (as in the case considered by Auer [2003]), Δ is exactly the same as in the K arm case. However, for some natural decision regions, such as a sphere, Δ is 0 so this bound is not applicable. Note that Δ 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.

Third, we provide a more general bound of $O^*(n\sqrt{T})$, which does not explicitly depend on Δ. Hence, this bound is applicable when $\Delta = 0$. It is also appropriate if we desire a bound that is not stated in terms of problem dependent constants. Using the result in Auer [2003], one can also derive a bound of the form $O(\text{poly}(n)\sqrt{T})$ for infinite decision sets by appealing to a covering argument (where the algorithm is run on an appropriately fine cover of D). However, this argument leads to significantly less sharp bound in terms of n, which we discuss later (after Theorem 3.2).

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 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 Dani et al. [2008] used a decision region which was a hypercube (so $\Delta > 0$ as this a polytope) — in fact, Δ actually scaled as $1/\sqrt{T}$. In order to negate the possibility of a polylogarithmic rate for a particular problem, we must hold $\Delta = 0$ as we scale the time, which we accomplish in this paper with a more delicate construction using an n-dimensional decision space constructed out of a Cartesian product of 2-dimensional spheres.

1.2 The Price of Bandit Information

It is natural to ask how much worse the regret is in the bandit setting as compared to a setting where we received full information about the complete loss function $f(\cdot)$ at the end of each round. In other words, what is the *price of bandit information*?

For the full information case, Dani et al. [2008] showed the regret is $O^*(\sqrt{nT})$ (which is tight up to log factors). In fact, in the stochastic case considered here, it is not too difficult to show that, in the full information case, the algorithm of "do the best in the past" achieves this rate. Hence, as the regret is $O^*(n\sqrt{T})$ in the bandit case and $O^*(\sqrt{nT})$ (both of which are tight up to log factors), we have characterized the price of bandit information as $\sqrt{n}$, which is a rather mild dependence on n for having such limited feedback.

We should also note that the work in Dani et al. [2008] considers the adversarial case, where the cost functions are chosen in an arbitrary manner rather than stochastically. Here, it was shown that the regret in the bandit setting is $O^*(n^{3/2}\sqrt{T})$, though it was conjectured that this bound was loose and the optimal rate should be identical to rate for the stochastic case, considered here.

It is striking that the convergence rate for the bandit setting is only a factor of $\sqrt{n}$ worse than in the full information case — in stark contrast to the K-arm bandit setting, where the gap in the dependence on K is exponential ($\sqrt{TK}$ vs. $\sqrt{T\log K}$). See Dani et al. [2008] for further discussion.

2 Preliminaries

Let $D \subset \mathbb{R}^n$ be a compact (but otherwise arbitrary) set of decisions. Without loss of generality, assume this set is of full rank. On each round, we must choose a decision $x_t \in D$. Each such choice results in a cost $\ell_t = c_t(x_t) \in [-1, 1]$.

We assume that, regardless of the history, the conditional expectation of c_t is a fixed linear function, *i.e.*, for all $x \in D$,

$$\mathbb{E}\left(c_t(x) \mid \mathcal{H}_t\right) = \mu \cdot x = \mu^\dagger x \in [-1, 1].$$

where $x \in D$ is arbitrary, and we denote the transpose of any column vector v by $v^\dagger$. (Naturally, the vector μ is unknown, though fixed.) Under these assumptions, the *noise sequence,*

$$\eta_t = c_t(x_t) - \mu \cdot x_t$$

is a martingale difference sequence.

A special case of particular interest is when the cost functions c_t are themselves linear functions sampled independently from some fixed distribution. 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 $\ell_t = c_t(x_t)$ received on that round, *not* the entire cost function $c_t(\cdot)$.

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 μ, *i.e.*,

$$x^* \in \operatorname*{argmin}_{x \in D} \mu^\dagger x$$

which exists since D is compact. Observe that if the mean μ 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).

3 Main Results

3.1 The Algorithm

In Figure 3.1 we present a generalized version of the LinRel algorithm of Auer [2003] to the case where D is infinite. We call our algorithm the ConfidenceEllipsoid Algorithm to emphasize the fact that it maintains an ellipsoidal region in which μ is contained with high probability. Due to this ellipsoidal shape, the algorithm may be implemented efficiently for the case when D is convex and when given certain oracle optimization access to D (i.e. the ability to optimize linear functions over D).

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 $\widehat{\mu}$ to the true mean cost vector μ can be constructed by minimizing the square loss:

$$\widehat{\mu} := \operatorname*{argmin}_{\nu} \mathcal{L}(\nu), \text{ where } \mathcal{L}(\nu) := \sum_{\tau < t} \left(\nu^\dagger x_\tau - \ell_\tau \right)^2.$$

Defining $A = \sum x_\tau x_\tau^\dagger$, we have that the least squares estimator is

$$\widehat{\mu} = A^{-1} \sum_{\tau < t} \ell_\tau x_\tau .$$

Algorithm 3.1: CONFIDENCEELLIPSOID(D, δ)

Initialization:

Compute an (almost) barycentric spanner $b_1, \ldots, b_n$ for D.

$A_1 = \sum_{i=1}^{n} b_i b_i^\dagger$

$\widehat{\mu}_1 = 0$

for $t \leftarrow 1$ **to** ∞

$\beta_t = \max\left(128 n \ln t \ln(t^2/\delta), \left(\frac{8}{3}\ln\left(\frac{t^2}{\delta}\right)\right)^2\right)$ (this is $O(n \log^2 t)$)

$B_t = \left\{\nu \colon (\nu - \widehat{\mu}_t)^\dagger A_t (\nu - \widehat{\mu}_t) \leq \beta_t\right\}$

$x_t = \operatorname*{argmin}_{x \in D} \min_{\nu \in B_t} (\nu^\dagger x)$

Incur and observe loss $\ell_t := c_t(x_t)$

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

$\widehat{\mu}_{t+1} = A_{t+1}^{-1} \sum_{\tau=1}^{t} \ell_\tau x_\tau$

Figure 1: The Algorithm ConfidenceEllipsoid: See text for details.

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

$$\{\nu \mid \mathcal{L}(\nu) - \mathcal{L}(\widehat{\mu}_t) \leq \beta\}$$

It is straightforward to see that:

$$\mathcal{L}(\nu) - \mathcal{L}(\widehat{\mu}) = (\nu - \widehat{\mu})^\dagger A (\nu - \widehat{\mu})$$

Thus the confidence region proposed above has the shape of an ellipsoid centered on $\widehat{\mu}$, where the axes are defined through A. This set is commonly referred to the as set of vectors ν with bounded Mahalanobis distance with respect to mean $\widehat{\mu}$ 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 $\widehat{\mu}_t$ at time t by

$$\widehat{\mu}_t = A_t^{-1} \sum_{\tau < t} \ell_\tau x_\tau \,.$$

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 $\widehat{\mu}_t$ 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

$$B_t := \{\nu \mid (\nu - \widehat{\mu}_t)^\dagger A_t(\nu - \widehat{\mu}_t) \leq \beta_t\}$$

In the proofs, we show that, with our choice of β_t, μ always remains inside this ellipsoid for all times t, with high probability.

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

$$x_t = \operatorname*{argmin}_{x \in D} \min_{\nu \in B_t} (\nu^\dagger x).$$

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 $\widehat{\mu}_t$ is an approximation to μ, the function $x \mapsto \widehat{\mu}_t \cdot x$ may be far from being a feasible (i.e. $[-1, 1]$-valued) cost function on D — however, it is bounded in $[-n, n]$.

3.2 Upper Bounds

In the traditional K-arm bandit literature, the regret is often characterized for a particular problem in terms of T, K, and problem dependent constants. In the K-arm bandit results of Auer et al. [2002], this problem dependent constant is the "gap" between the loss of the best arm and the second best arm.

We cannot naively use the same definition since if the decision space is, say a convex set, then there is no well defined notion of second best arm. Instead, we define the gap as follows. Let $\mathcal{E}$ denote the set of extremal points of the decision set D, where an *extremal point* of D is defined as a point which is not a proper convex combination of points in D. It is easy to see that any linear loss function on D always attains its minimum value at a point in $\mathcal{E}$. It is not too difficult to show that ConfidenceEllipsoid always plays extremal points, due to the strict convexity of the confidence region. Now define the set of suboptimal extremal points as:

$$\mathcal{E}_- = \{x \in \mathcal{E} : \mu \cdot x > \mu \cdot x^*\}$$

and define the gap, Δ, as

$$\Delta = \inf_{x \in \mathcal{E}_-} \mu \cdot x - \mu \cdot x^*$$

so the Δ is just the difference in costs between the optimal and next to optimal decision among the extremal points. Note that if D is a fixed polytope then the $\Delta > 0$. However, if D is a ball then $\Delta = 0$, as all points on the surface (a sphere) are extremal — so $\inf_{x \in \mathcal{E}_-} \mu \cdot x$ limits to $\mu \cdot x^*$ (and no point in $\mathcal{E}_-$ achieves this value).

We now state the first upper bound, which is a problem dependent bound stated in terms of Δ.

Theorem 3.1. *(Problem Dependent Upper Bound) Let* $0 < \delta < 1$. *Suppose the decision set* D *and the true mean* μ *have a gap* $\Delta > 0$. *Then for all sufficiently large* T, *the cumulative regret* R_T *of ConfidenceEllipsoid(D,δ) is with high probability at most* $\frac{n^2}{\Delta} \log^3 T$. *More precisely,*

$$\text{Prob}\left(\forall T, \;\; R_T \leq \frac{8n\beta_T \ln(T)}{\Delta}\right) \geq 1 - \delta,$$

where $\beta_T = \max\left(128n \ln T \ln(T^2/\delta), \left(\frac{8}{3} \ln\left(\frac{T^2}{\delta}\right)\right)^2\right)$.

Analogous to the K-arm case, when $\Delta > 0$, 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 decision region.

The following upper bound is stated without regard to the specific parameter Δ for a given problem. Furthermore, it also holds for the case when $\Delta = 0$.

Theorem 3.2. *(Problem Independent Upper Bound) Let $0 < \delta < 1$. Then for all sufficiently large T, the cumulative regret R_T of ConfidenceEllipsoid(D,δ) is with high probability at most $O^*(n\sqrt{T})$, where the O^* notation hides a polylogarithmic dependence on T. More precisely,*

$$\text{Prob}\left(\forall T,\ \ R_T \leq \sqrt{8nT\beta_T \ln T}\right) \geq 1 - \delta.$$

where $\beta_T = \max\left(128n \ln T \ln(T^2/\delta), \left(\frac{8}{3}\ln\left(\frac{T^2}{\delta}\right)\right)^2\right)$.

We note that Auer [2003] achieves a rate that is $O^*((\log|D|)^{3/2}\sqrt{nT})$, with a more complicated algorithm for any finite decision set D. Using his result, one can derive the less sharp bound of $O^*(n^{5/2}\sqrt{T})$ for arbitrary compact decision sets with two observations. First, through a covering argument, we need only consider D to be exponential in n. Second, Auer [2003] 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 4 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.

The following subsection shows our bound of $O^*(n\sqrt{T})$ is tight, in terms of both n and T. Also, as mentioned in the Introduction, tightly characterizing the dimensionality dependence allows us to show that the price of bandit information is only $(\sqrt{n})$.

3.3 Lower Bounds

Note that our upper bounds still leave open the possibility that there is a 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. We now provide that a lower bound showing in fact that this is not possible, by providing a $\Omega(n\sqrt{T})$ lower bound.

For the lower bound, we must consider a decision region with $\Delta = 0$, which rules out polytopes and finite sets (so the decision region of a hypercube, used by Dani et al. [2008], 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, $D_n = \{(x_1, \ldots, x_n) \colon x_1^2 + x_2^2 = x_3^2 + x_4^2 = \cdots = x_{n-1}^2 + x_n^2 = 1\}$. 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 be $\mu \cdot x$, where $n\mu \in D_n$. Since each cost function is only be evaluated at one point, any two distributions over $\{-1, +1\}$-valued cost functions with the same value of μ are equivalent for the purposes of our model.

Theorem 3.3. *(Lower Bound) If μ is chosen uniformly at random from the set D_n/n, and the cost for each $x \in D_n$ is in $\{-1, +1\}$ with mean $\mu \cdot x$, then, for every algorithm, for every $T \geq 1$,*

$$\mathbb{E}\, R = \mathbb{E}_{\mu}\, \mathbb{E}\,(R \mid \mu) \geq \frac{1}{10} n\sqrt{T}.$$

where the inner expectation is with respect to observed costs.

In addition to showing that a polylogarithmic rate is not achievable in general, this bound shows our upper bound is tight in terms of n and T. Again, contrast this with the K-arm case where the optimal asymptotic regret for any given problem is always logarithmic in T.

4 Upper Bound Analysis

Throughout the proof, without loss of generality, assume that the barycentric spanner is the standard basis $\vec{e}_1 \dots \vec{e}_n$ (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 μ lies in the "confidence ellipsoid" B_t. The following shows that event E occurs with high probability. More precisely,

Theorem 4.1. *(Confidence) Let $\delta > 0$. Then*

$$\text{Prob}\left(\forall t, \quad \mu \in B_t\right) \geq 1 - \delta.$$

Subsection 7.2 (in the Appendix) is devoted to establishing this confidence bound. The proof centers on a rather delicate construction In essence, the proof seeks to understand the growth of the quantity $(\widehat{\mu}_t - \mu)^\dagger A_t (\widehat{\mu}_t - \mu)$, which involves a rather technical construction of a martingale (using the matrix inversion lemma) along with a careful application of Freedman's inequality (Theorem 6.1).

The second main step in analyzing ConfidenceEllipsoid(D,δ) is to show that, as long as the aforementioned high-probability event holds, we have some control on the growth of the regret. The following bounds the sum of the squares of instantaneous regret.

Theorem 4.2. *(Sum of Squares Regret Bound) Let*

$$r_t = \mu \cdot x_t - \mu \cdot x^*$$

denote the instantaneous regret acquired by the algorithm on round t. If $\mu \in B_t$ for all $t \leq T$, then

$$\sum_{t=1}^{T} r_t^2 \leq 8 n \beta_T \ln T$$

This is proven in the Appendix (in Subsection 7.1). The idea of the proof involves a potential function argument on the log volume (i.e. the log determinant) of the "precision matrix" A_t (which tracks how accurate our estimates of μ are in each direction). The proof involves relating the growth of this volume to the regret.

At this point the proofs of Theorems 3.1 and 3.2 diverge. To show the former, we use the gap to bound the regret in terms of $\sum_{t=1}^{T} r_t^2$. For the latter, we simply appeal to the Cauchy-Schwarz inequality.

Using these two results we are able to prove our upper bounds as follows.

Proof of Theorem 3.1. Let us analyze $r_t = \mu \cdot x_t - \mu \cdot x^*$, the regret of ConfidenceEllipsoid on round t. Since ConfidenceEllipsoid always chooses a decision from $\mathcal{E}$, either $\mu \cdot x_t = \mu \cdot x^*$ or $x_t \in \mathcal{E}_-$, so that $\mu \cdot x_t - \mu \cdot x^* \geq \Delta$. Since $\Delta > 0$ it follows that either $r_t = 0$ or $r_t/\Delta \geq 1$ and in either case,

$$r_t \leq \frac{r_t^2}{\Delta}$$

By Theorem 4.2, we see that if $\mu \in B_t$, then

$$\begin{aligned} R_T &= \sum_{t=1}^{T} r_t \\ &\leq \sum_{t=1}^{T} \frac{r_t^2}{\Delta} \\ &\leq \frac{8n\beta_T \ln T}{\Delta} \end{aligned}$$

Applying Theorem 4.1, we see that this occurs with probability at least $1-\delta$, which completes the proof. □

Proof of Theorem 3.2. By Theorems 4.1 and 4.2, we know that with probability at least $1-\delta$, $\sum_{t=1}^{T} r_t^2 \leq 8n\beta_T \ln T$. Applying the Cauchy-Schwarz inequality, we have, with probability at least $1-\delta$

$$\begin{aligned} R_T &= \sum_{t=1}^{T} r_t \\ &\leq \left(T \sum_{t=1}^{T} r_t^2 \right)^{1/2} \\ &\leq \sqrt{8nT\beta_T \ln T} \end{aligned}$$

Substituting $\beta_T = \max\left(128n \ln T \ln(T^2/\delta), \left(\frac{8\ln(T^2/\delta)}{3}\right)^2\right)$ completes the proof. □

5 Lower Bound Analysis

This section analyzes the 2-dimensional case. The extension to the general case is provided in the Appendix.

Assume $n = 2$. Let us condition on the event that $\mu \in \{\mu_1, \mu_2\}$, where $\mu_1, \mu_2 \in D_2/2$ such that $\|\mu_1 - \mu_2\| = \varepsilon$. Note that μ is uniform over $\{\mu_1, \mu_2\}$ in this event. We show that, even conditioned on this additional information, the expected regret is $\Omega(\sqrt{T})$. The conclusion of Theorem 3.3 then follows by an averaging argument.

Let

$$b_t := \mathbf{Pr}\left(\mu = \mu_1 \mid \mathcal{H}_t\right) - \mathbf{Pr}\left(\mu = \mu_2 \mid \mathcal{H}_t\right)$$

be the bias towards μ_1 at time t. Note that $b_0 = 0$, and that the sequence (b_t) is a martingale with respect to $(\mathcal{H}_t)$.

Lemma 5.1. *For all t, for any sequence of decisions $x_1, \dots, x_t$ and outcomes $\ell_1, \dots, \ell_{t-1}$, the regret from round t satisfies*

$$\mathop{\mathbb{E}}_{\mu}(r_t \mid \mathcal{H}_t) \geq \frac{1}{16}\left(\varepsilon^2 + \frac{|b_{t+1} - b_t|^2}{\varepsilon^2}\right)\mathbf{1}\{|b_t| \leq 1/2\}$$

The proof of this Lemma is somewhat technical and is provided in the Appendix.

We are now ready to prove Theorem 3.3 in the $n = 2$ case. We generalize the argument to n-dimensions in the appendix.

Proof of Theorem 3.3 for $n = 2$. Let $\varepsilon = T^{-1/4}$. First, observe that, by Fubini's theorem and linearity of expectation,

$$\begin{aligned}
\mathbb{E}\,R = \mathop{\mathbb{E}}_{\mu}\mathop{\mathbb{E}}_{\mathcal{H}_T}(R \mid \mu) &= \mathop{\mathbb{E}}_{\mu}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\mathcal{H}_t}(r_t \mid \mu) \\
&= \sum_{t=1}^{T}\mathop{\mathbb{E}}_{\mathcal{H}_t}\mathop{\mathbb{E}}_{\mu}(r_t \mid \mathcal{H}_t) \\
&\geq \frac{1}{16}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\mathcal{H}_t}\left(\left(\varepsilon^2 + \frac{|b_{t+1} - b_t|^2}{\varepsilon^2}\right)\mathbf{1}\{|b_t| \leq 1/2\}\right) \qquad \text{.... by Lemma 5.1} \\
&\geq \frac{1}{16}\varepsilon^2 T\,\mathrm{Prob}\left(\text{for all t, } |b_t| \leq 1/2\right) + \frac{1}{16}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\mathcal{H}_t}\left(\frac{|b_{t+1} - b_t|^2}{\varepsilon^2}\mathbf{1}\{|b_t| \leq 1/2\}\right) \\
&= \frac{\sqrt{T}}{16}\left(\mathrm{Prob}\left(\text{for all t, } |b_t| \leq 1/2\right) + \sum_{t=1}^{T}\mathop{\mathbb{E}}_{\mathcal{H}_t}\left(|b_{t+1} - b_t|^2\mathbf{1}\{|b_t| \leq 1/2\}\right)\right)
\end{aligned}$$

Thus, if $\mathrm{Prob}\left(\text{for all } t \leq T\ |b_t| \leq 1/2\right) \geq 1/2 - 1/e$, 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 $|b_t| \geq 1/2$. By Freedman's Bernstein-type inequality for martingales (Theorem 6.1 in the Appendix) applied to the martingale $b_{t \wedge \sigma}$, where $\sigma = \min\{\tau \colon |b_\tau| \geq 1/2\}$, we have

$$\mathrm{Prob}\left((\exists t \leq T)\ |b_t| \geq \frac{1}{2} \text{ and } V \leq \frac{1}{32}\right) \leq 2\exp\left(\frac{-1/4}{1/8 + \varepsilon/3}\right) \leq \frac{2}{e^2} < \frac{1}{e}$$

where $V = \sum_{t=1}^{T}\mathbf{1}\{\forall\tau \leq t,\ |b_\tau| \leq 1/2\}\mathbb{E}\left(|b_{t+1} - b_t|^2 \mid \mathcal{H}_t\right)$. It follows that

$$\mathrm{Prob}\left(V > \frac{1}{32}\right) \geq 1/2.$$

In particular,

$$\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\mathcal{H}_t}\left(|b_{t+1} - b_t|^2\mathbf{1}\{|b_t| \leq 1/2\}\right) \geq \mathbb{E}\,V \geq \frac{1}{64}.$$

completing the proof. $\square$

References

R. Agrawal. Sample mean based index policies with O(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27:1054–1078, 1995.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, 2002. ISSN 0885-6125.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3:397–422, 2003. ISSN 1533-7928.

B. Awerbuch and R. Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the 36th ACM Symposium on Theory of Computing (STOC)*, 2004.

Donald A. Berry and Bert Fristedt. *Bandit Problems: Sequential Allocation of Experiments.* Springer, October 1985.

V. Dani, T. P. Hayes, and S. M. Kakade. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*. 2008.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, Feb. 1975.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4, 1985.

Colin McDiarmid. *Concentration. In Probabilistiic Methods for Algorithmic Discrete Mathematics.* Springer, 1998.

H. Robbins. Some aspects of the sequential design of experiments. In *Bulletin of the American Mathematical Society*, volume 55, 1952.

Appendix

6 Concentration of Martingales

We use the following Bernstein-type concentration inequality for martingales, due to Freedman Freedman [1975] (see also [McDiarmid, 1998, Theorem 3.15]).

Theorem 6.1 (Freedman). *Suppose $X_1, \ldots, X_T$ is a martingale difference sequence, and b is an uniform upper bound on the steps X_i. Let V denote the sum of conditional variances,*

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

Then, for every $a, v > 0$,

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

7 Upper Bound Proofs

7.1 Proof of Theorem 4.2

In this section, we prove Theorem 4.2, which says that the sum of the squares of the instantaneous regrets of the algorithm is small, assuming the evolving confidence ellipsoids always contain the true mean μ. A key insight is that on any round t in which $\mu \in B_t$, 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.

Observation 7.1. *Let $\nu \in B_t$ and $x \in D$. Then*

$$|(\nu - \widehat{\mu}_t)^\dagger x| \leq \sqrt{\beta_t x^\dagger A_t^{-1} x}$$

Proof. A_t is a symmetric positive definite matrix. Hence $A_t^{1/2}$ is a well-defined symmetric positive definite (and hence invertible) matrix. Now we have

$$\begin{aligned}
|(\nu - \widehat{\mu}_t)^\dagger x| &= |(\nu - \widehat{\mu}_t)^\dagger A_t^{1/2} A_t^{-1/2} x| \\
&= |(A_t^{1/2}(\nu - \widehat{\mu}_t))^\dagger A_t^{-1/2} x| \\
&\leq \|A_t^{1/2}(\nu - \widehat{\mu}_t)\| \|A_t^{-1/2} x\| && \text{by Cauchy-Schwarz} \\
&= \sqrt{(\nu - \widehat{\mu}_t)^\dagger A_t (\nu - \widehat{\mu}_t)} \sqrt{x^\dagger A_t^{-1} x} \\
&\leq \sqrt{\beta_t x^\dagger A_t^{-1} x} && \text{since } \nu \in B_t. \quad \square
\end{aligned}$$

Define

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

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 7.2. *Fix t. If $\mu \in B_t$, then*

$$r_t \leq 2 \min\left(\sqrt{\beta_t} w_t, 1\right)$$

Proof. Let $\tilde{\mu} \in B_t$ denote the vector which minimizes the dot product $\tilde{\mu}^\dagger x_t$. By choice of x_t, we have

$$\tilde{\mu}^\dagger x_t = \min_{\nu \in B_t} \min_{x \in D} \nu^\dagger x \leq \mu^\dagger x^*,$$

where the inequality used the hypothesis $\mu \in B_t$. Hence,

$$\begin{aligned}
r_t &= \mu^\dagger x_t - \mu^\dagger x^* \\
&\leq (\mu - \tilde{\mu})^\dagger x_t \\
&= (\mu - \widehat{\mu}_t)^\dagger x_t + (\widehat{\mu}_t - \tilde{\mu})^\dagger x_t \\
&\leq 2\sqrt{\beta_t} w_t
\end{aligned}$$

where the last step follows from Observation 7.1 since $\tilde{\mu}$ and μ are in B_t. Since $\ell_t \in [-1, 1]$, r_t is always at most 2 and the result follows. $\square$

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

Lemma 7.3. *Let $t \leq T$. If $\mu \in B_\tau$ for all $\tau \leq t$, then*

$$\sum_{\tau=1}^{t} \min\left(w_\tau^2, 1\right) \leq 2n \ln t.$$

To prove this, we need to track the change in the confidence ellipsoid from round t to round $t+1$. The following two facts prove useful to this end.

Lemma 7.4. *For every $t \leq T$,*

$$\det A_{t+1} = \prod_{\tau=1}^{t} (1 + w_t^2).$$

Proof. By the definition of A_{t+1}, we have

$$\begin{aligned}
\det A_{t+1} &= \det(A_t + x_t x_t^\dagger) \\
&= \det(A_t^{1/2}(I + A_t^{-1/2} x_t x_t^\dagger A_t^{-1/2}) A_t^{1/2}) \\
&= \det(A_t) \det(I + A_t^{-1/2} x_t (A_t^{-1/2} x_t)^\dagger) \\
&= \det(A_t) \det(I + v_t v_t^\dagger),
\end{aligned}$$

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

$$(I + v_t v_t^\dagger) v_t = v_t + v_t (v_t^\dagger v_t) = (1 + w_t^2) v_t$$

Hence $(1 + w_t^2)$ is an eigenvalue of $I + v_t v_t^\dagger$. Since $v_t v_t^\dagger$ is a rank one matrix, all the other eigenvalues of $I + v_t v_t^\dagger$ equal 1. It follows that $\det(I + v_t v_t^\dagger)$ is $(1 + w_t^2)$, 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. □

Lemma 7.5. *For all t, $\det A_t \leq t^n$.*

Proof. The rank one matrix $x_t x_t^\dagger$ has $x_t^\dagger x_t = \|x_t\|^2$ as its unique non-zero eigenvalue. Also, since we have identified the spanner with the standard basis, we have $\sum_{i=1}^{n} b_i b_i^\dagger = I$. Since the trace is a linear operator, it follows that

$$\operatorname{trace} A_t = \operatorname{trace}\left(I + \sum_{\tau<t} x_t x_t^\dagger\right) = n + \sum_{\tau<t} \operatorname{trace}(x_t x_t^\dagger) = n + \sum_{\tau<t} \|x_\tau\|^2 \leq nt.$$

Now, recall that $\operatorname{trace} A_t$ 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. □

Proof of Lemma 7.3. Using the fact that for $0 \leq y \leq 1$, $\ln(1+y) \geq y/2$, we have

$$\begin{aligned}
\sum_{\tau=1}^{t} \min(w_\tau^2, 1) &\leq \sum_{\tau=1}^{t} 2\ln(1 + w_\tau^2) \\
&= 2\ln(\det A_{t+1}) && \text{by Lemma 7.4} \\
&\leq 2n \ln t && \text{by Lemma 7.5} \quad \square
\end{aligned}$$

Finally, we are ready to prove that if μ always stays within the evolving confidence ellipsoid, then our regret is under control.

Proof of Theorem 4.2. Assume that $\mu \in B_t$ for all t. Then

$$\begin{aligned}
\sum_{t=1}^{T} r_t^2 &\leq \sum_{t=1}^{T} 4\beta_t \min(w_t^2, 1) && \text{by Lemma 7.2} \\
&\leq 4\beta_T \sum_{t=1}^{T} \min(w_t^2, 1) && \text{since } 1 < \beta_1 < \cdots < \beta_T \\
&\leq 8\beta_T n \ln T && \text{by Lemma 7.3}
\end{aligned}$$

as desired. $\square$

7.2 Proof of Theorem 4.1

In this section, we prove Theorem 4.1, which states that with high probability, for all t the true mean μ lies in the confidence ellipsoid B_t.

Recall that

$$\eta_t := c_t(x_t) - \mu^\dagger x_t = \ell_t - \mathbb{E}\left(\ell_t \mid \mathcal{H}_t\right)$$

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

Let us define

$$Z_t := (\widehat{\mu}_t - \mu)^\dagger A_t (\widehat{\mu}_t - \mu)$$

which measures the error of $\widehat{\mu}_t$ as an approximation to the true mean, μ, under the norm induced by A_t. Note that $\mu \in B_t$ if and only if $Z_t \leq \beta_t$. The next lemma bounds the growth of Z_t.

Lemma 7.6. *For all t,*

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

Proof. For notational convenience, define:

$$Y_t = A_t(\widehat{\mu}_t - \mu)$$

We have the following relations:

$$\begin{aligned}
Z_t &= Y_t^\dagger A_t^{-1} Y_t \\
Y_t &= \sum_{\tau < t} \eta_\tau x_\tau - \mu \\
Y_{t+1} &= Y_t + \eta_t x_t
\end{aligned}$$

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

Now examining the growth of Z_t, we have:

$$\begin{aligned}
Z_{t+1} &= Y_{t+1}^\dagger A_{t+1}^{-1} Y_{t+1} \\
&= (Y_t + \eta_t x_t)^\dagger A_{t+1}^{-1}(Y_t + \eta_t x_t) \\
&= Y_t^\dagger A_{t+1}^{-1} Y_t + 2\eta_t x_t^\dagger A_{t+1}^{-1} Y_t + \eta_t^2 x_t^\dagger A_{t+1}^{-1} x_t \qquad (1)
\end{aligned}$$

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

$$\begin{aligned} A_{t+1}^{-1} &= (A_t + x_t x_t^\dagger)^{-1} \\ &= A_t^{-1} - \frac{A_t^{-1} x_t x_t^\dagger A_t^{-1}}{1 + x_t^\dagger A_t^{-1} x_t} \\ &= A_t^{-1} - \frac{A_t^{-1} x_t x_t^\dagger A_t^{-1}}{1 + w_t^2} \end{aligned}$$

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

$$\begin{aligned} Y_t^\dagger A_{t+1}^{-1} Y_t &= Y_t^\dagger A_t^{-1} Y_t - \frac{(Y_t^\dagger A_t^{-1} x_t)^2}{1 + w_t^2} \\ &\leq Z_t. \end{aligned}$$

For the second term,

$$\begin{aligned} 2\eta_t x_t^\dagger A_{t+1}^{-1} Y_t &= 2\eta_t x_t^\dagger A_t^{-1} Y_t - 2\eta_t \frac{x_t^\dagger A_t^{-1} x_t x_t^\dagger A_t^{-1} Y_t}{1 + w_t^2} \\ &= 2\eta_t x_t^\dagger A_t^{-1} Y_t - 2\eta_t \frac{w_t^2 x_t^\dagger A_t^{-1} Y_t}{1 + w_t^2} \\ &= 2\eta_t \frac{x_t^\dagger A_t^{-1} Y_t}{1 + w_t^2} \\ &= 2\eta_t \frac{x_t^\dagger (\widehat{\mu}_t - \mu)}{1 + w_t^2} \end{aligned}$$

For the third term,

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

Putting these together, we have shown

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

By induction, it follows that

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

Finally, we check that $Z_1 \leq n$. To see this, recall from the algorithm that $A_1 = I$ and $\widehat{\mu}_1 = 0$.

Also, since $\vec{e_1}, \ldots, \vec{e_n} \in D$, by assumption, $\mu \cdot \vec{e_j} \in [-1, 1]$.

$$\begin{aligned} Z_1 &= (\widehat{\mu}_1 - \mu)^\dagger A_1 (\widehat{\mu}_1 - \mu) \\ &= \|\mu\|^2 \\ &= \sum_{j=1}^{n} (\mu^\dagger \vec{e_j})^2 \\ &\leq n. \end{aligned}$$

□

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

$$E_t = \mathrm{I}\{Z_\tau \leq \beta_\tau \text{ for all } \tau \leq t\} = \mathrm{I}\{\mu \in B_\tau \text{ for all } \tau \leq t\}$$

where $\mathrm{I}\{\cdot\}$ is the indicator function.

Lemma 7.7. *Define a random variable M_t by*

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

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:

$$\begin{aligned} \mathbb{E}(M_t \mid \mathcal{H}_t) &= 2E_t \frac{x_t^\dagger (\widehat{\mu}_t - \mu)}{1 + w_t^2} \mathbb{E}(\eta_t \mid \mathcal{H}_t) \\ &= 0 \end{aligned}$$

since the history fully determines $x_1, \ldots, x_t$, $\widehat{\mu}_1, \ldots, \widehat{\mu}_t$, $Z_1, \ldots, Z_t$, and $E_1, \ldots, E_t$, and since the noise functions η_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 7.8. *Given $\delta < 1$,*

$$\mathrm{Prob}\left(\forall t, \ \sum_{\tau=1}^{t-1} M_\tau \leq \beta_t / 2\right) \geq 1 - \delta,$$

where β_t is defined in the statement of ConfidenceEllipsoid(D,δ).

We defer the proof to Section 7.2.1. Equipped with this lemma, we can prove Theorem 4.1.

Proof of Theorem 4.1. It suffices to show that the high-probability event described in Lemma 7.8 is contained in the support of E_t for every t. We prove the latter by induction on t.

By Lemma 7.6 and the definition of β_1, we know that $Z_1 \leq n < \beta_1$. Hence E_1 is always 1 (equivalently, μ is always in B_1).

Now suppose the high-probability event of Lemma 7.8 holds, so in particular,

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

By inductive hypothesis, $E_\tau = 1$ for $\tau \le t-1$. Hence by Lemma 7.6 we have

$$\begin{aligned}
Z_t &\le n + 2\sum_{\tau=1}^{t-1} \eta_\tau \frac{x_\tau^\dagger(\widehat{\mu}_\tau - \mu)}{1+w_\tau^2} + \sum_{\tau=1}^{t-1} \eta_\tau^2 \frac{w_\tau^2}{1+w_\tau^2} \\
&= n + \sum_{\tau=1}^{t-1} M_\tau + \sum_{\tau=1}^{t-1} \eta_\tau^2 \frac{w_\tau^2}{1+w_\tau^2} \\
&\le n + \beta_t/2 + \sum_{\tau=1}^{t-1} \eta_\tau^2 \frac{w_\tau^2}{1+w_\tau^2} \\
&\le n + \beta_t/2 + \sum_{\tau=1}^{t-1} \min(w_\tau^2, 1) && \text{since } |\eta_\tau| \le 1 \\
&\le n + \beta_t/2 + 2n \ln t && \text{by Lemma 7.3} \\
&\le \beta_t.
\end{aligned}$$

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

7.2.1 Concentration

All that remains to complete the proof now is to show that our martingale $\sum_1^t M_\tau$ 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 $\sum_1^t M_\tau$ 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 6.1, which allows us to to bound the step sizes in terms of random variables, as long as the conditional variances remain under control.

Proof of Lemma 7.8. Let us first obtain upper bounds on the step sizes of our martingale.

$$\begin{aligned}
|M_t| &= 2|\eta_t| E_t \frac{|x_t^\dagger(\widehat{\mu}_t - \mu)|}{1+w_t^2} \\
&\le 2|\eta_t| E_t \frac{\sqrt{\beta_t x_t^\dagger A_t^{-1} x_t}}{1+w_t^2} \\
&= 2|\eta_t| E_t \frac{w_t \sqrt{\beta_t}}{1+w_t^2} \\
&\le 2|\eta_t| E_t \sqrt{\beta_t} \min(w_t, 1/2) \qquad (2)
\end{aligned}$$

where the first inequality follows trivially when $E_t = 0$, and by Observation 7.1 when $E_t = 1$. Additionally this gives a family of uniform upper bounds:

$$|M_\tau| \le \sqrt{\beta_t} \text{ for all } \tau \le t$$

since $|\eta_t| \le 1$ and (by choice) β_τ is a non-decreasing sequence.

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

$$
\begin{aligned}
V_t &:= \sum_{\tau=1}^{t} \mathbf{Var}\left(M_\tau \mid M_1 \dots M_{\tau-1} \right) \\
&\leq \sum_{\tau=1}^{t} 4E_\tau \beta_\tau \min\left(w_\tau^2, 1/4\right) && \text{by (2) and since } |\eta_\tau| \leq 1 \\
&\leq 4(\max_{\tau \leq t} \beta_\tau) \sum_{\tau=1}^{t} E_\tau \min(w_\tau^2, 1) \\
&\leq 4\beta_t \sum_{\tau \leq t} E_\tau \min(w_\tau^2, 1) \\
&\leq 8\beta_t n \ln\left(\max\{\tau \leq t \mid E_\tau = 1\}\right) && \text{by Lemma 7.3} \\
&\leq 8\beta_t n \ln t
\end{aligned}
$$

Since we have established that the sum of conditional variances, V_t, is *always* bounded by $8\beta_t n \ln t$, we can apply Theorem 6.1 with parameters $a = \beta_t/2$, $b = \sqrt{\beta_t}$ and $v = 8n\beta_t \ln t$, to get

$$
\begin{aligned}
\text{Prob}\left(\sum_{\tau=1}^{t-1} M_\tau \geq \beta_t/2\right) &= \text{Prob}\left(\sum_{\tau=1}^{t-1} M_\tau \geq \beta_t/2 \text{ and } V_t \leq 8n\beta_t \ln t\right) \\
&\leq \exp\left(\frac{-(\beta_t/2)^2}{2(8n\beta_t \ln t) + \frac{2}{3}(\beta_t/2)(\sqrt{\beta_t})}\right) \\
&= \exp\left(\frac{-\beta_t}{64n \ln t + \frac{4}{3}\sqrt{\beta_t}}\right) \\
&\leq \max\left\{\exp\left(\frac{-\beta_t}{128n \ln t}\right), \exp\left(\frac{-3\sqrt{\beta_t}}{8}\right)\right\} \\
&\leq \frac{\delta}{t^2}
\end{aligned}
$$

where the last inequality follows from the definition of β_t. Finally, we apply a union bound to get

$$
\begin{aligned}
\text{Prob}\left(\sum_{\tau=1}^{t-1} M_\tau \geq \frac{\beta_t}{2} \text{ for some } t\right) &\leq \sum_{t=1}^{\infty} \text{Prob}\left(\sum_{\tau=1}^{t-1} M_\tau \geq \frac{\beta_t}{2}\right) \\
&\leq \sum_{t=2}^{\infty} \frac{\delta}{t^2} \\
&\leq \delta\left(\frac{\pi^2}{6} - 1\right) \\
&\leq \delta
\end{aligned}
$$

completing the proof of Lemma 7.8. $\square$

8 Lower Bound Proofs

We first prove Lemma 5.1. Then we generalize our argument to the n dimensional case in the following subsection.

Proof. (Proof of Lemma 5.1). Let v_1 be the unit vector in the direction of $\mu_1 - \mu_2$, and let v_2 be the unit vector in the direction of $\mu_1 + \mu_2$. Note that v_1, v_2 is an orthonormal basis for the plane. Decompose $x_t = \alpha v_1 + \beta v_2$, and $\mathbb{E}(\mu \mid \mathcal{H}_t) = \gamma v_1 + \delta v_2$. Since $\mathbb{E}(\mu \mid \mathcal{H}_t) = \frac{\mu_1+\mu_2}{2} + b_t \frac{\mu_1-\mu_2}{2}$, we have $\gamma = \varepsilon b_t/2$ and $\delta = \frac{\sqrt{1-\varepsilon^2}}{2}$.

Let $p = \mathbf{Pr}\left(\mu = \mu_1 \mid \mathcal{H}_t\right)$. Then $b_t = 2p - 1$. Observe that

$$\begin{aligned}
b_{t+1} - b_t &= \frac{p(1+\ell_t\mu_1\cdot x_t) - (1-p)(1+\ell_t\mu_2\cdot x_t)}{p(1+\ell_t\mu_1\cdot x_t) + (1-p)(1+\ell_t\mu_2\cdot x_t)} - 2p + 1 \\
&= \frac{(2p-1) + p\ell_t\mu_1\cdot x_t - (1-p)\ell_t\mu_2\cdot x_t}{1 + p\ell_t\mu_1\cdot x_t + (1-p)\ell_t\mu_2\cdot x_t} - 2p + 1 \\
&= \frac{p\ell_t\mu_1\cdot x_t - (1-p)\ell_t\mu_2\cdot x_t - (2p-1)(p\ell_t\mu_1\cdot x_t + (1-p)\ell_t\mu_2\cdot x_t)}{1 + p\ell_t\mu_1\cdot x_t + (1-p)\ell_t\mu_2\cdot x_t} \\
&= \frac{2p(1-p)\ell_t\mu_1\cdot x_t - 2p(1-p)\ell_t\mu_2\cdot x_t}{1 + p\ell_t\mu_1\cdot x_t + (1-p)\ell_t\mu_2\cdot x_t} \\
&= \frac{2p(1-p)\ell_t(\mu_1-\mu_2)\cdot x_t}{1 + p\ell_t\mu_1\cdot x_t + (1-p)\ell_t\mu_2\cdot x_t}
\end{aligned}$$

Since $|\mu_i \cdot x_t| \le 1/2$, the denominator of the above expression is at least $1/2$. Since $p(1-p) \le 1/4$, it follows that

$$|b_{t+1} - b_t| \le |(\mu_1 - \mu_2)\cdot x_t| = \varepsilon|\alpha|. \tag{3}$$

Assume the game history is such that $|b_t| \le 1/2$. Otherwise, since the regret is non-negative, there is nothing to prove. Now we calculate

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mu}(r_t \mid \mathcal{H}_t) &= \frac{1}{2} + x_t\cdot\mathbb{E}(\mu \mid \mathcal{H}_t) \\
&= \frac{1}{2} + \alpha\gamma + \beta\delta \\
&= \frac{1}{2}(1 + \alpha\varepsilon b_t + \beta\sqrt{1-\varepsilon^2}) \\
&\ge \frac{1}{2}\left(1 + \alpha\varepsilon b_t + \left(\frac{\alpha^2}{2} - 1\right)\sqrt{1-\varepsilon^2}\right) && (4) \\
&\ge \frac{1}{2}\left(1 + \alpha\varepsilon b_t + \left(\frac{\alpha^2}{2} - 1\right)\left(1 - \frac{\varepsilon^2}{2}\right)\right) \\
&= \frac{1}{16}(\alpha^2 + \varepsilon^2) + \frac{1}{8}(\alpha^2 + 4b_t\alpha\varepsilon + \varepsilon^2) + \frac{1}{16}(\alpha^2 + \varepsilon^2 - 2\alpha^2\varepsilon^2) \\
&\ge \frac{1}{16}(\alpha^2 + \varepsilon^2) && (5) \\
&\ge \frac{1}{16}\left(\varepsilon^2 + \frac{|b_{t+1} - b_t|^2}{\varepsilon^2}\right) && (6)
\end{aligned}$$

Here (4) follows because $\alpha^2 + \beta^2 = 1$ implies that $1 + \beta \ge \alpha^2/2$, with equality iff $\beta = -1$. Inequality (5) follows because $|b_t| \le 1/2$ and $|\alpha|, |\varepsilon| \le 1$. Inequality (6) follows from (3), which completes the proof. □

8.1 The n-Dimensional Case

Now suppose $n > 2$ is even. Fix an index $1 \le i \le n/2$, and consider the contribution to the total expected regret from the choice of (x_{2i-1}, x_{2i}), *i.e.*, the component from the i'th circle.

Analogously to the 2-dimensional case, we condition on the i'th component of μ being one of two vectors, $\nu_1, \nu_2 \in S^1/n$. We further condition on the exact values of the other $n/2 - 1$ components of μ. We denote $\varepsilon = \|\nu_1 - \nu_2\|$

Let b_t denote the bias toward ν_1, given the history $\mathcal{H}_t$ of the game on rounds $1, \ldots, t-1$. That is,

$$b_t = \mathbf{Pr}\left(\mu_i = \nu_1 \mid \mathcal{H}_t\right) - \mathbf{Pr}\left(\mu_i = \nu_2 \mid \mathcal{H}_t\right)$$

Then we have the following analog of Lemma 5.1.

Lemma 8.1. *For all t, for any sequence of decisions $x_1, \ldots, x_t$ and outcomes $\ell_1, \ldots, \ell_{t-1}$, the regret from round t due to the ith component of x_t satisfies*

$$\mathop{\mathbb{E}}_{\mu}(r_t^{(i)} \mid \mathcal{H}_t) \geq \frac{1}{64}\left(\varepsilon^2 + \frac{|b_{t+1} - b_t|^2}{\varepsilon^2}\right) \mathbf{1}\{|b_t| \leq 1/2\}$$

It follows along the same lines as before that the expected total regret from the ith component is $\Omega(\sqrt{T})$. Summing over the $n/2$ possible values of i completes the proof.