Journal of Machine Learning Research 17 (2016) 1-34 Submitted 9/14; Published 4/16

Volumetric Spanners: An Efficient Exploration Basis for Learning

Elad Hazan EHAZAN@CS.PRINCETON.EDU
Princeton University, Department of Computer Science
Princeton, NJ 08540, USA

Zohar Karnin ZKARNIN@YMAIL.COM
Yahoo Haifa Labs
Matam, Haifa 31905, Israel

Editor: Alexander Rakhlin

Abstract

Numerous learning problems that contain exploration, such as experiment design, multi-arm bandits, online routing, search result aggregation and many more, have been studied extensively in isolation. In this paper we consider a generic and efficiently computable method for action space exploration based on convex geometry.

We define a novel geometric notion of an exploration mechanism with low variance called volumetric spanners, and give efficient algorithms to construct such spanners. We describe applications of this mechanism to the problem of optimal experiment design and the general framework for decision making under uncertainty of bandit linear optimization. For the latter we give efficient and near-optimal regret algorithm over general convex sets. Previously such results were known only for specific convex sets, or under special conditions such as the existence of an efficient self-concordant barrier for the underlying set. [1]

Keywords: barycentric spanner, volumetric spanner, linear bandits, hard margin linear regression

1. Introduction

A fundamental challenge in machine learning is environment exploration. A prominent example is the famed multi-armed bandit (MAB) problem, in which a decision maker iteratively chooses an action from a set of available actions and receives a payoff, without observing the payoff of all other actions she could have taken. The MAB problem displays an exploration-exploitation tradeoff, in which the decision maker trades exploring the action space vs. exploiting the knowledge already obtained to pick the best arm. The exploration challenge arises in structured bandit problems such as online routing and rank aggregation: how to choose the most informative path in a graph, or the most informative ranking?

Another example in which environment exploration is crucial is experiment design, and more generally the setting of active learning. In this setting it is important to correctly identify the most informative experiments/queries so as to efficiently construct a solution.

1. This paper is the full version of a merger of two extended abstracts (Hazan et al., 2014) and (Karnin and Hazan, 2014).

©2016 Elad Hazan, Zohar Karnin.

Exploration is hardly summarized by picking an action uniformly at random. Indeed, sophisticated techniques from various areas of optimization, statistics and convex geometry have been applied to designing ever better exploration algorithms. To mention a few: (Awerbuch and Kleinberg, 2008) devise the notion of *barycentric spanners*, and use this construction to give the first low-regret algorithms for complex decision problems such as online routing. (Abernethy et al., 2012) use self-concordant barriers to build an efficient exploration strategy for convex sets in Euclidean space. (Bubeck et al., 2012) apply tools from convex geometry, namely the John ellipsoid to construct optimal-regret algorithms for bandit linear optimization, albeit not always efficiently.

In this paper we consider a generic approach to exploration, and quantify what efficient exploration with *low variance* requires in general. Given a set in Euclidean space, a low-variance exploration basis is a subset with the following property: given noisy estimates of a linear function over the basis, one can construct an estimate for the linear function over the entire set without increasing the variance of the estimates.

By definition, such low variance exploration bases are immediately applicable to noisy linear regression: given a low-variance exploration basis, it suffices to learn the function values only over the basis in order to interpolate the value of the underlying linear regressor over the entire decision set. This fact can be used for active learning as well as for the exploration component of a bandit linear optimization algorithm.

Henceforth we define a novel construction for a low variance exploration basis called **volumetric spanners** and give efficient algorithms to construct them. We further investigate the convex geometry implications of our construction, and define the notion of a **minimal volumetric ellipsoid** of a convex body. We give structural theorems on the existence and properties of these ellipsoids, as well as constructive algorithms to compute them in several cases.

We complement our findings with two applications to machine learning. The first application is to the problem of experiment design, in which we give an efficient algorithm for hard-margin active linear regression with optimal bounds. Next, we advance a well-studied open problem that has exploration as its core difficulty: an efficient and near-optimal regret algorithm for bandit linear optimization (BLO). We expect that volumetric spanners and volumetric ellipsoids can be useful elsewhere in experiment design and active learning.

1.1 Informal statement of results

Experiment design: In the statistical field called *optimal design of experiments*, or just *optimal design* (Atkinson and Donev, 1992; Wu, 1978) , a statistician is faced with the task of choosing experiments to perform from a given pool, with the goal of producing the optimal result within the budget constraint.

Formally, consider a pool of possible experiments denoted $x_1, ..., x_n \in \mathbb{R}^d$. The goal of the designer is to choose a distribution over the pool of experiments, such that experiments chosen according to this distribution produce a hypothesis $\hat{\mathbf{w}}$ that is as close as possible to the true linear function behind the data. The distance between the hypothesis and true linear function can be measured in different ways, each corresponding to a different *optimality criteria*. The common property of the criteria is that they all minimize the variance of the hypothesis. Since the variance is not a scalar but a $d \times d$ matrix, the different

criteria differ by the fact that each one minimizes a different function $\Phi : \mathcal{R}^{d\times d} \to \mathcal{R}$ over the covariance matrix. Common criteria are the A-, D-, and E-optimality criteria. D-optimality, minimizes the determinant of the covariance matrix, and thus minimizes the volume of the confidence region. In A-optimality the trace of the covariance matrix, i.e., the total variance of the parameter estimates, is minimized. E-optimality minimizes the maximum eigenvalue of the covariance matrix, and thus minimizes the size of the major axis of the confidence region.

The above criteria do not directly characterize the quality of predictions on test data. A common criterion that directly takes the test data into account is that of G-optimality. Here the goal is to minimize the maximum variance of the predicted values. In other words, by denoting $\mathrm{Var}_S(x_i)$ the variance of the prediction of x_i after querying the points of S, the goal in G-optimality is to minimize $\max_i \mathrm{Var}_S(x_i)$. G-optimality and D-optimality are closely related in the sense that an exact solution to one is the solution to the other, see for example (Spruill and Studden, 1979).

In this paper we solve a problem closely related to the G-optimality criteria. Given a pool of data points $\mathcal{K} \in \mathcal{R}^d$, say representing a pool of patients, we aim to solve an active regression problem finding w.h.p a regressor minimizing the *worst-case error*, while minimizing the number of (noisy) queries to the regressor. The formal definition of the problem is given in Section 6. This problem differs from classic linear regression results as there, the *mean square error* is bounded. The difference between the described problem and solving the optimal design problem with the G-optimality criteria is two-fold. First, we do not aim to minimize the variance but to obtain a high probability bound. Second, in optimal design the quality of the distribution is measured when the budget tends to infinity. Specifically, notice that for a distribution over the possible experiments, rather than a deterministic subset of them, the corresponding covariance matrix is random. The discussed minimizations are done over the expected covariance matrix, where the expectation is taken over the subset of chosen experiments. When the budget tends to infinity the actual covariance matrix is close w.h.p to its expected counterpart. We call this the *infinite budget setting*. Ours is a finite budget setting where one does not aim to provide a distribution over the possible experiments but a deterministic subset of them of a fixed size.

For the finite budget setting various relaxations have been considered in the statistical literature, usually without an approximation guarantee. Our method differs from previous works of this spirit by: First, we do not impose a hard-budget constraint of experiments, but rather bound the number of experiments as a function of the desired approximation guarantee. Second, we obtain a computationally efficient algorithm with provable optimality results. Finally, as an added bonus our solution has the property of choosing very few data points to explore, potentially much less than the budget. A motivating example for this property is the medical experiment design. Here a data point is a human subject and it is more realistic to have few volunteers being thoroughly tested on as opposed to performing few tests over many volunteers. Our setting is arguably more natural for the medical-patient-experiment motivating example; in general there are numerous examples where the budget of experiments is not fixed but rather the tolerable error. Of equal importance is the fact that our setting allows to derive efficient algorithms with rigorous theoretical guarantees.

A related and recently popular model is called random design (Hsu et al., 2012; Audibert and Catoni, 2010; Györfi et al., 2006; Audibert and Catoni, 2011). In this setting the designer is given a set of measurements $\{x_i, y_i | i \in [n]\}$ for $x_i \in \mathbb{R}^d$ drawn from an unknown distribution $\mathcal{D}$. The goal is to predict as well as the best linear predictor measured according to the mean square error, i.e., minimize

$$\mathop{\mathbf{E}}_{(x,y)\in\mathcal{D}} \left[(x^\top w - y)^2 - (x^\top w^* - y)^2 \right]$$

where w^* is the optimal linear regressor. Various other performance metrics have been considered in the referenced papers, i.e., measuring the norm of the regressor vs. the optimal regressor in a norm proportional to the covariance matrix. However, in this setting an expected error is the criterion vs. our criterion of worst-case, or a high confidence bound on the error[2], which is more suitable for some experiment design settings.

Active learning: The most well-studied setting in active learning is pool-based active learning (McCallum and Nigam, 1998), in which the learner has access to a pool of examples, and can iteratively query labels of particular examples of her choice. Compared to passive learning, in which labelled examples are drawn from a fixed unknown distribution, known active learning algorithms can attain a certain generalization error guarantee albeit observing exponentially fewer labelled examples, e.g., (Cohn et al., 1994; Dasgupta et al., 2009; Hanneke, 2007; Balcan et al., 2009), under certain assumptions such as special hypothesis classes, realizability or large-margin. Active learning with noise is a much less studied topic: (Balcan et al., 2009) give an exponential improvement over passive learning of linear threshold functions, but under the condition that the noise is smaller than the desired accuracy. Real valued active learning with a soft-margin criteria was addressed in (Ganti and Gray, 2012). The reader is referred to (Dasgupta and Langford, 2009) for a more detailed survey of active learning literature.

Bandit Linear Optimization Bandit linear optimization (BLO) is a fundamental problem in decision making under uncertainty that efficiently captures structured action sets. The canonical example is that of online routing in graphs: a decision maker iteratively chooses a path in a given graph from source to destination, the adversary chooses lengths of the edges of the graph, and the decision maker receives as feedback the length of the path she chose but no other information (Awerbuch and Kleinberg, 2008). Her goal over many iterations is to attain an average travel time as short as that of the best fixed shortest path in the graph.

This decision problem is readily modeled in the "experts" framework, albeit with efficiency issues: the number of possible paths is potentially exponential in the graph representation. The BLO framework gives an efficient model for capturing such structured decision problems: iteratively a decision maker chooses a point in a convex set and receives as a payoff an adversarially chosen linear cost function. In the particular case of online routing, the decision set is taken to be the s-t-flow polytope, which captures the convex hull of all source-destination shortest paths in a given graph, and has a succinct representation

2. Our results though stated as a worst case error can be generalized to a high probability solution in the random design scenario.

with polynomially many constraints and low dimensionality. The linear cost function corresponds to a weight function on the graphs edges, where the length of a path is defined as the sum of weights of its edges.

The BLO framework captures many other structured problems efficiently, e.g., learning permutations, rankings and other examples (Abernethy et al., 2012). As such, it has been the focus of much research in the past few years. The reader is referred to the recent survey of (Bubeck and Cesa-Bianchi, 2012a) for more details on algorithmic results for BLO. Let us remark that certain online bandit problems do not immediately fall into the convex BLO model that we address, such as combinatorial bandits studied in (Cesa-Bianchi and Lugosi, 2012).

In this paper we contribute to the large literature on the BLO model by giving the first polynomial-time and near optimal-regret algorithm for BLO over general convex decision sets; see Section 7 for a formal statement. Previously efficient algorithms, with non-optimal-regret, were known over convex sets that admit an efficient self-concordant barrier (Abernethy et al., 2012), and optimal-regret algorithms were known over general sets (Bubeck et al., 2012) but these algorithms were not computationally efficient. Our result, based on volumetric spanners, is able to attain the best of both worlds.

1.2 Volumetric Ellipsoids and Spanners

We now describe the main convex geometric concepts we introduce and use for low variance exploration. To do so we first review some basic notions from convex geometry.

Let $\mathbb{R}^d$ be the d-dimensional vector space over the reals. Given a set of vectors $S = \{v_1, \ldots, v_t\} \subset \mathbb{R}^d$, we denote by $\mathcal{E}(S)$ the ellipsoid defined by S:

$$\mathcal{E}(S) = \left\{ \sum_{i \in S} \alpha_i v_i \ : \ \sum_i \alpha_i^2 \leq 1 \right\}.$$

By abuse of notation, we also say that $\mathcal{E}(S)$ is *supported* on the set S.

Ellipsoids play an important role in convex geometry and specific ellipsoids associated with a convex body have been used in previous works in machine learning for designing good exploration bases for convex sets $\mathcal{K} \subseteq \mathbb{R}^d$. For example, the notion of *barycentric spanners* which were introduced in the seminal work of Awerbuch and Kleinberg (Awerbuch and Kleinberg, 2008) corresponds to looking at the ellipsoid of maximum volume supported by exactly d points from [3] $\mathcal{K}$. Barycentric Spanners have since been used as an exploration basis in several works: Online bandit linear optimization (Dani et al., 2007), A high probability counterpart of online bandit linear optimization (Bartlett et al., 2008), repeated decision making of approximable functions (Kakade et al., 2009) and a stochastic version of bandit linear optimization (Dani et al., 2008). Another example is the work of Bubeck et al. (Bubeck et al., 2012) which looks at the minimum volume enclosing ellipsoid (MVEE) also known as the John ellipsoid (see Section 2 for more background on this fundamental object from convex geometry).

As will be clear soon, our definition of a *minimal volumetric ellipsoid* enjoys the best properties of the examples above enabling us to get more efficient algorithms. Similar to

3. While the definition of (Awerbuch and Kleinberg, 2008) is not phrased as such, their analysis shows the existence of barycentric spanners by looking at the maximum volume ellipsoid.

barycentric spanners, it is supported by a small (linear) set of points of $\mathcal{K}$. Simultaneously and unlike the barycentric counterpart, the volumetric ellipsoid contains the body $\mathcal{K}$, a property shared with the John ellipsoid.

Definition 1 (Volumetric Ellipsoids) *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a set in Euclidean space. For $S \subseteq \mathcal{K}$, we say that $\mathcal{E}(S)$ is a* volumetric ellipsoid *for $\mathcal{K}$ if it contains $\mathcal{K}$. We say that $\mathcal{E}(S)$ is a* minimal volumetric ellipsoid *if it is a containing ellipsoid defined by a set of minimal cardinality*

$$S \in \arg\min_{S' \in \mathcal{S}(\mathcal{K})} |S'|, \qquad \mathcal{S}(\mathcal{K}) = \left\{ S' \mid S' \subseteq \mathcal{K} \subseteq \mathcal{E}(S') \right\}$$

We say that $|S|$ is the order *of the minimal volumetric ellipsoid or of* [4] *$\mathcal{K}$ denoted* $\mathbf{order}(\mathcal{K})$.

We discuss various geometric properties of volumetric ellipsoids later. For now, we focus on their utility in designing efficient exploration bases. To make this concrete and to simplify some terminology later on (and also to draw an analogy to barycentric spanners), we introduce the notion of *volumetric spanners*. Informally, these correspond to sets S that span all points in a given set with coefficients having Euclidean norm at most one. Formally:

Definition 2 *Let $\mathcal{K} \subseteq \mathbb{R}^d$ and let $S \subseteq \mathcal{K}$. We say that S is a* volumetric spanner *for $\mathcal{K}$ if $\mathcal{K} \subseteq \mathcal{E}(S)$.*

Clearly, a set $\mathcal{K}$ has a volumetric spanner of cardinality t if and only if $\mathbf{order}(\mathcal{K}) \leq t$.

Our goal in this work is to bound the order of arbitrary sets. A priori, it is not even clear if there is a universal bound (depending only on the dimension and not on the set) on the order S for compact sets $\mathcal{K}$. However, barycentric spanners and the John ellipsoid show that the order of any compact set in $\mathcal{R}^d$ is at most $O(d^2)$. Our main structural result in convex geometry gives a nearly optimal linear bound on the order of all sets.

Theorem 3 (Main) *Any compact set $\mathcal{K} \subseteq \mathbb{R}^d$ admits a volumetric ellipsoid of order at most $12d$. Further, if $\mathcal{K} = \{v_1, \ldots, v_n\}$ is a discrete set, then a volumetric ellipsoid for $\mathcal{K}$ of order at most $12d$ can be constructed in $\mathrm{poly}(n, d)$ time.*

This structural result is achieved by sparsifying (via the methods given in Batson et al. 2012) the John contact points of $\mathcal{K}$. We emphasize the last part of the above theorem giving an algorithm for finding volumetric spanners of small size; this could be useful in using our structural results for algorithmic purposes. We also give a different algorithmic construction for the discrete case (a set of n vectors) in Section 5. While being sub-optimal by logarithmic factors (gives an ellipsoid of order $O(d(\log d)(\log n))$ this alternate construction has the advantage of being simpler and more efficient to compute.

4. We note that our definition allows for multi-sets, meaning that S may contain the same vector more than once.

1.3 Approximate Volumetric Spanners

Theorem 3 shows the existence of good volumetric spanners and also gives an efficient algorithm for finding such a spanner in the discrete case, i.e., when $\mathcal{K}$ is finite and given explicitly. However, the existence proof uses the John ellipsoid in a fundamental way and it is not known how to compute (even approximately) the John ellipsoid efficiently for the case of general convex bodies. For such computationally difficult cases, we introduce a natural relaxation of volumetric ellipsoids which can be computed efficiently for a bigger class of bodies and is similarly useful. The relaxation comes from requiring that the ellipsoid of small support contain all but an ε fraction of the points in $\mathcal{K}$ (under some distribution). In addition, we also require that the measure of points decays exponentially fast w.r.t their $\mathcal{E}(S)$-norm (see precise definition in next section); this property gives us tighter control on the set of points not contained in the ellipsoid. When discussing a measure over the points of a body the most natural one is the uniform distribution over the body. However, it makes sense to consider other measures as well and our approximation results in fact hold for a wide class of distributions.

Definition 4 *Let $S \subseteq \mathcal{R}^d$ be a set of vectors and let V be the matrix whose columns are the vectors of S. We define the semi-norm*

$$\|x\|_{\mathcal{E}(S)} = \sqrt{x^\top (VV^\top)^{-1} x} \ ,$$

where $(VV^\top)^{-1}$ is the Moore-Penrose pseudo-inverse of $VV^\top$. Let $\mathcal{K}$ be a convex set in $\mathcal{R}^d$, p a distribution over it, and let $\varepsilon > 0$. A (p, ε)-exp-volumetric spanner *of $\mathcal{K}$ is a set $S \subseteq \mathcal{K}$ such that for any $\theta > 1$*

$$\Pr_{x \sim p}[\|x\|_{\mathcal{E}(S)} \geq \theta] \leq \varepsilon^{-\theta}.$$

To understand the intuition behind the above definition, notice that for any point x in $\mathcal{E}(S)$ we have $\|x\|_{\mathcal{E}(S)} \leq 1$. Hence, if S is such that $\mathcal{K} \subseteq \mathcal{E}(S)$ we have that S is a (p, ε)-*exp-volumetric spanner* of $\mathcal{K}$ for $\varepsilon = 0$ and any p. It follows that the above provides an approximate volumetric spanner; in what follows we show that this particular type of approximation can be computed efficeintly for log-concave p and is sufficient in certain cases.

Theorem 5 *Let $\mathcal{K}$ be a convex set in $\mathcal{R}^d$ and p a log-concave distribution over it. By sampling $O(d+\log^2(1/\varepsilon))$ i.i.d. points from p one obtains, w.p. at least $1-\exp\left(-\sqrt{\max\{\log(1/\varepsilon), d\}}\right)$, a (p, ε)-exp-volumetric spanner for $\mathcal{K}$.*

1.4 Structure of the paper

In the next section we list the preliminaries and known results from measure concentration, convex geometry and online learning that we need. In Section 3 we show the existence of small size volumetric spanners. In Sections 4 and 5 we give efficient constructions of volumetric spanners for continuous and discrete sets, respectively. We describe the application to experiment design in Section 6. We then proceed to describe the application of our geometric results to bandit linear optimization in Section 7.

2. Preliminaries

We now describe several preliminary results we need from convex geometry and linear algebra. We start with some notation:

- A matrix $A \in \mathbb{R}^{d\times d}$ is positive semi-definite (PSD) when for all $x \in \mathcal{R}^d$ it holds that $x^\top Ax \geq 0$. Alternatively, when all of its eigenvalues are non-negative. We say that $A \succeq B$ if $A - B$ is PSD.

- Given an ellipsoid $\mathcal{E}(S) = \{\sum_i \alpha_i v_i \,:\, \sum_i \alpha_i^2 \leq 1\}$, we shall use the notation $\|x\|_{\mathcal{E}(S)} \stackrel{\Delta}{=} \sqrt{x^\top (VV^\top)^{-1} x}$ to denote the (Minkowski) semi-norm defined by the ellipsoid, where V is the matrix with the vectors v_i's as columns. Notice that $\mathcal{E}(S)$ is symmetric and convex hence it defines a norm.

- Throughout, we denote by I_d the $d \times d$ identity matrix.

We next describe properties of the John ellipsoid which plays an important role in our proofs.

2.1 The Fritz John Ellipsoid

Let $\mathcal{K} \subseteq \mathbb{R}^n$ be an arbitrary convex body. Then, the **John ellipsoid** of $\mathcal{K}$ is the minimum volume ellipsoid containing $\mathcal{K}$. This ellipsoid is unique and its properties have been the subject of important study in convex geometry since the seminal work of John (John, 1948) (see Ball 1997 and Henk 2012 for historic information).

Suppose that we have linearly transformed $\mathcal{K}$ such that its minimum volume enclosing ellipsoid (MVEE) is the unit sphere; in convex geometric terms, $\mathcal{K}$ is in *John's position*. The celebrated decomposition theorem by (John, 1948) gives a characterization of when a body is in John's position and will play an important role in our constructions (the version here is from Ball 1997).

Below we consider only symmetric convex sets, i.e., sets that admit the following property: if $x \in \mathcal{K}$ then also $-x \in \mathcal{K}$. The sets encountered in machine learning applications are most always symmetric, since estimating a linear function on a point x is equivalent to estimating it on its polar $-x$, and negating the outcome.

Theorem 6 (Ball 1997) *Let $\mathcal{K} \in \mathcal{R}^d$ be a symmetric set such that its MVEE is the unit sphere. Then there exist $m \leq d(d+1)/2 - 1$* contact points *of $\mathcal{K}$ and the sphere $u_1, \ldots, u_m$ and non-negative weights $c_1, \ldots, c_m$ such that $\sum_i c_i u_i = 0$ and $\sum c_i u_i u_i^T = I_d$.*

The contact points of a convex body have been extensively studied in convex geometry and they also make for an appealing exploration basis in our context. Indeed, (Bubeck et al., 2012) use these contact points to attain an optimal-regret algorithm for BLO. Unfortunately we know of no efficient algorithm to compute, or even approximate, the John ellipsoid for a general convex set, thus the results by (Bubeck et al., 2012) do not yield efficient algorithms for BLO.

For our construction of volumetric spanners we need to compute the MVEE of a discrete symmetric set, which can be done efficiently. We make use of the following (folklore) result:

Theorem 7 (folklore, see e.g., Khachiyan 1996; Damla Ahipasaoglu et al. 2008) *Let $\mathcal{K} \subseteq \mathcal{R}^d$ be a set of n points. It is possible to compute an ε-approximate MVEE for $\mathcal{K}$ (an enclosing ellipsoid of volume at most $(1+\varepsilon)$ that of the MVEE) that is also supported on a subset of $\mathcal{K}$ in time $O(n^{3.5} \log \frac{1}{\varepsilon})$.*

The run-time above is attainable via the ellipsoid method or path-following interior point methods (see references in theorem statement). An approximation algorithm rather than an exact one is necessary in a real-valued computation model, and the logarithmic dependence on the approximation guarantee is as good as one can hope for in general.

The above theorem allows us to efficiently compute a linear transformation such that the MVEE of $\mathcal{K}$ is essentially the unit sphere. We can then use linear programming to compute an approximate decomposition like in John's theorem as follows.

Theorem 8 *Let $\{x_1, \ldots, x_n\} = \mathcal{K} \subseteq \mathcal{R}^d$ be a set of n points and assume that:*

1. *$\mathcal{K}$ is symmetric.*
2. *The John Ellipsoid of $\mathcal{K}$ is the unit sphere.*

Then it is possible, in $O((\sqrt{n}+d)n^3)$ time, to compute non-negative weights $c_1, \ldots, c_n$ such that (1) $\sum_i c_i x_i = 0$ and (2) $\sum_{i=1}^n c_i x_i x_i^\top = I_d$.

Proof

Denote the MVEE of $\mathcal{K}$ by $\mathcal{E}$ and let V be its corresponding $d \times d$ matrix, meaning V is such that $\|y\|_{\mathcal{E}}^2 = y^\top V^{-1} y \leq 1$ for all $y \in \mathcal{K}$. By our assumptions $I_d = V$.

As $\mathcal{K}$ is symmetric and its MVEE is the unit ball, according to Theorem 6, there exist $m \leq d(d+1)/2 - 1$ contact points $u_1, \ldots, u_m$ of $\mathcal{K}$ with the unit ball and a vector $c' \in \mathcal{R}^m$ such that $c' \geq 0$, $\sum c'_i = d$ and $\sum c'_i u_i u_i^T = I_d$. It follows that the following LP has a feasible solution: Find $c \in \mathcal{R}^n$ such that $c \geq 0$, $\sum c_i \leq d$ and $\sum c_i u_i u_i^T = I_d$. The described LP has $O(n + d^2)$ constraints and n variables. It can thus be solved in time $O(d + \sqrt{n})n^3)$ via interior point methods. ■

We next state the results from probability theory that we need.

2.2 Distributions and Measure Concentration

For a set $\mathcal{K}$, let $x \sim \mathcal{K}$ denote a uniformly random vector from $\mathcal{K}$.

Definition 9 *A distribution over $\mathcal{R}^d$ is log-concave if its probability distribution function (pdf) p is such that for all $x, y \in \mathcal{R}^d$ and $\lambda \in [0,1]$,*

$$p(\lambda x + (1-\lambda)y) \geq p(x)^\lambda p(y)^{1-\lambda}$$

Two log-concave distributions of interest to us are (1) the uniform distribution over a convex body and (2) a distribution over a convex body where $p(x) \propto \exp(L^\top x)$, where L is some vector in $\mathcal{R}^d$. The following result shows that given oracle access to the pdf of a log-concave distribution we can sample from it efficiently. An oracle to a pdf accepts as input a point $x \in \mathcal{R}^d$ and returns the value $p(x)$.

Lemma 10 (Lovász and Vempala 2007, Theorems 2.1 and 2.2) *Let p be a log-concave distribution over $\mathbb{R}^d$ and let $\delta > 0$. Then, given oracle access to p, i.e., and oracle computing its pdf for any point in $\mathbb{R}^d$, there is an algorithm which approximately samples from p such that:*

1. *The total variation distance between the produced distribution and the distribution defined by p is no more than δ. That is, the difference between the probabilities of any event in the produced and actual distribution is bounded by δ.*
2. *The algorithm requires a pre-processing time of $\tilde{O}(d^5)$.*
3. *After pre-processing, each sample can be produced in time $\tilde{O}(d^4/\delta^4)$, or amortized time of $\tilde{O}(d^3/\delta^4)$ if more than d samples are needed.*

Definition 11 (Isotropic position) *A random variable x is said to be in isotropic position*
(or isotropic) if

$$\mathbf{E}[x] = 0, \qquad \mathbf{E}[xx^\top] = I_d.$$

A set $\mathcal{K} \subseteq \mathbb{R}^d$ is said to be in isotropic position if $x \sim \mathcal{K}$ is isotropic. Similarly, a distribution p is isotropic if $x \sim p$ is isotropic.

Henceforth we use several results regarding the concentration of log-concave isotropic random vectors. In these results we use the matrix operator norm (i.e., spectral norm) defined as $\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$. We use slight modification where the center of the distribution is not necessarily in the origin. For completeness we present the proof of the modified theorems in Appendix A

Theorem 12 (Theorem 4.1 in Adamczak et al. 2010) *Let p be a log-concave distribution over $\mathcal{R}^d$ in isotropic position. There is a constant C such that for all $t, \delta > 0$, the following holds for $n = \frac{Ct^4 d \log^2(t/\delta)}{\delta^2}$. For independent random vectors $x_1, \ldots, x_n \sim p$, with probability at least $1 - \exp(-t\sqrt{d})$,*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top - I_d \right\| \leq \delta.$$

Corollary 13 *Let p be a log-concave distribution over $\mathcal{R}^d$ and $x \sim p$. Assume that x is such that $\mathbf{E}[xx^T] = I_d$. Then, there is a constant C such that for all $t \geq 1, \delta > 0$, the following holds for $n = \frac{Ct^4 d \log^2(t/\delta)}{\delta^2}$. For independent random vectors $x_1, \ldots, x_n \sim p$, with probability at least $1 - \exp(-t\sqrt{d})$,*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top - I_d \right\| \leq \delta.$$

Theorem 14 (Theorem 1.1 in Guédon and Milman 2011) *There exist constants c, C such that the following holds. Let p be a log-concave distribution over $\mathcal{R}^d$ in isotropic position and let $x \sim p$. Then, for all $\theta \geq 0$,*

$$Pr\left[\left| \|x\| - \sqrt{d} \right| > \theta\sqrt{d} \right] \leq C \exp(-c\sqrt{d} \cdot \min(\theta^3, \theta)).$$

Corollary 15 *Let p be a log-concave distribution over $\mathcal{R}^n$ and let $x \sim p$. Assume that $\mathbf{E}[xx^T] = I_d$. Then for some universal C, c it holds for any $\theta \geq 3$ that*

$$\Pr\left[\|x\| > \theta\sqrt{d}\right] \leq C\exp\left(-c\theta\sqrt{d}\right)$$

The following theorem provides a concentration bound for random vectors originating from an arbitrary distribution.

Theorem 16 (Rudelson 1999) *Let X be a vector-valued random variable over $\mathbb{R}^d$ with $\mathbf{E}[XX^\top] = \Sigma$ and $\|\Sigma^{-1/2}X\|^2 \leq R$. Then, for independent samples $X_1, \ldots, X_M$ from X, and*
$M \geq CR\log(R/\varepsilon)/\epsilon^2$ the following holds with probability at least $1/2$:

$$\left\|\frac{1}{M}\sum_{i=1}^{M} X_i X_i^\top - \Sigma\right\| \leq \epsilon\|\Sigma\|.$$

Finally, we also make use of barycentric spanners in our application to BLO and we briefly describe them next.

2.3 Barycentric Spanners

Definition 17 (Awerbuch and Kleinberg 2008) *A barycentric spanner of $\mathcal{K} \subseteq \mathcal{R}^d$ is a set of d points $S = \{u_1, \ldots, u_d\} \subseteq \mathcal{K}$ such that any point in $\mathcal{K}$ may be expressed as a linear combination of the elements of S using coefficients in $[-1, 1]$. For $C > 1$, S is a C-approximate barycentric spanner of $\mathcal{K}$ if any point in $\mathcal{K}$ may be expressed as a linear combination of the elements of S using coefficients in $[-C, C]$*

In (Awerbuch and Kleinberg, 2008) it is shown that any compact set has a barycentric spanner. Moreover, they show that given an oracle with the ability to solve linear optimization problems over $\mathcal{K}$, an approximate barycentric spanner can be efficiently obtained. In the following sections we will use this constructive result.

Theorem 18 (Proposition 2.5 in Awerbuch and Kleinberg 2008) *Let $\mathcal{K}$ be a compact set in $\mathcal{R}^d$ that is not contained in any proper linear subspace. Given an oracle for optimizing linear functions over $\mathcal{K}$, for any $C > 1$, it is possible to compute a C-approximate barycentric spanner for $\mathcal{K}$, using $O(d^2 \log_C(d))$ calls to the optimization oracle.*

3. Existence of Volumetric Ellipsoids and Spanners

In this section we show the existence of low order volumetric ellipsoids proving our main structural result, Theorem 3. Before we do so, we first state a few simple properties of volumetric ellipsoids (recall the Definition of **order** from Definition 1):

- The definition of **order** is linear invariant: for any invertible linear transformation $T : \mathbb{R}^d \to \mathbb{R}^d$ and $K \subseteq \mathbb{R}^d$, $\mathbf{order}(\mathcal{K}) = \mathbf{order}(T\mathcal{K})$.

  Proof Let $S \subseteq \mathcal{K}$ be such that $\mathcal{K} \subseteq \mathcal{E}(S)$. Then, clearly $T\mathcal{K} \subseteq \mathcal{E}(TS)$. Thus, $\mathbf{order}(T\mathcal{K}) \leq \mathbf{order}(\mathcal{K})$. The same argument applied to T^{-1} and $T\mathcal{K}$ shows that

Figure 1: In $\mathbb{R}^2$ the order of the volumetric ellipsoid of the equilateral triangle centered at the origin is at least 3. If the vertices are $[0,1],[-\frac{\sqrt{3}}{2},-\frac{1}{2}],[\frac{\sqrt{3}}{2},-\frac{1}{2}]$, then the eigenpoles of the ellipsoid of the bottom two vertices are $[0.\frac{2}{3}],[2,0]$. The second figure shows one possibility for a volumetric ellipsoid by adding $\frac{3}{4}$ of the first vertex to the previous ellipsoid. This shows the ellipsoid to be non-unique, as it can be rotated three ways.

$\mathbf{order}(\mathcal{K}) \leq \mathbf{order}(T\mathcal{K})$. ■

- The minimal volumetric ellipsoid is not unique in general; see example in Figure 1. Further, it is in general different from the John ellipsoid.

- For non-degenerate bodies $\mathcal{K}$, their order is naturally lower bounded by d, and there are examples in which it is strictly larger than d (e.g., Figure 1).

In the proof Theorem 3 we require a modification of a result by (Batson et al., 2012) providing a method to sparsity a distribution over vectors while approximately maintaining its covariance matrix. We show that this technique can be applied over the distribution derived from John's decomposition of $\mathcal{K}$ in order to obtain a small set from which we construct a volumetric spanner. We begin by presenting the result given in (Batson et al., 2012), then its modification, and then proceed to the proof of Theorem 3.

Theorem 19 (Theorem 3.1 of Batson et al. 2012) *Let $v_1,\ldots,v_m$ be vectors in $\mathcal{R}^d$ and let $c > 1$. Assume that $\sum_i v_i v_i^\top = I_d$. There exist scalars $s_i \geq 0$ for $i \in [m]$ with $|\{i|s_i > 0\}| \leq cd$ such that*

$$I_d \preceq \sum_i s_i v_i v_i^\top \preceq \frac{c+1+2\sqrt{c}}{c+1-2\sqrt{c}} I_d$$

Furthermore, these scalars can be found in time $O(cd^3m)$.

Lemma 20 *Let $u_1,\ldots,u_m$ be unit vectors and let $p \in \Delta(m)$ be a distribution over $[m]$ such that $d\sum_{i=1}^m p_i u_i u_i^\top = I_d$. Then, there exists a (possibly multi) set $S \subseteq \{u_1,\ldots,u_m\}$ such that $\sum_{v\in S} vv^\top \succeq I_d$ and $|S| \leq 12d$. Moreover, such a set can be computed in time $O(md^3)$.*

Proof Let $v_i = \sqrt{p_i}u_i$. We fix c as some constant whose value will be determined later. Clearly for these vectors v_i it holds that $d\sum_i v_i v_i^\top = I_d$. It follows from the above theorem that there exist some scalars $s_i \geq 0$ for which at most cd are non-zeros and

$$I_d \preceq d\sum_i s_i v_i v_i^\top \preceq \frac{c+1+2\sqrt{c}}{c+1-2\sqrt{c}} I_d \tag{1}$$

Our set S will be composed by taking each u_i $\lceil ds_i p_i \rceil$ many times. Plugging in equation (1) shows that indeed

$$\sum_{w\in S} ww^\top \succeq I_d$$

and it remains to bound the size of S i.e., $\sum \lceil ds_i p_i \rceil$. By taking the trace of the expression and dividing by d we get that

$$\sum_i s_i \text{Trace}(v_i v_i^\top) \preceq \frac{c+1+2\sqrt{c}}{c+1-2\sqrt{c}}$$

Plugging in the expressions for v_i along with u_i being unit vectors (hence $\text{Trace}(u_i u_i^\top) = 1$) lead to

$$\sum_i s_i p_i \preceq \frac{c+1+2\sqrt{c}}{c+1-2\sqrt{c}}$$

It follows that

$$\sum \lceil ds_i p_i \rceil \leq d\frac{c+1+2\sqrt{c}}{c+1-2\sqrt{c}} + |\{i|s_i \geq 0\}| \leq d\left(c + \frac{c+1+2\sqrt{c}}{c+1-2\sqrt{c}}\right)$$

By optimizing c we get $\sum \lceil ds_i p_i \rceil \leq 12d$, and the lemma is proved. ■

Proof of Theorem 3 Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a compact set. Without loss of generality assume that $\mathcal{K}$ is symmetric and contains the origin; we can do so as in the following we only look at outer products of the form $vv^\top$ for vectors $v \in \mathcal{K}$. Further, as $\textbf{order}(\mathcal{K})$ is invariant under linear transformations, we can compute, as detailed in Theorem 7 the MVEE[5] of $\mathcal{K}$ and transform the space into one where this ellipsoid is the Euclidean unit sphere. That is, move $\mathcal{K}$ into John's position, at $\text{poly}(n, d)$ time.

According to Theorem 8 it is then possible to compute a distribution p over $\mathcal{K}$ with

$$d\sum_{x\in\mathcal{K}} p_x xx^\top = I_d$$

Lemma 20 provides a way to compute a (multi-)set S of size at most $12d$ with

$$\sum_{v\in S} vv^\top \succeq I_d$$

5. Notice that the Theorem provides an approximation of $1+\varepsilon$ of the MVEE where the running time scales as $\log(1/\varepsilon)$. In what follows it is easy to see that the precision of all equalities is affected up to a multiplicative factor of $1 \pm \varepsilon$ by this issue. This eventually translates into a set of size $12(1+\varepsilon)d$ rather than $12d$. We omit these technical details for a more readable proof.

Since $\mathcal{K}$ is contained in the unit ball we get that S is a volumetric spanner for $\mathcal{K}$ as required. The total running time required to find s includes the computation of the transformation into John's position, John's decomposition, and its sparsification. The running time amounts to $O(n^{3.5} + dn^3 + nd^3)$. ■

4. Approximate Volumetric Spanners

In this section we provide a construction for (p, ε)-exp-volumetric spanner (as in Definition 4), proving Theorem 5. We start by providing a more technical definition of a spanner. Note that unlike previous definitions, the following is not impervious to linear operators and will only be used to aid our construction.

Definition 21 *A β-relative-spanner is a discrete subset $S \subseteq \mathcal{K}$ such that for all $x \in \mathcal{K}$, $\|x\|^2_{\mathcal{E}(S)} \leq \beta \|x\|^2$.*

A first step is a spectral characterization of relative spanners:

Lemma 22 *Let $S = \{v_1, ..., v_T\} \subseteq \mathcal{K}$ span $\mathcal{K}$ and be such that*

$$W = \sum_{i=1}^{T} v_i v_i^\top \succeq \frac{1}{\beta} I_d$$

Then S is a β-relative-spanner.

Proof Let $V \in \mathbb{R}^{d \times T}$ be a matrix whose columns are the vectors of S. As $VV^\top = W \succeq \frac{1}{\beta} I_d$ we have that

$$\beta I_d \succeq (VV^\top)^{-1}$$

It follows that

$$\|x\|_{\mathcal{E}(S)} = x^\top (VV^\top)^{-1} x \leq \beta \|x\|^2$$

as required. ■

Proof [Proof of Theorem 5] We analyze the algorithm of sampling i.i.d points according to p, previously defined within Theorem 5, assuming the vectors are sampled exactly according to the log-concave distribution. The result involving an approximate sample, which is necessary for implementing the algorithm in the general case, is an immediate application of Lemma 10 and Corollary 13.

Our analysis of the algorithm is for $T = C(d + \log^2(1/\varepsilon))$ samples, where C is some sufficiently large constant. Assume first that $\mathbf{E}_{x \sim p}[xx^\top] = I_d$. Let $W = \sum_{i=1}^{T} u_i u_i^\top$. Then, for $C > 0$ large enough, by Corollary 13, $\|\frac{1}{T} W - I_d\| \leq 1/2$ w.p. at least $1 - \exp(-\sqrt{d})$. Therefore, S spans $\mathbb{R}^d$ and

$$\frac{1}{T} W \succeq I_d - \frac{1}{2} I_d = \frac{1}{2} I_d$$

Thus according to Lemma 22, S is a $(2/T)$-relative spanner. Consider the case where $\Sigma = \mathbf{E}_{x\sim p}[xx^\top]$ is not necessarily the identity. By the above analysis we get that

$$\Sigma^{-1/2}S = \{\Sigma^{-1/2}u_1, \ldots, \Sigma^{-1/2}u_T\}$$

form a $(2/T)$-relative spanner for $\Sigma^{-1/2}\mathcal{K}$. This is since the r.v defined as $\Sigma^{-1/2}x$ where $x \sim p$ is log-concave. The latter along with Corollary 15 implies that for any $\theta \geq 1$,

$$\Pr_{x\sim p}\left[\|\Sigma^{-1/2}x\| \geq 3\theta\sqrt{d}\right] \leq c_1 \exp\left(-c_2\theta\sqrt{d}\right) \tag{2}$$

for some universal constants $c_1, c_2 > 0$. It follows that for our set S and any $\theta \geq 1$,

$$\begin{aligned}
\Pr_{x\sim p}\left[\|x\|_{\mathcal{E}(S)} > \theta\right] &= \Pr_{x\sim p}\left[\|\Sigma^{-1/2}x\|_{\mathcal{E}(\Sigma^{-1/2}S)} > \theta\right] && \|x\|_{\mathcal{E}(S)} = \|\Sigma^{-1/2}x\|_{\mathcal{E}(\Sigma^{-1/2}S)} \\
&\leq \Pr_{x\sim p}\left[\|\Sigma^{-1/2}x\| > \theta\sqrt{T/2}\right] && \Sigma^{-1/2}S \text{ is a } 2/T\text{-relative-spanner} \\
&= \Pr_{x\sim p}\left[\|\Sigma^{-1/2}x\| > 3\theta\sqrt{\tfrac{dC}{18}}\cdot\sqrt{1+\tfrac{\log^2(1/\varepsilon)}{d}}\right] && T = C(d+\log^2(1/\varepsilon)) \\
&\leq c_1\exp\left(-c_2\theta\sqrt{d}\sqrt{\tfrac{C}{18}}\cdot\sqrt{1+\tfrac{\log^2(1/\varepsilon)}{d}}\right) && \text{Equation (2), } C \geq 18 \\
&\leq \exp\left(-\theta\sqrt{d+\log^2(1/\varepsilon)}\right) && C \text{ sufficiently large} \\
&\leq \varepsilon^{-\theta}
\end{aligned}$$

■

In our application of volumetric spanners to BLO, we also need the following relaxation of volumetric spanners where we allow ourselves the flexibility to scale the ellipsoid:

Definition 23 *A ρ-ratio-volumetric spanner S of $\mathcal{K}$ is a subset $S \subseteq \mathcal{K}$ such that for all $x \in \mathcal{K}$, $\|x\|_{\mathcal{E}(S)} \leq \rho$.*

One example for such an approximate spanner with $\rho = \sqrt{d}$ is a barycentric spanner (Definition 17). In fact, it is easy to see that a C-approximate barycentric spanner is a $C\sqrt{d}$-ratio-volumetric spanner . The following is immediate from Theorem 18.

Corollary 24 *Let $\mathcal{K}$ be a compact set in $\mathcal{R}^d$ that is not contained in any proper linear subspace. Given an oracle for optimizing linear functions over $\mathcal{K}$, for any $C > 1$, it is possible to compute a $C\sqrt{d}$-ratio-volumetric spanner S of $\mathcal{K}$ of cardinality $|S| = d$, using $O(d^2 \log_C(d))$ calls to the optimization oracle.*

5. Fast Volumetric Spanners for Discrete Sets

In this section we describe a different algorithm that constructs volumetric spanners for discrete sets. The order of the spanners we construct here is suboptimal (in particular, there is a dependence on the size of the set $\mathcal{K}$ which we didn't have before). However, the algorithm is particularly simple and efficient to implement (takes time linear in the size of the set).

Algorithm 1 Fast Volumetric Spanner construction

1: Input $\mathcal{K} = \{x_1, ..., x_n\} \subseteq \mathbb{R}^d$, $C \in \mathbb{R}$.
2: Set $T = \emptyset$
3: **while** $|K| > Cd \log d$ **do**
4: Compute $\Sigma = \sum_i x_i x_i^\top$ and let $u_i = \Sigma^{-1/2} x_i$, $p_i = 1/2n + \|u_i\|^2/2d$.
5: Set $S \leftarrow \emptyset$.
6: **for** $i = 1, ..., Cd \log d$ **do**
7: sample with replacement from $[n]$ according to $p_1, \ldots, p_n$: $S \leftarrow S \cup \{i\}$ w.p. p_i
8: **end for**
9: **if** $|\{i \ , \ \|x_i\|_{\mathcal{E}(S)} \leq 1\}| < \frac{n}{2}$ **then**
10: Set $S \leftarrow \emptyset$ and GOTO step (6).
11: **end if**
12: Set $T \leftarrow T \cup S$
13: Set the remainder as $\mathcal{K} \leftarrow \{x_i \text{ s.t. } \|x_i\|_{\mathcal{E}(S)} > 1\}$
14: **end while**
15: **return** $T \leftarrow T \cup \mathcal{K}$

Theorem 25 *Given a set of vectors $\mathcal{K} = \{x_1, \ldots, x_n\} \in \mathbb{R}^d$, Algorithm 1 outputs a volumetric spanner of size $O((d \log d)(\log n))$ and has an expected running time of $O(nd^2)$.*

Proof Consider a single iteration of the algorithm with input $v_1, \ldots, v_n \in \mathbb{R}^d$. We claim that the random set S obtained in Step 7 satisfies the following condition with constant probability:

$$\Pr_{x \in \mathcal{K}} \left[\|x\|_{\mathcal{E}(S)} \leq 1 \right] \geq 1/2 \tag{3}$$

Suppose the above statement is true. Then, the lemma follows easily as it implies that for the next iteration there are fewer than $n/2$ vectors. Hence, after $(\log n)$ recursive calls we will have a volumetric spanner. The total size of the set will be $O((d \log d)(\log n))$. To see the time complexity, consider a single run of the algorithm. The most computationally intensive steps are computing Σ and $\Sigma^{-1/2}$ which take time $O(nd^2)$ and $O(d^3)$ respectively. We also need to compute $(\sum_{v \in S} vv^\top)^{-1}$ (to compute the $\mathcal{E}(S)$ norm) which takes time $O(d^3 \log d)$, and compute the $\mathcal{E}(S)$ norm of all the vectors which requires $O(nd^2)$. As $n = \Omega(d \log(d))$, it follows that a single iteration runs of a total expected time of $O(nd^2)$. Since the size of n is split in half between iterations, the claim follows.

We now prove that Equation 3 holds with constant probability

$$x_j^\top \left(\sum_{v \in S} vv^\top \right)^{-1} x_j = u_j^\top \left(\sum_{v \in S'} vv^\top \right)^{-1} u_j. \tag{4}$$

where $S' = \{\Sigma^{-1/2}v | v \in S\}$ is the (linearly) shifted version of S. Therefore, it suffices to show that with sufficiently high probability, the right hand side of the above equation is bounded by 1 for at least $n/2$ indices $j \in [n]$.

Note that $p_i = 1/2n + \|u_i\|^2/2d$ form a probability distribution: $\sum_i p_i = 1/2 + (\sum_i \|u_i\|^2)/2d = 1$. Let $X \in \mathbb{R}^d$ be a random variable with $X = u_i/\sqrt{p_i}$ with probability p_i for $i \in [n]$. Then, $\mathbf{E}[XX^\top] = I_d$. Further, for any $i \in [n]$

$$\|u_i\|^2/p_i \leq 2d.$$

Therefore, by Theorem 16, if we take $M = Cd(\log d)$ samples $X_1, \ldots, X_M$ for C sufficiently large, then with probability of at least $1/2$, it holds that

$$\sum_{i=1}^{M} X_i X_i^\top \succeq (M/2) I_d.$$

Let $T \subseteq [n]$ be the multiset corresponding to the indices of the sampled vectors $X_1, \ldots, X_M$. The above inequality implies that

$$\sum_{i \in T} \frac{1}{p_i} u_i u_i^\top \succeq (M/2) I_d.$$

Now,

$$\sum_{v \in S'} vv^\top \succeq (\min_i p_i) \sum_{v \in S'} \frac{1}{p_i} vv^\top \succeq (\min_i p_i)(M/2) I_d \succeq (M/4n) I_d.$$

Therefore,

$$\begin{aligned}
\sum_{i=1}^{n} u_i^\top \left(\sum_{v \in S'} vv^\top \right)^{-1} u_i &= \sum_{i=1}^{n} \operatorname{Tr}\left(\left(\sum_{v \in S'} vv^\top \right)^{-1} \left(u_i u_i^\top \right) \right) \\
&= \operatorname{Tr}\left(\left(\sum_{v \in S'} vv^\top \right)^{-1} \left(\sum_{i=1}^{n} u_i u_i^\top \right) \right) \\
&= \operatorname{Tr}\left(\left(\sum_{v \in S'} vv^\top \right)^{-1} \right) \\
&\leq \frac{4nd}{M} \leq \frac{4n}{C \log d} \leq \frac{n}{2 \log d},
\end{aligned}$$

for C sufficiently large. Therefore, by Markov's inequality and Equation 4, it follows that Equation 3 holds with high probability. The theorem now follows. ■

6. Experiment design using volumetric spanners

The active linear regression (ALR) problem is formally defined as follows. The input is a pool of n data points $\mathcal{K} = \{x_1, \ldots, x_n\} \subseteq \mathcal{R}^d$, a tolerable error $\varepsilon > 0$ and a confidence

parameter δ. A *query* to a point x returns an unbiased noisy estimate of $\langle w^*, x\rangle$ for some unknown vector $w^* \in \mathcal{R}^d$, with variance bounded by 1. Our objective is to actively choose points to query, and based on these queries, obtain a vector $w \in \mathcal{R}^d$ that with probability at least $1-\delta$ has a *max-error* of

$$\max_{x\in\mathcal{K}} |\langle w^*, x\rangle - \langle w, x\rangle| < \varepsilon$$

We aim to minimize the *query complexity* of the solution, which is the mean number of queries required by the process. To avoid tedious definitions and details we assume that

1. all of the data points $\{x_i\}$ and w^* are bounded in the Euclidean unit sphere
2. the noise is bounded in absolute value by one with probability one.

Our results can be extended to remove the first assumption, simply by allowing an additional parameter which is the radius of the minimal enclosing Euclidian ball.

It is also possible to remove the second assumption, although this requires more robust estimators, e.g., the median-of-means estimator[6].

6.1 A lower bound for passive linear regression

In this section we provide an example for a set $\mathcal{K}$ where the passive learning algorithm must use $\Omega(\frac{n}{\varepsilon^2})$ observations to obtain a regressor w with additive error of at most ε on all of the data points. We start by formally defining the passive setup. Here, a query returns a random point x chosen uniformly from the set $\mathcal{K}$ and an unbiased noisy measurement of $\langle w^*, x\rangle$, with variance of at most 1. As before we assume that all points, including w^* are contained in the ℓ_2 unit sphere.

The set $\mathcal{K}$ in our example is defined in the following manner. Let $Y \subseteq \mathcal{R}^d$ be an arbitrary set of size $n-1$ such that for all $x \in Y$, $\langle x, e_1\rangle = 0$. Let $\mathcal{K} = Y \cup \{e_1\}$. Here, e_1 is the first vector in the standard basis for $\mathcal{R}^d$.

Theorem 26 *Any algorithm in the passive setting achieving an additive error of at most ε in all of the data points of $\mathcal{K}$ whose success probability is $1-\delta$ requires $\Omega(\log(1/\delta)n/\varepsilon^2)$ queries.*

The theorem is an immediate corollary of the following lemma.

Lemma 27 *For $\mathcal{K}$ defined above, any policy distinguishing between the case where $\langle w^*, e_1\rangle = -\varepsilon$ and $\langle w^*, e_1\rangle = \varepsilon$ with probability larger than $1-\delta$ must use $\Omega(n\log(1/\delta)/\varepsilon^2)$ queries.*

Proof We begin by mentioning that a query of a point x where $\langle x, e_1\rangle = 0$ provides no information to the sign of $\langle w^*, e_1\rangle$, hence does not help distinguish between the two hypotheses. The following lemma provides a lower bound to the number of queries at point e_1 required to estimate the $\langle w^*, e_1\rangle$ up to a sufficiently small additive error and with sufficient confidence. It is a folklore lemma in statistics and appears e.g., in (Mannor and Tsitsiklis, 2004) in a much more general form.

6. The median-of-means estimator is defined as follows: Given m queries to a single dimensional distribution, the estimator of its mean is not taken to be the average of the queries; the m queries are instead split into $\log(1/\delta)$ equal sized buckets, each bucket has its average computed and the estimator is the median of these averages.

Lemma 28 (Theorem 1 of Mannor and Tsitsiklis 2004) *Let $\mathcal{D}$ be a distribution over $[-1,1]$. Let $\varepsilon > 0$ be such that for $X \sim \mathcal{D}$, $|\mathbf{E}[X]| \geq \varepsilon$. Let T be the expected number of queries required by any algorithm that queries i.i.d copies of $X \sim \mathcal{D}$ until being able to distinguish, with probability at least $1-\delta$ between the cases $\mathbf{E}[X] \leq -\varepsilon$ and $\mathbf{E}[X] \geq \varepsilon$. Then for universal constants $\varepsilon_0 > 0$, $\delta_0 > 0$, c_1, c_2 it holds that if $\varepsilon < \varepsilon_0$ and $\delta < \delta_0$ then $T \geq \frac{c_1 \log(c_2/\delta)}{\varepsilon^2}$.*

It follows that the expected number of queries needed in order to distinguish between the two hypotheses with probability $\geq 1-\delta$ is at least $\frac{c_1 n \log(c_2/\delta)}{\varepsilon^2}$, as the probability of observing a query to the inner product with e_1 is $1/n$. ■

6.2 Our ALR solution

In this section prove the following.

Theorem 29 *There exists an algorithm to the ALR problem with success probability of at least*
$1-\delta$ *and with the following properties:*

1. *The algorithm requires a preprocessing stage for building a volumetric spanner over $\mathcal{K}$*
2. *It's running time (after preprocessing) is $\tilde{O}\left(\frac{nd\log(1/\delta)}{\varepsilon^2}\right)$*
3. *It's query complexity is at most $O\left(\frac{d\log(n)\log(1/\delta)}{\varepsilon^2}\right)$.*

The intuition behind the algorithm is the following. We begin with a preprocessing stage of computing a volumetric spanner S for the set of points $\mathcal{K}$. Given this spanner we can implement a procedure that outputs, for all of the points of $\mathcal{K}$ simultaneously, an unbiased estimator of $\langle w^*, x\rangle$ with variance of at most $|S|$. To demonstrate the usefulness of this estimator, consider averaging $|S|\log(n/\delta)/\varepsilon^2$ i.i.d outputs of S. Standard concentration bound show that w.p at least $1-\delta$ the estimates of all points in $\mathcal{K}$ are correct up to an additive error of ε. Rather than computing a noisy output for w^* on the points and recovering a hypothesis w from that we use a technique by (Clarkson et al., 2012) that given an oracle for a function over a set of data points constructs a hypothesis w using a small number of queries to the oracle.

6.2.1 Constructing a low variance estimator

The main component of our method is a black box providing a noisy estimate of $\langle x, w^*\rangle$ to all of the data point of $\mathcal{K}$ simultaneously. The intuition behind the algorithm we describe next, is that given sufficiently many queries to the noisy estimator, a union bound argument can ensure an accurate estimate in all of the data points simultaneously.

We begin with the description of this black box providing the estimates. The main tool used for this 'all-point-estimator' is a volumetric spanner for the set $\mathcal{K}$. Algorithm 2 provides the formal description of how a volumetric spanner can be used to obtain these estimates.

The following lemma provides the analysis of Algorithm 2.

Algorithm 2 Sample($\mathcal{K}$)

1: Input: set $\mathcal{K} = \{x_1, ..., x_n\}$, Volumetric spanner for $\mathcal{K}$ denoted S, and measurement oracle that given x returns an unbiased estimator $\widehat{\langle x, w^* \rangle}$ with variance at most one for some fixed w^*.
2: Choose a point $v_j \in S$ uniformly at random, query its inner product $\hat{\ell} = \widehat{\langle v_j, w^* \rangle}$
3: let V be the $d \times |S|$ matrix whose columns are the elements of S, and let $V^\dagger \in \mathcal{R}^{|S| \times d}$ be its Moore-Penrose pseudo inverse.
4: For $x \in \mathcal{K}$, let $\alpha_x = V^\dagger x$ and let $\hat{\ell}_x \leftarrow (\alpha_x)_j \hat{\ell} \cdot |S|$
5: **return** estimates $\{\hat{\ell}_x\}_{x \in \mathcal{K}}$

Lemma 30 *Algorithm 2 queries a single point from $\mathcal{K}$. Its estimates have the properties of (1) being unbiased and (2) have a variance of at most $12d$. More formally, we have*

$$\forall x \in \mathcal{K} \ . \ \mathbf{E}[\hat{\ell}_x] = \langle x, w^* \rangle \ , \ \mathbf{Var}(\hat{\ell}_x) \leq |S| \leq 12d$$

Proof

$$\begin{aligned} \mathbf{E}\left[\hat{\ell}_x\right] &= \textstyle\sum_{j \in S} \Pr[v_j] \cdot (\alpha_x)_j \, \mathbf{E}\, \widehat{\langle v_j, w^* \rangle} \cdot |S| \\ &= \textstyle\sum_{j \in S} (\alpha_x)_j \, \mathbf{E}\, \widehat{\langle v_j, w^* \rangle} \\ &= (V^\dagger x)^T V^T w^* = \langle x, w^* \rangle \end{aligned}$$

For the variance, recall that $x \in \mathcal{K}$ and S is a volumetric spanner of $\mathcal{K}$ indicates that $\|\alpha_x\|_2 \leq 1$:

$$\begin{aligned} \mathbf{E}[\hat{\ell}_x^2] &= \textstyle\sum_{j \in S} \Pr[v_j] \cdot (\alpha_x)_j^2 \, \mathbf{E}\left[\widehat{\langle v_j, w^* \rangle}^2\right] \cdot |S|^2 \\ &\leq |S| \textstyle\sum_{j \in S} (\alpha_x)_j^2 \leq |S| \end{aligned}$$

By Theorem 3 we can efficiently construct volumetric spanners of size $|S| = 12d$. ■

6.2.2 Algorithm and its analysis

In this section we present an algorithm for the ALR problem, following the primal-dual paradigm as in (Clarkson et al., 2012), and specifically their meta algorithm 3. This latter meta-algorithm can be used to solve any convex constrained feasibility problem, of which our ALR problem is a special case. The idea is to apply a low regret algorithm to a distribution over the constraints. The distribution over the constraints is changed according to a multiplicative update rule. The specific meta-algorithm we apply uses random estimates of the constraints that enable faster running time. We proceed to spell out the details.

To avoid extraneous notions we will assume henceforth w.l.o.g that $\mathcal{K}$ is symmetric meaning that $x \in \mathcal{K}$ iff $-x \in \mathcal{K}$. This is without loss of generality since an unbiased estimator for $\langle -x, w^* \rangle$ is obtained by negating the estimator for $\langle x, w^* \rangle$.

We write the ALR problem as the following mathematical program:

$$\min_{\|w\|\leq 1} g(w) \quad s.t. \qquad g(w) = \max_{x\in\mathcal{K}} c_x(w) \tag{5}$$
$$c_x(w) = \langle x, w\rangle - \langle x, w^*\rangle$$

Note that by definition $g(w^*) = 0$, which is the optimal solution to the problem as $\mathcal{K}$ is symmetric. In addition, an ε approximate solution to the ALR instance, assuming $\|w^*\| \leq 1$, corresponds to a vector $\hat{w}$ with $g(\hat{w}) \leq \varepsilon$.

Algorithm 3 Primal-Dual Algorithm for ALR

Input: T
Let $w_1 \leftarrow 0$, $q_0 \leftarrow \mathbf{1}_n$, $\eta \leftarrow \frac{1}{100}\sqrt{\frac{\log(n)}{T}}$.
for $t = 1$ to T **do**
  Query **Sample**$(\mathcal{K})$ $12d$ times to obtain unit-variance zero-mean estimators $\tilde{a}_t(i)$ for all constraints c_i's:
$$\tilde{a}_t(i) \stackrel{\text{def}}{=} \langle x_i, w_t\rangle - \widehat{\langle x_i, w^*\rangle}$$
  where $\widehat{\langle x_i, w^*\rangle}$ is the estimate of $\langle x_i, w^*\rangle$ given by **Sample**$(\mathcal{K})$.
  for $i \in [n]$ **do**
    $a_t(i) \leftarrow \text{clip}(\tilde{a}_t(i), 1/\eta)$, where
$$\text{clip}(\alpha, \beta) = \begin{cases} \min\{\alpha, |\beta|\} & \alpha \geq 0 \\ \max\{-\alpha, -|\beta|\} & \alpha < 0 \end{cases}$$
    $q_t(i) \leftarrow q_{t-1}(i)(1 - \eta a_t(i) + \eta^2 a_t(i)^2)$
  end for
  Choose $i_t \in [n]$ at random with $\Pr[i_t = i] \propto q_t(i)$
  $w_t \leftarrow w_{t-1} - \frac{1}{\sqrt{t}}\nabla_w c_{i_t}$, where $\nabla_w c_{i_t} = x_{i_t}$
end for
return $\bar{w} = \frac{1}{T}\sum_t w_t$

The following theorem bounds the performance of Algorithm 3. It immediately follows from Lemma 4.1 in (Clarkson et al., 2012).

Theorem 31 *Algorithm 3 runs in time $\tilde{O}(\frac{dn}{\varepsilon^2})$ and requires $O(\frac{d\log n}{\varepsilon^2})$ queries to the procedure* **Sample**$(\mathcal{K})$. *It returns, with probability of at least $\frac{1}{2}$, a vector w such that* $\max_{x\in\mathcal{K}}\langle w - w^*, x\rangle \leq \varepsilon$.

Proof

Notice that Algorithm 3 is an instantiation of Alg 3 from (Clarkson et al., 2012) applied to mathematical Program 5 with the following arguments:

1. The primal decision set $\{\|w\| \leq 1\}$ and (linear) cost functions $c_x(w)$, admits an iterative low regret algorithm, namely online gradient descent, with expected regret

$\mathbf{E}[R(T)] \leq 2\sqrt{T}$. This follows since the norms of x, w (for all $x \in \mathcal{K}$) are bounded by one. See e.g., Theorem 1 by (Zinkevich, 2003).

The expression $T_\epsilon(LRA)$ (here LRA stands for Low Regret Algorithm, and not the Active Linear Regression problem we are addressing) refers to the number T such that the average regret of online gradient descent is at most ϵ, in our case, this is $\frac{4}{\epsilon^2}$.

2. Meta-algorithm 3 in (Clarkson et al., 2012) assumes an oracle to a procedure **Sample**($\mathcal{K}$) that returns a vector of length $|\mathcal{K}|$ whose entries are unbiased estimators of $\langle x, w^* \rangle$, for all $x \in \mathcal{K}$ whose variance is upper bounded by 1. Recall that such a procedure, with variance $12d$ rather than 1, was given in Section 6.2.1. By averaging $12d$ such samples we obtain unit-variance estimates.

Thus, Lemma 4.1 (Clarkson et al., 2012) implies that the algorithm returns w.p. $\frac{1}{2}$ an ϵ-approximate solution in number of iterations bounded by

$$\max\{T_\epsilon(LRA), \frac{\log n}{\epsilon^2}\} \leq \frac{\log n}{\epsilon^2}$$

Each iteration involves elementary operations that can be implemented in time $\tilde{O}(nd)$ and $O(d)$ queries to **Sample**.

■

6.2.3 Validation and high probability algorithm

Algorithm 3 provides a method to obtain an approximated solution to the ALR problem with probability 1/2. We now describe a method to amplify the success probability to $1-\delta$. The idea is to use a validation procedure and repeat the algorithm multiple times. It is easy to see that with a validation process, repeating Algorithm 3 for $O(\log(1/\delta))$ many times will increase the probability of success to $1 - \delta$.

We now describe a validation procedure that is in itself random, in the sense that it may err but its error probability is manageable. Algorithm 4 is given as input a hypothesis w and a ALR problem. It verifies, w.h.p., that w is an ε-approximate solution to the ALR problem accurate.

Algorithm 4 Verification

1: **Input:** Volumetric spanner S, parameters $\varepsilon, \delta > 0$, hypothesis $w \in \mathcal{R}^d$.
2: run **Sample**($\mathcal{K}$) $T = 2\ln(2n/\delta)|S|/\varepsilon^2$ times and obtain for each data point in $\mathcal{K}$, T i.i.d samples of $\langle w^*, x \rangle$
3: for each $x \in \mathcal{K}$ let $\tilde{f}(x)$ be the average of the above T samples.
4: declare w as accurate iff for all x, $|\langle w, x \rangle - \tilde{f}(x)| < 2\varepsilon$

Lemma 32 *Algorithm 4 has the following properties:*

- *It requires $O(\log(n/\delta)|S|/\varepsilon^2)$ queries to the oracle* **Sample**($\mathcal{K}$)

- *If the worst-case error of w is bounded by ε then w.p. at least $1-\delta$ it will be declared as accurate*
- *If the worst-case error of w is larger than 3ε then w.p. at least $1-\delta$ it will be declared as inaccurate.*

Proof We recall that the process **Sample**($\mathcal{K}$) returns unbiased estimates of $\langle w^*, x\rangle$ for all of the data points where the estimates are bounded in absolute value by $|S|$. Fix some point $x \in \mathcal{K}$. Chernoff's inequality dictates that

$$\Pr\left[\left|\langle w^*, x\rangle - \tilde{f}(x)\right| > \varepsilon\right] \leq 2\exp\left(-\frac{\varepsilon^2}{2|S|}\cdot\frac{2\ln(2n/\delta)|S|}{\varepsilon^2}\right) = \delta/n$$

Hence, w.p at least $1-\delta$ it holds for all $x \in \mathcal{K}$ simultaneously that $|\langle w^*, x\rangle - \tilde{f}(x)| < \varepsilon$. The claim immediately follows by using the triangle inequality in order to bound $|\langle w, x\rangle - \tilde{f}(x)|$. ■

Corollary 33 *There exists an algorithm that runs in time $\tilde{O}\left(\frac{dn\log\frac{1}{\delta}}{\varepsilon^2}\right)$ and returns, with probability of at least $1-\delta$, a vector w such that $\max_{x\in\mathcal{K}}\langle w - w^*, x\rangle \leq \varepsilon$.*

Proof Given the parameter δ, we run $\log(1/2\delta)$ independent copies of Algorithm 3 with parameter $\varepsilon' = \varepsilon/3$. Each such copy will produce a hypothesis w. We check for each such hypothesis w whether it is $\varepsilon' = \varepsilon/3$ accurate using Algorithm 4, with success probability $1-\delta/2\log(1/2\delta)$. With probability $1/2\delta$, all of the occurrences of the validation procedures will not err. Also, with probability at least $1-2\delta$ at least one hypothesis will be sufficiently accurate. Hence, by union bound we have with probability at least $1-\delta$ that at least one hypothesis will be declared accurate and any of the hypotheses declared accurate will be at least $3\varepsilon' = \varepsilon$ accurate. This concludes the quality of the output of the algorithm. The running time analysis is trivial. ■

7. Bandit Linear Optimization

Recall the problem of Bandit Linear Optimization (BLO): iteratively at each time sequence t, the environment chooses a loss vector L_t that is not revealed to the player. The player chooses a vector $x_t \in \mathcal{K}$ where $\mathcal{K} \subseteq \mathbb{R}^d$ is convex, and once she commits to her choice, the loss $\ell_t = x_t^\top L_t$ is revealed. The objective is to minimize the loss and specifically, the regret, defined as the strategy's loss minus the loss of the best fixed strategy of choosing some $x^* \in \mathcal{K}$ for all t. We henceforth assume that the loss vectors L_t's are chosen from the polar of $\mathcal{K}$, meaning from $\{L : |L^\top x| \leq 1\ \forall x \in \mathcal{K}\}$. In particular this means that the losses are bounded in absolute value, although a different choice of assumption (i.e., ℓ_∞ bound on the losses) can yield different regret bounds, see discussion in (Audibert et al., 2011).

The problem of BLO is a natural generalization of the classical Multi-Armed Bandit problem and extremely useful for efficiently modeling decision making under partial feedback for structured problems. As such the research literature is rich with algorithms and insights

into this fundamental problem (see surveys Bubeck and Cesa-Bianchi 2012b and Hazan 2014). In this section we focus on the first efficient and optimal-regret algorithm, and thus immediately jump to Algorithm 5. We make the following assumptions over the decision set $\mathcal{K}$:

1. The set $\mathcal{K}$ is equipped with a membership oracle. This implies via the results by (Lovász and Vempala, 2007) (Lemma 10) that there exists an efficient algorithm for sampling from a given log-concave distribution over $\mathcal{K}$. Via the discussion in previous sections, this also implies that we can construct approximate (both types of approximations, see Definitions 23 and 4) volumetric spanners efficiently over $\mathcal{K}$.

2. The losses are bounded in absolute values by 1. That is, the loss functions are always chosen (by an oblivious adversary) from a convex set $\mathcal{Z}$ such that $\mathcal{K}$ is contained in its polar, i.e., $\forall L \in \mathcal{Z}, x \in \mathcal{K}$, $|L^\top x| \leq 1$. This implies that the set $\mathcal{K}$ admits for any $\varepsilon > 0$ an ε-net, w.r.t the norm defined by $\mathcal{Z}$, whose size we denote by $|K|_\varepsilon \leq (\varepsilon/2)^{-d}$.

For Algorithm 5 we prove the optimal regret of Theorem 34.

Remark: notice that to obtain a (p, ε)-exp-volumetric spanner for a log-concave distribution p over a body $\mathcal{K}$ we simply choose sufficiently many i.i.d samples from p. Since in Algorithm 5 p_t is always log-concave, it follows that S'_t consists of i.i.d samples from p_t, meaning that if we would not have required S''_t, the exploration and exploration strategies would be the same! Since we still require the set S''_t, there exists a need for a separate exploration strategy. Interestingly, the $2\sqrt{d}$-ratio-volumetric spanner is obtained by taking a barycentric spanner, which is the exploration strategy given in (Dani et al., 2007).

Algorithm 5 GeometricHedge with Volumetric Spanners Exploration

$\mathcal{K}$, parameters γ, η, horizon T.
$p_1(x)$ uniform distribution over $\mathcal{K}$.
for $t = 1$ to T **do**
  Let S'_t be a $(p_t, \exp(-(4\sqrt{d} + \log(2T))))$-exp-volumetric spanner of $\mathcal{K}$.
  Let S''_t be a $2\sqrt{d}$-ratio-volumetric spanner of $\mathcal{K}$
  Set S_t as the union of S'_t, S''_t.
  $\hat{p}_t(x) = (1-\gamma)p_t(x) + \frac{\gamma}{|S_t|}1_{x \in S_t}$
  sample x_t according to $\hat{p}_t$ (via the tools described in Lemma 10)
  observe loss $\ell_t \triangleq L_t^\top x_t$
  Let $C_t \triangleq \mathbf{E}_{x \sim \hat{p}_t}[xx^\top]$
  $\hat{L}_t \triangleq \ell_t C_t^{-1} x_t$
  $p_{t+1}(x) \propto p_t(x) e^{-\eta \hat{L}_t^\top x}$
end for

Theorem 34 *Under the assumptions stated above, and let* $s = \max_t |S_t|$, $\eta = \sqrt{\frac{\log |\mathcal{K}|_{1/T}}{dT}}$ *and let* $\gamma = s\sqrt{\frac{\log(|\mathcal{K}|_{1/T})}{dT}}$. *Algorithm 5 given parameters* γ, η *suffers a regret bounded by*

$$O\left((s+d)\sqrt{\frac{T \log |\mathcal{K}|_{1/T}}{d}}\right)$$

We note that while the size $\log(|\mathcal{K}|_{1/T})$ can be bounded by $d\log(T)$, in certain scenarios such as s-t paths in graphs it is possible to obtain sharper upper bounds that immediately imply better regret via Theorem 34.

Corollary 35 *There exist an efficient algorithm for BLO for any convex set $\mathcal{K}$ with regret of*

$$O\left(\sqrt{dT\log|\mathcal{K}|_{1/T}}\right) = O\left(d\sqrt{T\log(T)}\right)$$

Proof The spanner in Step 4 of the algorithm does not have to be explicitly constructed. According to Theorem 5, to obtain such as spanner it suffices to sample sufficiently many points from the distribution p_t, hence this portion of the exploration strategy is identical to the exploitation strategy.

According to Corollary 24, a $2\sqrt{d}$-ratio-volumetric spanner of size d can be efficiently constructed, given a linear optimization oracle which in turn can be efficiently implemented by the membership oracle for $\mathcal{K}$. Hence, it follows that for the purpose of the analysis, $s = d$ and the bound follows. ■

To prove the theorem we follow the general methodology used in analyzing the performance of the geometric hedge algorithm. The major deviation from standard technique is the following sub-exponential tail bound, which we use to replace the the standard absolute bound for $|\hat{L}_t x|$. After giving its proof and a few auxiliary lemmas, we give the proof of the main theorem.

Lemma 36 *Let $x \sim p_t$, $x_t \sim \hat{p}_t$ and let $\hat{L}_t$ be defined according to x_t. It holds, for any $\theta > 1$ that*

$$\Pr\left[|\hat{L}_t^\top x| > \frac{\theta s}{\gamma}\right] \le \exp(-2\theta)/T$$

Proof

$$\begin{aligned}
\Pr\left[|\hat{L}_t^\top x| > \frac{\theta s}{\gamma}\right] &\le \Pr\left[\|x\|_{\mathcal{E}(S_t)} \cdot \|x_t\|_{\mathcal{E}(S_t)} \ge \theta\right] && \text{Lemma 37}\\
&\le \Pr\left[\|x\|_{\mathcal{E}(S_t)} \ge \sqrt{\theta} \ \bigvee \ \|x_t\|_{\mathcal{E}(S_t)} \ge \sqrt{\theta}\right]\\
&\le \Pr\left[\|x\|_{\mathcal{E}(S_t)} \ge \sqrt{\theta}\right] + \Pr\left[\|x_t\|_{\mathcal{E}(S_t)} \ge \sqrt{\theta}\right]\\
&\le 2\Pr\left[\|x\|_{\mathcal{E}(S_t)} \ge \sqrt{\theta}\right]
\end{aligned}$$

To justify the last inequality notice that $x \sim p_t$ and $x_t \sim \hat{p}_t$ where $\hat{p}_t$ is a convex sum of p_t and a distribution q_t for which $\Pr_{y\sim q_t}\left[\|y\|_{\mathcal{E}(S_t)} \ge \sqrt{\theta} > 1\right] = 0$. Before we continue recall that we can assume that $\sqrt{\theta} \le 2\sqrt{d}$, since S_t'' is a $2\sqrt{d}$-ratio-volumetric spanner .

$$\begin{aligned}
\Pr\left[|\hat{L}_t^\top x| > \frac{\theta s}{\gamma}\right] &\le 2\Pr\left[\|x\|_{\mathcal{E}(S_t)} \ge \sqrt{\theta}\right]\\
&\le 2\exp(-\sqrt{\theta}(4\sqrt{d} + \log 2T)) && \text{property of exp-volumetric spanner}\\
&\le \tfrac{1}{T}\exp(-2\sqrt{\theta\cdot 4d})\\
&\le \tfrac{1}{T}\exp(-2\theta) && \text{since } \theta \le 4d
\end{aligned}$$

■

Lemma 37 *For all $x \in \mathcal{K}$ it holds that $|\hat{L}_t^\top x| \leq \frac{|S_t|\|x\|_{\mathcal{E}(S_t)}\|x_t\|_{\mathcal{E}(S_t)}}{\gamma}$.*

Proof Let $x \in \mathcal{K}$. Denote by V_t the matrix whose columns are the elements of S_t and recall that $\|y\|^2_{\mathcal{E}(S_t)} = y^\top (V_t V_t^\top)^{-1} y$. Since $C_t \triangleq \mathbf{E}_{x\sim\hat{p}_t}[xx^\top]$, it holds that

$$C_t \succeq \frac{\gamma}{|S_t|}\sum_{v\in S_t} vv^\top = \frac{\gamma}{|S_t|} V_t V_t^\top$$

since both matrices are full rank, it holds that

$$C_t^{-1} \preceq \frac{|S_t|}{\gamma}(V_t V_t^\top)^{-1}$$

Notice that due to the Cauchy-Schwartz inequality,

$$|x^\top \hat{L}_t| = |\ell_t| \cdot |x^\top C_t^{-1} x_t| \leq |\ell_t| \cdot \|x^\top C_t^{-1/2}\| \cdot \|C_t^{-1/2} x_t\|$$

The matrix $C_t^{-1/2}$ is defined as C_t is positive definite. Now,

$$\|x^\top C_t^{-1/2}\|^2 = x^\top C_t^{-1} x \leq x^\top \frac{|S_t|}{\gamma}(V_t V_t^\top)^{-1} x = \frac{|S_t|}{\gamma}\|x\|^2_{\mathcal{E}(S_t)}$$

Since the analog can be said for $\|C_t^{-1/2} x_t\|$ (as $x_t \in \mathcal{K}$), it follows that

$$|x^\top \hat{L}_t| \leq |\ell_t| \frac{|S_t|\|x\|_{\mathcal{E}(S_t)}\|x_t\|_{\mathcal{E}(S_t)}}{\gamma} \leq \frac{|S_t|\|x\|_{\mathcal{E}(S_t)}\|x_t\|_{\mathcal{E}(S_t)}}{\gamma}$$

The last inequality is since we assume the rewards are in $[-1, 1]$. ■

Implementation for general convex bodies. In the case where the set $\mathcal{K}$ is a general convex body, the analysis must include the fact that we can only approximately sample a log-concave distribution over $\mathcal{K}$. As the main focus of our work is to prove a polynomial solution we present only a simple analysis yielding a running time polynomial in the dimension d and horizon T. It is likely that a more thorough analysis can substantially reduce the running time.

Corollary 38 *In the general case where an approximate sampling is required, Algorithm 5 can be implemented with a running time of $\tilde{O}(d^5 + d^3 T^6)$ per iteration.*

Proof

Fix an error parameter δ, and let us run Algorithm 5 with the approximate samplers p'_t guaranteed by Theorem 10. Then, in each use of the sampler we are replacing the true distribution we should be using p_t, with a distribution p'_t such that statistical distance between p_t, p'_t is at most δ. Let us now analyze the error incurred by this approximation by bounding

the loss from the first round onwards. Suppose the algorithm ran with the approximate sampler in the first round but the exact sampler in each round afterwards. Then, as the statistical distance between the distributions is at most δ and the loss in each round is bounded by 1 and there are T rounds, the net difference in expected regret between using p_1 and p'_1 will be at most $\delta \cdot T$. Similarly, if we ran the algorithm with $p'_1, \ldots, p'_{i-1}, p'_i, p_{i+1}, \cdots, p_T$ as opposed to[7] $p'_1, \ldots, p'_{i-1}, p_i, p_{i+1}, \cdots, p_T$ (we are changing the i'th distribution from exact to approximate), the net difference in expected regret would be at most $\delta \cdot T$. Therefore, the total additional loss we may incur for using the approximate oracles is at most $T \cdot (\delta T) = \delta T^2$. Thus, if we take $\delta = \Delta/T^2$, where Δ is the regret bound from Theorem 34, we get a regret bound of 2Δ. The required value of δ is bounded by $T^{-1.5}$. Applying Theorem 10 leads to a running time of $\tilde{O}(d^5 + d^3T^6)$ per iteration. ■

7.1 Proof of Theorem 34

We continue the analysis of the Geometric Hedge algorithm similarly to (Dani et al., 2007; Bubeck et al., 2012), under certain assumptions over the exploration strategy. For convenience we will assume that the set of possible arms $\mathcal{K}$ is finite. This assumption holds w.l.o.g since if $\mathcal{K}$ is infinite, a $\sqrt{1/T}$-net of it can be considered as described earlier (this will have no effect on the computational complexity of our algorithm, but a mere technical convenience in the proof below).

Before proving the theorem we will require three technical lemmas. In the first we show that $\hat{L}_t$ is an unbiased estimator of L_t. In the second, we bound a proxy of its variance. In the third, we bound a proxy of the expected value of its exponent.

Lemma 39 *In each t, $\hat{L}_t$ is an unbiased estimator of L_t*

Proof

$$\hat{L}_t = \ell_t C_t^{-1} x_t = (L_t^\top x_t) C_t^{-1} x_t = C_t^{-1}(x_t x_t^\top) L_t$$

Hence,

$$\mathop{\mathbf{E}}_{x_t \sim p_t}[\hat{L}_t] = C_t^{-1} \mathop{\mathbf{E}}_{x_t \sim p_t}[x_t x_t^\top] L_t = C_t^{-1} C_t L_t = L_t$$

■

Lemma 40 *Let $t \in [T]$, $x \sim p_t$ and $x_t \sim \hat{p}_t$. It holds that $\mathbf{E}[(\hat{L}_t^\top x)^2] \leq d/(1-\gamma) \leq 2d$*

Proof For convenience, denote by q_t the uniform distribution over S_t - the exploration strategy at round t. First notice that for any $x \in \mathcal{K}$,

$$\begin{aligned}
\mathop{\mathbf{E}}_{x_t \sim \hat{p}_t}[(\hat{L}_t^\top x)^2] &= x^\top \mathbf{E}_{x_t \sim \hat{p}_t}[\hat{L}_t \hat{L}_t^\top] x = x^\top \mathbf{E}_{x_t \sim \hat{p}_t}[\ell_t^2 C_t^{-1} x_t x_t^\top C_t^{-1}] x \\
&= \ell_t^2 x^\top C_t^{-1} \mathbf{E}_{x_t \sim \hat{p}_t}[x_t x_t^\top] C_t^{-1} x = \ell_t^2 x^\top C_t^{-1} x \\
&\leq x^\top C_t^{-1} x \qquad (6)
\end{aligned}$$

7. Here, p_j's are interpreted as the distribution given by the algorithm based on the distribution from previous round and p'_j is the approximate oracle for this distribution p_j.

Next,

$$\mathop{\mathbf{E}}_{x\sim\hat{p}_t}[x^\top C_t^{-1}x] = \mathop{\mathbf{E}}_{x\sim\hat{p}_t}[C_t^{-1}\bullet xx^\top] = C_t^{-1}\bullet \mathop{\mathbf{E}}_{x\sim\hat{p}_t}[xx^\top] = C_t^{-1}\bullet C_t = \mathbf{Tr}(I_d) = d$$

Where we used linearity of expectation and denote $A\bullet B = \mathbf{Tr}(AB)$. Since C_t^{-1} is positive semi definite,

$$(1-\gamma)\mathop{\mathbf{E}}_{x\sim p_t}[x^\top C_t^{-1}x] \le (1-\gamma)\mathop{\mathbf{E}}_{x\sim p_t}[x^\top C_t^{-1}x] + \gamma\mathop{\mathbf{E}}_{x\sim q_t}[x^\top C_t^{-1}x] = \mathop{\mathbf{E}}_{x\sim\hat{p}_t}[x^\top C_t^{-1}x] = d \quad (7)$$

The lemma follows from combining Equations 6 and 7. ■

Lemma 41 *Denote by $\mathbf{1}_\phi$ the random variable taking a value of 1 if event ϕ occurred and 0 otherwise. Let $t\in[T]$, $x_t\sim\hat{p}_t$ and $x\sim p_t$. For $\hat{L}_t$ defined by x_t it holds that*

$$\mathbf{E}\left[\exp(-\eta\hat{L}_t^\top x)\mathbf{1}_{-\eta\hat{L}_t^\top x>1}\right] \le \frac{2}{T}$$

Proof Let f, F be the pdf and cdf of the random variable $Y = -\eta\hat{L}_t^\top x$ correspondingly. From Lemma 36 and the fact that $1/\eta = s/\gamma$ ($s = \max_t|S_t|$) we have that for any $\theta\ge 1$,

$$1 - F(\theta) \le \frac{1}{T}e^{-2\theta}$$

and we'd like to prove that under this condition,

$$\mathbf{E}[e^Y\mathbf{1}_{Y>1}] = \int_{\theta=1}^{\infty} e^\theta f(\theta)d\theta \le \frac{2}{T}$$

which follows from the definition of the cdf and pdf:

$$\begin{aligned}
\mathbf{E}[e^Y\mathbf{1}_{Y>1}] &= \textstyle\int_{\theta=1}^{\infty} e^\theta f(\theta)d\theta \\
&= \textstyle\sum_{k=1}^{\infty}\int_{\theta=k}^{k+1} e^\theta f(\theta)d\theta \\
&\le \textstyle\sum_{k=1}^{\infty} e^{k+1}\int_{\theta=k}^{k+1} f(\theta)d\theta \\
&\le \textstyle\sum_{k=1}^{\infty} e^{k+1}(F(k+1)-F(k)) \\
&\le \textstyle\sum_{k=1}^{\infty} e^{k+1}(1-F(k)) \\
&\le \textstyle\sum_{k=1}^{\infty} e^{k+1}\cdot\frac{1}{T}e^{-2k} \qquad \text{Lemma 36} \\
&= \tfrac{e}{T}\textstyle\sum_{k=1}^{\infty} e^{-k} = \tfrac{e}{T}\cdot\tfrac{e^{-1}}{1-e^{-1}} \le \tfrac{2}{T}
\end{aligned}$$

■

Proof [Proof of Theorem 34] For convenience we define within this proof for $x\in\mathcal{K}$, $\hat{\ell}_{1:t-1}(x) \triangleq \sum_{i=1}^{t-1}\hat{L}_i^\top x$ and let $\hat{\ell}_t(x) \triangleq \hat{L}_t^\top x$. Let $W_t = \sum_{x\in\mathcal{K}}\exp(-\eta\hat{\ell}_{1:t-1}(x))$. For all

$t \in [T]$:

$$\begin{aligned}
\mathbf{E}\left[\frac{W_{t+1}}{W_t}\right] &= \mathbf{E}\left[\sum_{x\in\mathcal{K}} \frac{\exp(-\eta\hat{\ell}_{1:t-1}(x))\exp(-\eta\hat{\ell}_t(x))}{W_t}\right] \\
&= \mathbf{E}_{x_t\sim\hat{p}_t}\left[\sum_{x\in\mathcal{K}} p_t(x)\exp(-\eta\hat{\ell}_t(x))\right] \\
&= \mathbf{E}_{x_t\sim\hat{p}_t, x\sim p_t}[\exp(-\eta\hat{\ell}_t(x))] \leq \\
&\leq 1 - \eta\,\mathbf{E}[\hat{L}_t^\top x] + \eta^2\,\mathbf{E}[(\hat{L}_t^\top x)^2] + \mathbf{E}\left[\exp(-\eta\hat{L}_t^\top x)\mathbf{1}_{-\eta\hat{L}_t^\top x>1}\right]
\end{aligned}$$

using the inequality $\exp(y) \leq 1 + y + y^2 + \exp(y)\cdot\mathbf{1}_{y>1}$

$$\leq 1 - \eta\,\mathbf{E}[\hat{L}_t^\top x] + \eta^2\,\mathbf{E}[(\hat{L}_t^\top x)^2] + \tfrac{2}{T} \qquad \text{Lemma 41}$$

Since $\hat{L}_t$ is an unbiased estimator of L_t (Lemma 39) and according to Lemma 40, $\mathbf{E}[(\hat{L}_t^\top x)^2] \leq 2d$, we get:

$$\mathbf{E}\left[\frac{W_{t+1}}{W_t}\right] \leq 1 - \eta L_t^\top \mathop{\mathbf{E}}_{x\sim p_t}[x] + 2\eta^2 d + \frac{2}{T} \tag{8}$$

We now use Jensen's inequality:

$$\begin{aligned}
\mathbf{E}[\log(W_T)] - \mathbf{E}[\log(W_1)] &= \mathbf{E}[\log(W_T/W_1)] \\
&= \textstyle\sum_{t=1}^{T-1} \mathbf{E}[\log(W_{t+1}/W_t)] \\
&\leq \textstyle\sum_{t=1}^{T-1} \log(\mathbf{E}[W_{t+1}/W_t]) && \text{Jensen} \\
&\leq \textstyle\sum_{t=1}^{T-1} \log\left(1 - \eta L_t^\top \mathbf{E}_{x\sim p_t}[x] + 2\eta^2 d + \frac{2}{T}\right) && (8) \\
&\leq \textstyle\sum_{t=1}^{T-1} -\eta L_t^\top \mathbf{E}_{x\sim p_t}[x] + 2\eta^2 d + \frac{2}{T} && \ln(1+y) \leq y \\
& && \text{for all } y > -1 \\
&\leq 2 + 2\eta^2 T d - \eta \textstyle\sum_t \mathbf{E}_{x\sim p_t}[L_t^\top x]
\end{aligned}$$

Now, since $\log(W_1) = \log(|\mathcal{K}|)$ and $W_T \geq \exp(-\eta\hat{\ell}_{1:T}(x^*))$ for any $x^* \in \mathcal{K}$, by shifting sides of the above it holds for any $x^* \in \mathcal{K}$ that

$$\sum_t \mathop{\mathbf{E}}_{x\sim p_t}[L_t^\top x] - \sum_t L_t^\top x^* \leq \sum_t \mathop{\mathbf{E}}_{x\sim p_t}[L_t^\top x] + \mathbf{E}[\log W_T] \leq \frac{\log(|\mathcal{K}|)+2}{\eta} + 2\eta T d$$

Finally, by noticing that

$$\sum_t \mathop{\mathbf{E}}_{x\sim \hat{p}_t}[L_t^\top x] - \sum_t \mathop{\mathbf{E}}_{x\sim p_t}[L_t^\top x] \leq \gamma T$$

we obtain a bound of

$$\mathbf{E}[\text{Regret}] = \mathbf{E}[\sum_t L_t^\top x_t] - \sum_t L_t^\top x^* = \sum_t \mathop{\mathbf{E}}_{x\sim\hat{p}_t}[L_t^\top x] - \text{Loss}(x^*) \leq \frac{\log(|\mathcal{K}|)+2}{\eta} + 2\eta T d + \gamma T$$

on the expected regret. By plugging in the values of η, γ we get the bound of

$$O\left((s+d)\sqrt{\frac{T\log(|\mathcal{K}|)}{d}}\right)$$

as required. ■

8. Conclusion and Open Question

We have described a geometric mechanism for exploration in machine learning problems and its application to experiment design as well as bandit linear optimization.

The following question in high-dimensional geometry remains open: what is the worst-case order of a given set in $\mathbb{R}^d$ (cardinality of its minimal volumetric ellipsoid, as per Definition 1)? A gap remains between our lower bound of $d+1$ and upper bound of $12d$.

Acknowledgments

We would like to thank Raghu Meka for helpful discussions in areas related to high dimensional geometry that were key to the results of this paper.

Elad Hazan gratefully acknowledges support from the European Research Council, in the form of a Marie Curie fellowship and an ERC Starting grant project SUBLRN.

Appendix A. Concentration bounds for non centered isotropic log concave distributions

We begin by proving an auxiliary lemma used in the proof of Corollary 13.

Lemma 42 *Let $\delta > 0$, $t \geq 1$, let d be a positive integer and let $n = \frac{Ct^4 d\log^2(t/\delta)}{\delta^2}$ for some sufficiently large universal constant C. Let $y_1, \ldots, y_n$ be i.i.d d-dimensional vectors from an isotropic log-concave distribution. Then*

$$\Pr\left[\left\|\frac{1}{n}\sum y_i\right\| > \delta\right] \leq \exp(-t\sqrt{d})$$

Proof For convenience let $S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n y_i$. Since the y's are independent, S_n is also log-concave distributed. Notice that $\mathbf{E}[S_n] = 0$ and $\mathbf{E}[S_n S_n^T] = \frac{1}{n}\sum \mathbf{E}[y_i y_i^T] = I_d$ hence S_n is isotropic. Now,

$$\begin{aligned}\Pr\left[\left\|\frac{1}{n}\sum y_i\right\| > \delta\right] &= \Pr\left[\|S_n\| > \sqrt{n}\delta\right] \\ &= \Pr\left[\|S_n\| > \sqrt{d}\cdot\sqrt{Ct^4\log^2(t/\delta)}\right] \\ &\leq \Pr\left[\|S_n\| > \sqrt{d} + \sqrt{d}\cdot\tfrac{1}{2}t\sqrt{C}\right]\end{aligned}$$

The last inequality holds for $t \geq 1$ and $C \geq 4$. It now follows from Theorem 14 that

$$\Pr\left[\left\|\frac{1}{n}\sum y_i\right\| > \delta\right] \leq c_1 \exp(-c_2 t\sqrt{Cd})$$

where c_1, c_2 are some universal constants. Since $t\sqrt{d} \geq 1$, setting $C \geq (\frac{1+\log(c_1)}{c_2})^2$ proves the claim. ■

Proof [Proof of Corollary 13] Let $a = \mathbf{E}[x]$ and let $\tilde{a} = \frac{1}{n}\sum x_i$. Notice that

$$\mathbf{E}[(x-a)(x-a)^T] = \mathbf{E}[xx^T] - \mathbf{E}[x]a^T - a\,\mathbf{E}[x] + aa^T = I_d - aa^T$$

is a PSD matrix hence $\|a\| \leq 1$. Consider the following equality.

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-a)(x_i-a)^\top = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top - a\tilde{a}^\top - \tilde{a}a^\top + aa^\top$$

According to Lemma 42, w.p. at least $1 - \exp(-t\sqrt{d})$,

$$\|\tilde{a} - a\| \leq \delta$$

in which case, since $\|a\| \leq 1$ and according to the triangle inequality,

$$\left\|\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top - I_d\right\| \leq \left\|\frac{1}{n}\sum_{i=1}^{n}(x_i-a)(x_i-a)^\top - (I_d - aa^\top)\right\| + 2\delta$$

According to Theorem 12, w.p. at least $1 - \exp(-t\sqrt{d})$

$$\left\|\frac{1}{n}\sum_{i=1}^{n}(x_i-a)(x_i-a)^\top - (I_d - aa^\top)\right\| \leq \delta$$

and the corollary follows. ■

Proof [Proof of Corollary 15] Let $\mathbf{E}[x] = a$. Consider the r.v $y = x - a$. It holds that $\mathbf{E}[y] = 0$ and $\mathbf{E}[yy^T] = I_d - aa^T$. Notice that we can derive that

$$\|a\| \leq 1 \tag{9}$$

As $\mathbf{E}[yy^T]$ is a PSD matrix. Also, it is easy to verify that y is log-concave distributed. We now consider the r.v[8] $z = (I_d - aa^t)^{-1/2}y$. It is easy to verify that the distribution of z is also log-concave and isotropic. It follows, from Theorem 14 that for any $\theta \geq 2$

$$\Pr\left[\|z\| > \theta\sqrt{d}\right] \leq \Pr\left[\|z\| - \sqrt{d} > \frac{1}{2}\theta\sqrt{d}\right] \leq C' \exp(-c\theta\sqrt{d})$$

By using Equation 9 we get that for $\theta > 3$

$$\Pr\left[\|x\| > \theta\sqrt{d}\right] \leq \Pr\left[\|y\| > \theta\sqrt{d} - 1\right] \leq \Pr\left[\|z\| > (\theta - 1/\sqrt{d})\sqrt{d}\right] \leq C' \exp(c' - c'\theta\sqrt{d}).$$

The last inequality holds since $\theta - 1/\sqrt{d} \geq 2$. ■

8. if $I_d - aa^T$ is not of full rank then y is in fact supported in an affine subspace of rank $d-1$ and we can continue the analysis there.

References

J.D. Abernethy, E. Hazan, and A. Rakhlin. Interior-point methods for full-information and bandit online learning. *IEEE Transactions on Information Theory*, 58(7):4164–4175, 2012.

Radoslaw Adamczak, Alexander E. Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of American Mathematical Society*, 23:535–561, 2010.

A.A.C. Atkinson and A.A.N. Donev. *Optimum Experimental Designs*. Oxford science publications. OXFORD University Press, 1992. ISBN 9780198522546. URL `http://books.google.co.il/books?id=cmmOA_-M7SOC`.

Jean-Yves Audibert and Olivier Catoni. Linear regression through pac-bayesian truncation. *arXiv preprint arXiv:1010.0072*, 2010.

Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *The Annals of Statistics*, pages 2766–2794, 2011.

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Minimax policies for combinatorial prediction games. In Sham M. Kakade and Ulrike von Luxburg, editors, *COLT*, volume 19 of *Journal of Machine Learning Research - Proceedings Track*, pages 107–132. JMLR.org, 2011.

Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114, 2008.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, January 2009. ISSN 0022-0000. doi: 10.1016/j.jcss.2008.07.003. URL `http://dx.doi.org/10.1016/j.jcss.2008.07.003`.

Keith Ball. An elementary introduction to modern convex geometry. In *Flavors of Geometry*, pages 1–58. Univ. Press, 1997.

Peter L. Bartlett, Varsha Dani, Thomas P. Hayes, Sham Kakade, Alexander Rakhlin, and Ambuj Tewari. High-probability regret bounds for bandit online linear optimization. In *Proceedings of The 21st Conference on Learning Theory (COLT)*, pages 335–342, 2008.

Joshua Batson, Daniel A Spielman, and Nikhil Srivastava. Twice-ramanujan sparsifiers. *SIAM Journal on Computing*, 41(6):1704–1721, 2012.

S. Bubeck and N. Cesa-Bianchi. *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*, volume 5 of *Foundations and Trends in Machine Learning*. NOW, 2012a.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5:1–122, 2012b.

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Sham M. Kakade. Towards minimax policies for online linear optimization with bandit feedback. *Journal of Machine Learning Research - Proceedings Track*, 23:41.1–41.14, 2012.

Nicolò Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

Kenneth L. Clarkson, Elad Hazan, and David P. Woodruff. Sublinear optimization for machine learning. *Journal of the ACM (JACM)*, 59(5):23:1–23:49, November 2012.

David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

S. Damla Ahipasaoglu, Peng Sun, and Michael J. Todd. Linear convergence of a modified frank–wolfe algorithm for computing minimum-volume enclosing ellipsoids. *Optimization Methods and Software*, 23(1):5–19, 2008.

Varsha Dani, Sham M Kakade, and Thomas P Hayes. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems*, pages 345–352, 2007.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of The 21st Conference on Learning Theory (COLT)*, pages 355–366, 2008.

Sanjoy Dasgupta and John Langford. Active learning tutorial. *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 2009. URL `http://hunch.net/~active_learning/`.

Sanjoy Dasgupta, Adam Tauman Kalai, and Claire Monteleoni. Analysis of perceptron-based active learning. *Journal of Machine Learning Research*, 10:281–299, 2009.

Ravi Ganti and Alexander G. Gray. Upal: Unbiased pool based active learning. In *AISTATS*, volume 22 of *Journal of Machine Learning Research - Proceedings Track*, pages 422–431. JMLR.org, 2012.

Olivier Guédon and Emanuel Milman. Interpolating thin-shell and sharp large-deviation estimates for lsotropic log-concave measures. *Geometric and Functional Analysis*, 21(5): 1043–1068, 2011. ISSN 1016-443X.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Steve Hanneke. Teaching dimension and the complexity of active learning. In *Proceedings of The 20th Conference on Learning Theory (COLT)*, pages 66–81. Springer-Verlag, 2007.

Elad Hazan. Introduction to online convex optimization. *Manuscript*, 2014. URL `http://ocobook.cs.princeton.edu/OCObook.pdf`.

Elad Hazan, Zohar Karnin, and Raghu Meka. Volumetric spanners: an efficient exploration basis for learning. In Maria-Florina Balcan and Csaba Szepesvári, editors, *COLT*, volume 35 of *Journal of Machine Learning Research - Proceedings Track*, pages 408–422. JMLR.org, 2014.

Martin Henk. Löwner-John ellipsoids. *Documenta Mathematica*, pages 95–106, 2012.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. *Journal of Machine Learning Research - Proceedings Track*, 23:9.1–9.24, 2012.

F. John. Extremum Problems with Inequalities as Subsidiary Conditions. In K. O. Friedrichs, O. E. Neugebauer, and J. J. Stoker, editors, *Studies and Essays: Courant Anniversary Volume*, pages 187–204. Wiley-Interscience, New York, 1948.

Sham M Kakade, Adam Tauman Kalai, and Katrina Ligett. Playing games with approximation algorithms. *SIAM Journal on Computing*, 39(3):1088–1106, 2009.

Zohar Karnin and Elad Hazan. Hard-margin active linear regression. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 883–891, 2014.

Leonid G Khachiyan. Rounding of polytopes in the real number model of computation. *Mathematics of Operations Research*, 21(2):307–320, 1996.

László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, 2004.

Andrew McCallum and Kamal Nigam. Employing em and pool-based active learning for text classification. In *Proceedings of the 15th International Conference on Machine Learning (ICML)*, pages 350–358, 1998.

M. Rudelson. Random vectors in the isotropic position. *Journal of Functional Analysis*, 164(1):60–72, 1999.

MC Spruill and WJ Studden. A kiefer-wolfowitz theorem in a stochastic process setting. *The Annals of Statistics*, 7(6):1329–1332, 1979.

Chien-Fu Wu. Some algorithmic aspects of the theory of optimal designs. *The Annals of Statistics*, 6(6):1286–1301, 1978.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 928–936, 2003.