Journal of Machine Learning Research 12 (2011) 1287-1311 Submitted 8/09; Revised 10/12; Published 4/11

Better Algorithms for Benign Bandits

Elad Hazan EHAZAN@IE.TECHNION.AC.IL
Department of IE&M
Technion–Israel Institute of Technology
Haifa 32000, Israel

Satyen Kale* SKALE@YAHOO-INC.COM
Yahoo! Research
4301 Great America Parkway
Santa Clara, CA, USA

Editor: Nicolo Cesa-Bianchi

Abstract

The online multi-armed bandit problem and its generalizations are repeated decision making problems, where the goal is to select one of several possible decisions in every round, and incur a cost associated with the decision, in such a way that the total cost incurred over all iterations is close to the cost of the best fixed decision in hindsight. The difference in these costs is known as the *regret* of the algorithm. The term *bandit* refers to the setting where one only obtains the cost of the decision used in a given iteration and no other information.

A very general form of this problem is the non-stochastic bandit linear optimization problem, where the set of decisions is a convex set in some Euclidean space, and the cost functions are linear. Only recently an efficient algorithm attaining $\tilde{O}(\sqrt{T})$ regret was discovered in this setting.

In this paper we propose a new algorithm for the bandit linear optimization problem which obtains a tighter regret bound of $\tilde{O}(\sqrt{Q})$, where Q is the total variation in the cost functions. This regret bound, previously conjectured to hold in the full information case, shows that it is possible to incur much less regret in a slowly changing environment even in the bandit setting. Our algorithm is efficient and applies several new ideas to bandit optimization such as reservoir sampling.

Keywords: multi-armed bandit, regret minimization, online learning

1. Introduction

Consider a person who commutes to work every day. Each morning, she has a choice of routes to her office. She chooses one route every day based on her past experience. When she reaches her office, she records the time it took her on that route that day, and uses this information to choose routes in the future. She doesn't obtain any information on the other routes she could have chosen to work. She would like to minimize her total time spent commuting in the long run; however, knowing nothing of how traffic patterns might change, she opts for the more pragmatic goal of trying to minimize the total time spent commuting in comparison with the time she would have spent had she full knowledge of the future traffic patterns but had to choose the same fixed route every day. This difference in cost (using time as a metric of cost) measures how much she regrets not knowing traffic patterns and avoiding the hassle of choosing a new path every day.

∗. Work done while the author was at Microsoft Research.

This scenario, and many more like it, are modeled by the multi-armed bandit problem and its generalizations. It can be succinctly described as follows: iteratively an online learner has to choose an action from a set of n available actions. She then suffers a cost (or receives a reward) corresponding to the action she took and no other information as to the merit of other available actions. Her goal is to minimize her *regret*, which is defined as the difference between her total cost and the total cost of the best single action knowing the costs of all actions in advance.

Various models of the "unknown" cost functions have been considered in the last half a century. Robbins (1952) pioneered the study of various stochastic cost functions, followed by Hannan (1957), Lai and Robbins (1985) and others. It is hard to do justice to the numerous contributions and studies and we refer the reader to the book of Cesa-Bianchi and Lugosi (2006) for references. In their influential paper, Auer et al. (2003) considered an adversarial non-stochastic model of costs, and gave an efficient algorithm that attains the optimal regret[1] in terms of the number of iterations, T, a bound of $\tilde{O}(\sqrt{T})$.[2] The sublinear (in T) regret bound implies that on average, the algorithm's cost converges to that of the best fixed action in hindsight.

The latter paper (Auer et al., 2003) was followed by a long line of work (Awerbuch and Kleinberg, 2004; McMahan and Blum, 2004; Flaxman et al., 2005; Dani et al., 2008) which considered the more general case of bandit online linear optimization over a convex domain. In this problem, the learner has to choose a sequence of points from the convex domain and obtains their cost from an unknown linear cost function. The objective, again, is to minimize the regret, that is, the difference between the total cost of the algorithm and that of the best fixed point in hindsight. This generality is crucial to allow for *efficient* algorithms for problems with a large decision space, such as online shortest path problem considered at the beginning. This line of work finally culminated in the work of Abernethy et al. (2008), who obtained the first algorithm to give $\tilde{O}(\sqrt{T})$ regret with polynomial running time.

Even though the $\tilde{O}(\sqrt{T})$ dependence on T was a great achievement, this regret bound is weak from the point of view of real-world bandit scenarios. Rarely would we encounter a case where the cost functions are truly adversarial. Indeed, the first work on this problem assumed a stochastic model of cost functions, which is a very restrictive assumption in many cases. One reasonable way to retain the appeal of worst-case bounds while approximating the steady-state nature of the stochastic setting is to consider the *variation* in the cost vectors.

For example, our office commuter doesn't expect the traffic gods to conspire against her every day. She might expect a certain predictability in traffic patterns. Most days the traffic pattern is about the same, except for some fluctuations depending on the day of the week, time of the day, etc. Coming up with a stochastic model for traffic patterns would be simply too onerous. An algorithm that quickly learns the dominant pattern of the traffic, and achieves regret bounded by the (typically small) variability in day-to-day traffic, would be much more desirable. Such regret bounds naturally interpolate between the stochastic models of Robbins and the worst case models of Auer *et al*.

In this paper[3] we present the first such bandit optimization algorithm in the worst-case adversarial setting, with regret bounded by $\tilde{O}(\sqrt{Q})$, where Q is the total observed variation in observed costs, defined as the sum of squared deviations of the cost vectors from their mean. This regret

1. Strictly speaking, here we talk about *expected* regret, as all algorithms that attain non-trivial guarantees must use randomization.
2. We use the notation $\tilde{O}$ to hide all constant terms (such as dependence on the dimension of the problem, or the diameter of the decision set) and other lower order terms which grow at a poly-logarithmic rate with T.
3. A preliminary version of this result was presented in Hazan and Kale (2009a).

degrades gracefully with increasing Q, and in the worst case, we recover the regret bound $\tilde{O}(\sqrt{T})$ of Abernethy et al. (2008). Our algorithm is efficient, running in polynomial time per iteration.

The conjecture that the regret of online learning algorithms should be bounded in terms of the total variation was put forth by Cesa-Bianchi et al. (2007) in the full information model (where the online player is allowed to observe the costs of actions she did not choose). This conjecture was recently resolved on the affirmative in Hazan and Kale (2008), in two important online learning scenarios, viz. online linear optimization and expert prediction. In addition, in Hazan and Kale (2009b), we give algorithms with regret bounds of $O(\log(Q))$ for the Universal Portfolio Selection problem and its generalizations. In this paper, we prove the surprising fact that such a regret bound of $\tilde{O}(\sqrt{Q})$ is possible to obtain even when the only information available to the player is the cost she incurred (in particular, we may not even be able to estimate Q accurately in this model).

To prove our result we need to overcome the following difficulty: all previous approaches for the non-stochastic multi-armed bandit problem relied on the main tool of "unbiased gradient estimator", that is, the use of randomization to extrapolate the missing information (cost function). The variation in these unbiased estimators is unacceptably large even when the underlying cost function sequence has little or no variation.

To overcome this problem we introduce two new tools: first, we use historical costs to construct our gradient estimators. Next, in order to construct these estimators, we need an accurate method of accumulating historical data. For this we use a method from data streaming algorithms known as "reservoir sampling". This method allows us to maintain an accurate "sketch" of history with very little overhead.

An additional difficulty which arises is the fact that a learning rate parameter η needs to be set based on the total variation Q to obtain the $\tilde{O}(\sqrt{Q})$ regret bound. Typically, in other scenarios where square root regret bound in some parameter is desired, a simple η-halving trick works, but requires the algorithm to be able to compute the relevant parameter after every iteration. However, as remarked earlier, even estimating Q is non-trivial problem. We do manage to bypass this problem by using a novel approach that implicitly mimics the η-halving procedure.

2. Preliminaries

Throughout the paper we use the standard asymptotic $O()$ notation to hide dependence on (absolute, problem-independent) constants. For convenience of notation, we use the $\tilde{O}()$ notation to hide dependence on (problem-dependent) constants as well as polylog(T) factors: $g = \tilde{O}(f)$ if $g < c f \log^d(T)$ for some problem-dependent constant $c > 0$ and and a problem-independent constant $d > 0$. Specifically, in the $\tilde{O}()$ notation we also hide terms which depend on the dimension n, since this is fixed: we use this notation in order highlight the dependence on the parameter which grows with time, viz. the quadratic variation. Unless specified otherwise, all vectors live in $\mathbb{R}^n$, and all matrices in $\mathbb{R}^{n\times n}$. The vector norm $\|\cdot\|$ denotes the standard ℓ_2 norm.

We consider the online linear optimization model in which iteratively the online player chooses a point $\mathbf{x}_t \in \mathcal{K}$, where $\mathcal{K} \subseteq \mathbb{R}^n$ is a convex compact set called the *decision set*. After her choice, an adversary supplies a linear cost function $\mathbf{f}_t$, and the player incurs a cost of $\mathbf{f}_t(\mathbf{x}_t)$. In this paper we assume an *oblivious* adversary, which can choose arbitrary cost functions $\mathbf{f}_t$ in advance, with prior knowledge of the player's algorithm, but the adversary does not have access to the random bits used by the player (see Dani and Hayes, 2006 for more details on various models of adversaries).

With some abuse of notation, we use $\mathbf{f}_t$ to also denote the cost vector such that $\mathbf{f}_t(\mathbf{x}) = \mathbf{f}_t^\top \mathbf{x}$. The *only* information available to the player is the cost incurred, that is, the scalar $\mathbf{f}_t(\mathbf{x}_t)$. Denote the total number of game iterations by T. The standard game-theoretic measure of performance is regret, defined as

$$\text{Regret}_T \;=\; \sum_{t=1}^{T} \mathbf{f}_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{K}} \sum_{t=1}^{T} \mathbf{f}_t(\mathbf{x}).$$

We make some normalizations on the cost vectors and the convex domain $\mathcal{K}$ to keep the presentation clean. We assume that the cost vectors are scaled so that their norms are bounded by one, that is, $\|\mathbf{f}_t\| \le 1$. This scaling only changes the regret bound by a constant factor: if G is a known upper bound on the norms of the cost vectors, we can scale down the cost vectors by G and run the algorithm; the actual regret is G times larger than the bound obtained here. Next, we assume $\mathcal{K}$ is scaled to fit inside the unit ball (in the ℓ_2 norm) centered at the origin, that is, for all $\mathbf{x} \in \mathcal{K}$, we have $\|\mathbf{x}\| \le 1$. We also assume that for some parameter $\gamma \in (0,1)$, all the vectors $\gamma\mathbf{e}_1, \ldots, \gamma\mathbf{e}_n$, where $\mathbf{e}_i$ is the standard basis vector with 1 in the i-th coordinate and 0 everywhere else, are in the decision set $\mathcal{K}$.

The above assumptions can be met by translating and scaling $\mathcal{K}$ appropriately. This only changes the regret bound by a constant factor: if D is a known upper bound on the diameter of $\mathcal{K}$, then we can translate $\mathcal{K}$ so that it contains the origin and scaled coordinate vectors $\gamma'\mathbf{e}_i$ for some $\gamma' > 0$, and then scale it down by D to make its diameter 1 and run the algorithm; the actual regret is D times larger than the bound obtained here. In certain specific cases, one can certainly obtain tighter constants by using a set of n linearly independent vectors contained inside $\mathcal{K}$, but here we make this simplifying assumption for the sake of cleanliness of presentation.

We denote by Q_T the total quadratic variation in cost vectors, that is,

$$Q_T \;:=\; \sum_{t=1}^{T} \|\mathbf{f}_t - \mu\|^2,$$

where $\mu = \frac{1}{T}\sum_{t=1}^{T} \mathbf{f}_t$ is the mean of all cost vectors.

A symmetric matrix $\mathbf{A}$ is called positive semidefinite (denoted by $\mathbf{A} \succeq \mathbf{0}$ if all its eigenvalues are non-negative. If all eigenvalues of $\mathbf{A}$ are strictly positive, the matrix is called positive definite. For symmetric matrices we denote by $\mathbf{A} \preceq \mathbf{B}$ the fact that the matrix $\mathbf{B} - \mathbf{A}$ is positive semidefinite. For a positive definite matrix $\mathbf{A}$ we denote its induced norm by $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^\top \mathbf{A}\mathbf{x}}$. We make use of the following simple generalization of the Cauchy-Schwarz inequality:

$$\mathbf{x}^\top \mathbf{y} \;\le\; \|\mathbf{x}\|_{\mathbf{A}} \cdot \|\mathbf{y}\|_{\mathbf{A}^{-1}}. \tag{1}$$

This inequality follows by applying the usual Cauchy-Schwarz inequality to the vectors $\mathbf{A}^{1/2}\mathbf{x}$ and $(\mathbf{A}^{1/2})^{-1}\mathbf{y}$, where $\mathbf{A}^{1/2}$ is the matrix square root of the positive definite matrix $\mathbf{A}$, that is, a matrix $\mathbf{B}$ which satisfies $\mathbf{B}\mathbf{B} = \mathbf{A}$.

For a twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we denote its gradient by ∇f and its Hessian by $\nabla^2 f$.

2.1 Reservoir Sampling

A crucial ingredient in our algorithm is a sampling procedure ubiquitously used in streaming algorithms known as "reservoir sampling" (Vitter, 1985). In a streaming problem the algorithm gets to

see a stream of data in one pass, and not allowed to re-visit previous data. Suppose the elements of the stream are real numbers $\mathbf{f}_1, \mathbf{f}_2, \ldots$ and our goal is to maintain a randomized estimate of the current mean $\mu_t := \frac{1}{t}\sum_{\tau=1}^{t} \mathbf{f}_\tau$. The main constraint which precludes the trivial solution is that we desire a sampling scheme that touches (i.e., observes the value of) very few elements in the stream.

The reservoir sampling method is to maintain a randomly chosen (without replacement) subset S of size k (also called a "reservoir") from the stream, and then use the average of the sample as an estimator. This works as follows. Initialize S by including the first k elements $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_k$ in the stream. For every subsequent element $\mathbf{f}_t$, we decide to include it in S with probability $\frac{k}{t}$. If the decision to include it is made, then a random element of S is replaced by $\mathbf{f}_t$.

The following lemma is standard (see Vitter, 1985) and we include a simple inductive proof for completeness:

Lemma 1 *For every $t \geq k$, the set S is random subset chosen without replacement uniformly from $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_t\}$.*

Proof We prove this by induction on t. The statement is trivially true for $t = k$. Assume that the statement is true for some $t \geq m$, and we now show $t+1$. Let S be an arbitrary subset of size k of $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_t\}$. We now show that the probability that the chosen set in the $t+1$-th round is S is $\frac{1}{\binom{t+1}{k}}$. For this, we have two cases: $\mathbf{f}_{t+1} \notin S$ and $\mathbf{f}_{t+1} \in S$. In the first case, the probability that S is the chosen subset at the end of t-th round is $\frac{1}{\binom{t}{k}}$ by the induction hypothesis. The conditional probability that it survives the $t+1$-th is $1 - \frac{k}{t+1}$, so the overall probability that S is the chosen set at the end of $t+1$-th round is $(1 - \frac{k}{t+1}) \cdot \frac{1}{\binom{t}{k}} = \frac{1}{\binom{t+1}{k}}$.

In the second case, S will be the chosen set at the end of $t+1$-th round if the set S' chosen at the end of the t-th round is one of the $t+1-k$ sets obtained from S by replacing $\mathbf{f}_{t+1}$ by an element of $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_t\} \setminus S$, which gets replaced by $\mathbf{f}_{t+1}$ in the $t+1$-th round. The probability of this happening is $\frac{t+1-k}{\binom{t}{k}} \cdot \frac{k}{t+1} \cdot \frac{1}{k} = \frac{1}{\binom{t+1}{k}}$, as required. ■

Now suppose we define $\tilde{\mu}_t$ to be the average of the k chosen numbers in S, then the following lemma is immediate:

Lemma 2 *For every $t \geq k$, we have $\mathbf{E}[\tilde{\mu}_t] = \mu_t$ and $VAR[\tilde{\mu}_t] \leq \frac{1}{kt}\sum_{\tau=1}^{t}(\mathbf{f}_\tau - \mu_t)^2 = \frac{1}{kt}Q_t$.*

The bound on the variance follows because the variance of a single randomly chosen element of the stream is $\frac{1}{t}\sum_{\tau=1}^{t}(\mathbf{f}_\tau - \mu_t)^2$. So the variance of the average of k randomly chosen elements *with* replacement is $\frac{1}{kt}\sum_{\tau=1}^{t}(\mathbf{f}_\tau - \mu_t)^2$. Since we choose the k elements in the sample *without* replacement, the variance is only smaller.

The main reason reservoir sampling is useful in our context is because it samples every element *obliviously*, that is, a decision to sample is made without looking at the element. This implies that the expected number of elements touched by the reservoir sampling based estimation procedure for μ_t is $k + \sum_{t=k+1}^{T} \frac{k}{t} = O(k \log(T))$, which is very small compared to the length of the stream, T, if k is set to some small value, like $O(\log(T))$ as in our applications.

2.2 Self-concordant Functions and the Dikin Ellipsoid

In this section we give a few definition and properties of self-concordant barriers that we will crucially need in the analysis. Our treatment of this subject follows Abernethy et al. (2008), who in-

troduced self-concordant functions to online learning. Self-concordance in convex optimization is a beautiful and deep topic, and we refer the reader to Nesterov and Nemirovskii (1994) and Ben-Tal and Nemirovski (2001) for a thorough treatment on the subject.

Definition 1 *A convex function $\mathcal{R}(\mathbf{x})$ defined on the interior of the convex compact set $\mathcal{K}$, and having three continuous derivatives, is said to be a ϑ-self-concordant barrier (where $\vartheta > 0$ is the self-concordance parameter) if the following conditions hold:*

1. *(Barrier property) $\mathcal{R}(\mathbf{x}_i) \to \infty$ along every sequence of points $\mathbf{x}_i$ in the interior of $\mathcal{K}$ converging to a boundary point of $\mathcal{K}$.*

2. *(Differential properties) $\mathcal{R}$ satisfies*

$$|\nabla^3 \mathcal{R}(\mathbf{x})[\mathbf{h},\mathbf{h},\mathbf{h}]| \;\leq\; 2(\mathbf{h}^\top [\nabla^2 \mathcal{R}(\mathbf{x})]\mathbf{h})^{3/2},$$

$$|\nabla \mathcal{R}(\mathbf{x})^\top \mathbf{h}| \;\leq\; \vartheta^{1/2} \left[\mathbf{h}^\top \nabla^2 \mathcal{R}(\mathbf{x})\mathbf{h}\right]^{1/2}.$$

where $\mathbf{x}$ is a point in the interior of $\mathcal{K}$, and $\mathbf{h}$ is an arbitrary vector in $\mathbb{R}^n$. Here, $\nabla \mathcal{R}(\mathbf{x}), \nabla^2 \mathcal{R}(\mathbf{x})$ denote the Gradient and Hessian, respectively, of $\mathcal{R}$ at point $\mathbf{x}$, and

$$\nabla^3 \mathcal{R}(\mathbf{x})[\mathbf{h}_1,\mathbf{h}_2,\mathbf{h}_3] \;=\; \frac{\partial^3}{\partial t_1 \partial t_2 \partial t_3} \mathcal{R}(\mathbf{x} + t_1\mathbf{h}_1 + t_2\mathbf{h}_2 + t_3\mathbf{h}_3)\Big|_{t_1=t_2=t_3=0}$$

Any n-dimensional closed convex set admits an $O(n)$-self-concordant barrier. However, such a barrier may not necessarily be efficiently computable.

More concretely, the standard logarithmic barrier for a half-space $\mathbf{u}^\top \mathbf{x} \leq b$ is given by

$$\mathcal{R}(\mathbf{x}) \;=\; -\log(b - \mathbf{u}^\top \mathbf{x}),$$

and is 1-self-concordant. For polytopes defined by m halfspaces, the standard logarithmic barrier (which is just the sum of all barriers for the defining half-spaces) has the self-concordance parameter $\vartheta = m$.

Definition 2 *For a given $\mathbf{x} \in \mathcal{K}$, and any $\mathbf{h} \in \mathbb{R}^n$, define the norm induced by the Hessian, and its dual norm, to be*

$$\|\mathbf{h}\|_{\mathbf{x}} := \sqrt{\mathbf{h}^\top [\nabla^2 \mathcal{R}(\mathbf{x})]\mathbf{h}}, \;\; \textit{and}$$
$$\|\mathbf{h}\|_{\mathbf{x}}^\star := \sqrt{\mathbf{h}^\top [\nabla^2 \mathcal{R}(\mathbf{x})]^{-1}\mathbf{h}}.$$

Definition 3 *The* Dikin ellipsoid *of radius r centered at $\mathbf{x}$ is the set*

$$W_r(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y} - \mathbf{x}\|_{\mathbf{x}} \leq r\}.$$

When a radius is unspecified, it is assumed to be 1; so "the Dikin ellipsoid at $\mathbf{x}$" refers to the Dikin ellipsoid of radius 1 centered at $\mathbf{x}$.

Definition 4 *For any two distinct points* $\mathbf{x}$ *and* $\mathbf{y}$ *in the interior of* $\mathcal{K}$*, the* Minkowsky function $\pi_{\mathbf{x}}(\mathbf{y})$ *on* $\mathcal{K}$ *is*

$$\pi_{\mathbf{x}}(\mathbf{y}) = \inf\{t \geq 0 : \mathbf{x} + t^{-1}(\mathbf{y} - \mathbf{x}) \in \mathcal{K}\}.$$

The Minkowsky function measures the distance from $\mathbf{x}$ to $\mathbf{y}$ as a portion of the total distance on the ray from $\mathbf{x}$ to the boundary of $\mathcal{K}$ through the point $\mathbf{y}$, and hence $\pi_{\mathbf{x}}(\mathbf{y}) \in [0,1]$.

The following facts about the Dikin ellipsoid and self concordant barriers will be used in the sequel (we refer to Nemirovskii, 2004 for proofs):

1. $W_1(\mathbf{x}) \subseteq \mathcal{K}$ for any $\mathbf{x} \in \mathcal{K}$. This is crucial for most of our sampling steps (the "ellipsoidal sampling"), since we sample from the Dikin ellipsoid $W_1(\mathbf{x}_t)$. Since $W_1(\mathbf{x}_t)$ is contained in $\mathcal{K}$, the sampling procedure yields feasible points.

2. The lengths of the principal axes of the ellipsoid $W_1(\mathbf{x})$ are $2/\sqrt{\lambda_i}$, where λ_i, for $i = 1, 2, \ldots, n$ are the eigenvalues of $\nabla^2\mathcal{R}(\mathbf{x})$. Thus, the fact that $W_1(\mathbf{x}) \subseteq \mathcal{K}$ and that $\mathcal{K}$ is contained in the unit ball implies that $2/\sqrt{\lambda_i} \leq 2$ for all i, or in other words, $1/\lambda_i \leq 1$ for all i. This implies that $[\nabla^2\mathcal{R}(\mathbf{x})]^{-1} \leq \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Thus, we can relate the $\|\cdot\|_{\mathbf{x}}^{\star}$ norm to the standard ℓ_2 norm $\|\cdot\|$: for any vector $\mathbf{h}$,

$$\|\mathbf{h}\|_{\mathbf{x}}^{\star} \;=\; \sqrt{\mathbf{h}^{\top}[\nabla^2\mathcal{R}(\mathbf{x})]^{-1}\mathbf{h}} \;\leq\; \sqrt{\mathbf{h}^{\top}\mathbf{I}\mathbf{h}} \;=\; \|\mathbf{h}\|. \tag{2}$$

3. In the interior of the Dikin ellipsoid at $\mathbf{x}$, the Hessian of $\mathcal{R}$ is "almost constant": for any $\mathbf{h} \in \mathbb{R}^n$ such that $\|\mathbf{h}\|_{\mathbf{x}} < 1$, we have

$$(1 - \|\mathbf{h}\|_{\mathbf{x}})^2\nabla^2\mathcal{R}(\mathbf{x}) \;\preceq\; \nabla^2\mathcal{R}(\mathbf{x}+\mathbf{h}) \;\preceq\; (1 - \|\mathbf{h}\|_{\mathbf{x}})^{-2}\nabla^2\mathcal{R}(\mathbf{x}).$$

4. For any ϑ-self-concordant barrier on $\mathcal{K}$, and for any two distinct points $\mathbf{x}$ and $\mathbf{y}$ in the interior of $\mathcal{K}$, it holds that (see Nemirovskii, 2004)

$$\mathcal{R}(\mathbf{y}) - \mathcal{R}(\mathbf{x}) \leq \vartheta \ln\left(\frac{1}{1 - \pi_{\mathbf{x}}(\mathbf{y})}\right). \tag{3}$$

Definition 5 *Let* $\mathbf{x}^{\circ}$ *be the analytic center of* $\mathcal{K}$ *with respect to the self-concordant barrier* $\mathcal{R}$*, that is, the point inside* $\mathcal{K}$ *in which* $\nabla\mathcal{R}(\mathbf{x}^{\circ}) = 0$*. For any* $\delta > 0$*, define the convex body* $\mathcal{K}_{\delta} \subseteq \mathcal{K}$ *by*

$$\mathcal{K}_{\delta} \;:=\; \{\mathbf{x} | \pi_{\mathbf{x}^{\circ}}(\mathbf{x}) \leq (1 - \delta)\}.$$

The following properties holds for $\mathcal{K}_{\delta}$:

Lemma 3 *For any* $\mathbf{x} \in \mathcal{K}$*, there exists a* $\mathbf{u} \in \mathcal{K}_{\delta}$ *such that* $\|\mathbf{x} - \mathbf{u}\| \leq 2\delta$ *which satisfies*

$$\mathcal{R}(\mathbf{u}) - \mathcal{R}(\mathbf{x}^{\circ}) \;\leq\; \vartheta \ln\frac{1}{\delta}.$$

Proof Let $\mathbf{u} = \delta\mathbf{x}^{\circ} + (1 - \delta)\mathbf{x}$. Since

$$\mathbf{x} \;=\; \mathbf{x}^{\circ} + \frac{\|\mathbf{x} - \mathbf{x}^{\circ}\|}{\|\mathbf{u} - \mathbf{x}^{\circ}\|}(\mathbf{u} - \mathbf{x}^{\circ}),$$

and $\mathbf{x} \in \mathcal{K}$, we have that

$$\pi_{\mathbf{x}^\circ}(\mathbf{u}) \;\leq\; \frac{\|\mathbf{u}-\mathbf{x}^\circ\|}{\|\mathbf{x}-\mathbf{x}^\circ\|} \;=\; \frac{\|(1-\delta)(\mathbf{x}-\mathbf{x}^\circ)\|}{\|\mathbf{x}-\mathbf{x}^\circ\|} \;=\; 1-\delta,$$

which implies that $\mathbf{u} \in \mathcal{K}_\delta$. Next, note that

$$\|\mathbf{x}-\mathbf{u}\| = \|\delta(\mathbf{x}-\mathbf{x}^\circ)\| \;\leq\; 2\delta,$$

since $\mathcal{K}$ is contained in the unit ball.

Finally, since $\mathcal{R}$ is a ϑ-self-concordant barrier, we have by (3)

$$\mathcal{R}(\mathbf{u}) - \mathcal{R}(\mathbf{x}^\circ) \;\leq\; \vartheta \log \frac{1}{1-\pi_{\mathbf{x}^\circ}(\mathbf{u})} \;\leq\; \vartheta \log \frac{1}{1-(1-\delta)} \;=\; \vartheta \ln \frac{1}{\delta}.$$

■

3. The Main Theorem and Algorithm

Main result. Before describing the algorithm, let us state the main result of this paper formally.

Theorem 4 *Let $\mathcal{K}$ be the underlying decision set in an online linear optimization instance, such that $\mathcal{K}$ admits an efficiently computable ϑ-self-concordant barrier. Then there exists a polynomial time algorithm for this online linear optimization problem (Algorithm 1 below coupled with the halving procedure of Section 5) whose expected regret is bounded as follows. Let Q_T be the total variation of a cost function sequence in the online linear optimization instance. Then*

$$\mathbf{E}[\mathit{Regret}_T] \;=\; O\left(n\sqrt{\vartheta Q_T \log T} + n\log^2 T + n\vartheta \log(T)\right).$$

This theorem can be used with the well known logarithmic barrier to derive regret bounds for the online-shortest-paths problem and other linearly constrained problems, and of course applicable much more generally.

The non-stochastic multi-armed bandit problem. A case of particular interest, which has been studied most extensively, is the "basic" multi-armed bandit (MAB) problem where in each iteration the learner pulls the arm of one out of n slot machines and obtains an associated reward, assumed to be in the range $[0,1]$. The learner's objective is to minimize his regret, that is, the difference between his cumulative reward and that of the best fixed slot machine in hindsight.

This is a special case of the more general problem considered earlier and corresponds to taking the convex set $\mathcal{K}$ to be the n-dimensional simplex of probability distributions over the arms. Since the n-dimensional simplex admits a simple n-self-concordant barrier, an immediate corollary of our main theorem is:

Corollary 5 *There exists an efficient algorithm for the multi-armed-bandit problem whose expected regret is bounded by*

$$\mathbf{E}[\mathit{Regret}_T] \;=\; O\left(n^2\sqrt{Q_T \log(T)} + n^{1.5}\log^2(T) + n^{2.5}\log(T)\right).$$

The additional factor of $\sqrt{n}$ factor is because our results assume that $\|\mathbf{f}_t\| \leq 1$, and so we need to scale the costs down by $\sqrt{n}$ to apply our bounds. In comparison, the best previously known bounds for this problem is $O(\sqrt{nT})$ (Audibert and Bubeck, 2010). Even though our bound is worse in the dependence on n, the dependence on the parameter which grows, viz. T, is much better.

3.1 Overview of the Algorithm

The underlying scheme of our algorithm follows the recent approach of Abernethy et al. (2008), who use the Follow-The-Regularized-Leader (FTRL) methodology with self concordant barrier functions as a regularization (see also exposition in Hazan and Kale 2008). At the top level, at every iteration this algorithm simply chooses the point that would have minimized the total cost so far, including an additional regularization cost function $\mathcal{R}(\mathbf{x})$, that is, we predict with the point

$$\mathbf{x}_t = \arg\min_{\mathcal{K}} \left[\eta \sum_{\tau=1}^{t-1} \tilde{\mathbf{f}}_\tau^\top(\mathbf{x}) + \mathcal{R}(\mathbf{x}) \right],$$

where η is a learning rate parameter.

Here, $\tilde{\mathbf{f}}_t$ is an estimator for the vector $\mathbf{f}_t$, which is carefully constructed to have low variance. In the full-information setting, when we can simply set $\tilde{\mathbf{f}}_t = \mathbf{f}_t$, such an algorithm can be shown to achieve low regret (see exposition in Abernethy et al., 2008 and references therein). In the bandit setting, a variety of "one-point-gradient-estimators" are used (Flaxman et al., 2005; Abernethy et al., 2008) which produce an unbiased estimator $\tilde{\mathbf{f}}_t$ of $\mathbf{f}_t$ by evaluating $\mathbf{f}_t$ at just one point.

In order to obtain variation based bounds on the regret, we modify the unbiased estimators of previous approaches by incorporating our experience with previous cost vector as a "prior belief" on the upcoming cost vector. Essentially, we produce an unbiased estimator of the *difference between the average cost vector in the past and the current cost vector*.

This brings out the issue that the past cost vectors are unfortunately also unknown. However, since we had many opportunities to learn about the past and it is an aggregate of many functions, our knowledge about the past cumulative cost vector is much better than the knowledge of any one cost vector in particular. We denote by $\tilde{\mu}_t$ our estimator of $\frac{1}{t}\sum_{\tau=1}^{t}\mathbf{f}_\tau$. The straightforward way of maintaining this estimator would be to average all previous estimators $\tilde{\mathbf{f}}_t$. However, this estimator is far from being sufficiently accurate for our purposes.

Instead, we use the reservoir sampling idea of Section 2.1 to construct this $\tilde{\mu}_t$. For each coordinate $i \in [n]$, we maintain a reservoir of size k, $S_{i,1}, S_{i,2}, \ldots, S_{i,k}$. The estimator for $\mu_t(i)$ is then $\tilde{\mu}_t(i) = \frac{1}{k}\sum_{j=1}^{k} S_{i,j}$. The first nk rounds we devote to initialize these reservoirs with samples from the stream. This increases the overall regret of our algorithm by a constant of nk.

Our current approach is to use separate exploration steps in order to construct $\tilde{\mu}_t$. While it is conceivable that there are more efficient methods of integrating exploration and exploitation, as done by the algorithm in the other iterations, reservoir sampling turns out to be extremely efficient and incurs only a logarithmic penalty in regret.

The general scheme is given in Algorithm 1. It is composed of exploration steps, called SIMPLEXSAMPLE steps, and exploration-exploitation steps, called ELLIPSOIDSAMPLE steps. Note that we use the notation $\mathbf{y}_t$ for the actual point in $\mathcal{K}$ chosen by the algorithm in either of these steps.

It remains to precisely state the SIMPLEXSAMPLE and ELLIPSOIDSAMPLE procedures. The SIMPLEXSAMPLE procedure is the simpler of the two. It essentially performs reservoir sampling on all the coordinates with a reservoir of size k. The initial nk time iterations are used to initialize this reservoir to have correct expectation (i.e., in these iterations we sample with probability one and fill all buckets S_{ij}), and incur an additional additive regret of at most nk.

Now, the SIMPLEXSAMPLE procedure is invoked with probability $\frac{nk}{t}$ for any time period $t > nk$. Once invoked, it samples a coordinate $i_t \in [n]$ with the uniform distribution. The point $\mathbf{y}_t$ chosen

Algorithm 1 Bandit Online Linear Optimization

Input: $\eta > 0$, ϑ-self-concordant $\mathcal{R}$, reservoir size parameter k
Initialization: for all $i \in [n]$, $j \in [k]$, set $S_{i,j} = 0$. Set $\mathbf{x}_1 = \arg\min_{\mathbf{x}\in\mathcal{K}} [\mathcal{R}(\mathbf{x})]$ and $\tilde{\mu}_0 = 0$. Let $\pi : \{1, 2 \ldots, nk\} \to \{1, 2, \ldots, nk\}$ be a random permutation.
for $t = 1$ to T **do**
  Set $r = 1$ with probability $\min\left\{\frac{nk}{t}, 1\right\}$, and 0 with probability $1 - \min\left\{\frac{nk}{t}, 1\right\}$.
  if $r = 1$ **then**
    // SIMPLEXSAMPLE step
    if $t \leq nk$ **then**
      Set $i_t = (\pi(t) \mod n) + 1$.
    else
      Set i_t uniformly at random from $\{1, 2, \ldots, n\}$.
    end if
    Set $\tilde{\mu}_t \leftarrow$ SIMPLEXSAMPLE(i_t).
    Set $\tilde{\mathbf{f}}_t = 0$.
  else
    // ELLIPSOIDSAMPLE step
    Set $\tilde{\mu}_t = \tilde{\mu}_{t-1}$.
    Set $\tilde{\mathbf{f}}_t \leftarrow$ ELLIPSOIDSAMPLE$(\mathbf{x}_t, \tilde{\mu}_t)$.
  end if
  Update $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{K}} \underbrace{\left[\eta \sum_{\tau=1}^{t} \tilde{\mathbf{f}}_\tau^\top \mathbf{x} + \mathcal{R}(\mathbf{x})\right]}_{\Phi_t(\mathbf{x})}$
end for

by the algorithm is the corresponding vertex $\gamma \mathbf{e}_{i_t}$ of the (γ-scaled) n-dimensional simplex (which is assumed to be contained inside of $\mathcal{K}$) to obtain the coordinate $\mathbf{f}_t(i_t)$ as the cost.

It then chooses one of the samples $S_{i_t,1}, S_{i_t,2}, \ldots, S_{i_t,k}$ uniformly at random and replaces it with the value $\mathbf{f}_t(i_t)$, and updates $\tilde{\mu}_t$. This exactly implements the reservoir sampling for each coordinate, and detailed in Algorithm 2.

As for the ELLIPSOIDSAMPLE procedure, it is a modification of the sampling procedure of Abernethy et al. (2008). The point $\mathbf{y}_t$ chosen by the algorithm is uniformly at random chosen from the endpoints of the principal axes of the Dikin ellipsoid $W_1(\mathbf{x}_t)$ centered at $\mathbf{x}_t$. The analysis of Abernethy et al. (2008) already does the hard work of making certain that the ellipsoidal sampling is unbiased and has low variation with respect to the regularization. However, to take advantage of the low variation in the data, we incorporate the previous information in the form of $\tilde{\mu}$. This modification seems to be applicable more generally, not only to the algorithm of Abernethy et al. (2008). However, plugged into this recent algorithm we obtain the best possible regret bounds and also an efficient algorithm.

Before we proceed with the analysis, let us make a small formal claim that our estimates of μ_t are unbiased for $t \geq nk$:

Algorithm 2 SIMPLEXSAMPLE(i_t)

Predict $\mathbf{y}_t = \gamma \mathbf{e}_{i_t}$, that is, the i_t-th standard basis vector scaled by γ.
Observe the cost $\mathbf{f}_t^\top \mathbf{y}_t = \mathbf{f}_t(i_t)$.
if some bucket for i_t is empty **then**
    Set j to the index of the empty bucket.
else
    Set j uniformly at random from $\{1,\dots,k\}$.
end if
Update the sample $S_{i_t,j} = \frac{1}{\gamma}\mathbf{f}_t(i_t)$.
if $t < nk$ **then**
    Return $\tilde{\mu}_t = 0$.
else
    Return $\tilde{\mu}_t$ defined as: $\forall i \in \{1,2,\dots,n\}$, set $\tilde{\mu}_t(i) := \frac{1}{k}\sum_{j=1}^{k} S_{i,j}$.
end if

Algorithm 3 ELLIPSOIDSAMPLE($\mathbf{x}_t, \tilde{\mu}_t$)

Let $\{\mathbf{v}_1,\dots,\mathbf{v}_n\}$ and $\{\lambda_1,\dots,\lambda_n\}$ be the set of orthogonal eigenvectors and eigenvalues of $\nabla^2 \mathcal{R}(\mathbf{x}_t)$.
Choose i_t uniformly at random from $\{1,\dots,n\}$ and $\varepsilon_t = \pm 1$ with probability $1/2$.
Predict $\mathbf{y}_t = \mathbf{x}_t + \varepsilon_t \lambda_{i_t}^{-1/2} \mathbf{v}_{i_t}$.
Observe the cost $\mathbf{f}_t^\top \mathbf{y}_t$.
Return $\tilde{\mathbf{f}}_t$ defined as:

$$\tilde{\mathbf{f}}_t = \tilde{\mu}_t + \tilde{\mathbf{g}}_t$$

Where $\tilde{\mathbf{g}}_t := n\left(\mathbf{f}_t^\top \mathbf{y}_t - \tilde{\mu}_t^\top \mathbf{y}_t\right) \varepsilon_t \lambda_{i_t}^{1/2} \mathbf{v}_{i_t}$.

Claim 1 *For all $t \geq nk$, and for all $i = 1,2,\dots,n$, the reservoir for i, $S_i = \{S_{i,1}, S_{i,2}, \dots, S_{i,k}\}$ is a random subset of size k chosen without replacement from $\{\mathbf{f}_1(i), \mathbf{f}_2(i), \dots, \mathbf{f}_t(i)\}$. Hence, we have $\mathbf{E}[\tilde{\mu}_t] = \mu_t$.*

Proof For $t = nk$ the claim follows because the choice of the random permutation π ensures that the set of times $\{t : (\pi(t) \mod n) + 1 = i\}$ is a random subset of size k chosen without replacement from $\{1,2,\dots,nk\}$.

For $t > nk$ the claim follows from the properties of reservoir sampling, as we show now. This is because SIMPLEXSAMPLE simulates reservoir sampling. We just showed that at time $t = nk$, the claim is true. Then, at time $t = nk+1$ and onwards, reservoir sampling performs select-and-replace with probability $\frac{k}{t}$ (i.e., it selects $\mathbf{f}_t(i)$ with probability $\frac{k}{t}$ and replaces a random element of the previous S_i with it). The algorithm does exactly the same thing: SIMPLEXSAMPLE is invoked with probability $\frac{nk}{t}$, and with probability $\frac{1}{n}$, we have $i_t = i$. Thus, the overall probability of sample-and-replace is $\frac{k}{t}$, exactly as in reservoir sampling. ■

4. Analysis

In this section, we prove a regret bound, in a slightly easier setting where we know an upper bound Q on the total variation Q_T. The main theorem proved here is the following:

Theorem 6 *Let Q be an estimated upper bound on Q_T. Suppose that Algorithm 1 is run with $\eta = \min\left\{\sqrt{\frac{\log T}{n^2 Q}}, \frac{1}{25n}\right\}$ and $k = \log(T)$. Then, if $Q_T \leq Q$, the expected regret is bounded as follows:*

$$\mathbf{E}[Regret_T] \;=\; O\left(n\sqrt{\vartheta Q \log T} + n\log^2(T) + n\vartheta\log(T)\right).$$

Although this bound requires an estimate of the total variation, we show in the Section 5 how to remove this dependence, thereby proving Theorem 4. In this section we sketch the simpler proof of Theorem 6 and give precise proofs of the main lemmas involved.

Proof For clarity, we present the proof as a series of lemmas whose complete proofs appear after this current proof.

We first relate the expected regret of Algorithm 1 which plays the points $\mathbf{y}_t$, for $t = 1, 2, \ldots$ with the $\mathbf{f}_t$ cost vectors to the expected regret of another algorithm that plays the points $\mathbf{x}_t$ with the $\tilde{\mathbf{f}}_t$ cost vectors.

Lemma 7 *For any $\mathbf{u} \in \mathcal{K}$,*

$$\mathbf{E}\left[\sum_{t=1}^{T} \mathbf{f}_t^\top(\mathbf{y}_t - \mathbf{u})\right] \;\leq\; \mathbf{E}\left[\sum_{t=1}^{T} \tilde{\mathbf{f}}_t^\top(\mathbf{x}_t - \mathbf{u})\right] + 2n\log^2(T).$$

Intuitively, this bound holds since in every ELLIPSOIDSAMPLE step, the expectation of $\tilde{\mathbf{f}}_t$ and $\mathbf{y}_t$ (conditioned on all previous randomization) are $\mathbf{f}_t$ and $\mathbf{x}_t$ respectively, the expected costs for both algorithms is the same in such rounds. In the SIMPLEXSAMPLE steps, we have $\tilde{\mathbf{f}}_t = 0$ and we can bound $|\mathbf{f}_t^\top(\mathbf{y}_t - \mathbf{u})|$ by 2. The expected number of such steps is $O(nk\log(T)) = O(n\log^2(T))$, which yields the extra additive term.

We therefore turn to bounding $\sum_{t=1}^{T} \tilde{\mathbf{f}}_t^\top(\mathbf{x}_t - \mathbf{u})$. For this, we apply standard techniques (originally due to Kalai and Vempala 2005) which bounds the regret of any follow-the-leader type algorithm by terms which depend on the stability of the algorithm, measured by how close the successive predictions $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$ are:

Lemma 8 *For any sequence of cost vectors $\tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_T \in \mathbb{R}^n$, the FTRL algorithm with a ϑ-self concordant barrier $\mathcal{R}$ has the following regret guarantee: for any $\mathbf{u} \in \mathcal{K}$, we have*

$$\sum_{t=1}^{T} \tilde{\mathbf{f}}_t^\top(\mathbf{x}_t - \mathbf{u}) \;\leq\; \sum_{t=1}^{T} \tilde{\mathbf{f}}_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1}) + \frac{2}{\eta}\vartheta\log T.$$

We now turn to bounding the term $\tilde{\mathbf{f}}_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1})$. The following main lemma gives such bounds, and forms the main part of the theorem. We go into detail of its proof in the next section, as it contains the main new ideas.

Lemma 9 *Let t be an* ELLIPSOIDSAMPLE *step. Then we have*

$$\tilde{\mathbf{f}}_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1}) \;\leq\; 64\eta n^2\|\mathbf{f}_t - \mu_t\|^2 + 64\eta n^2\|\mu_t - \tilde{\mu}_t\|^2 + 2\mu_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1}). \tag{4}$$

A similar but much easier statement can be made for SIMPLEXSAMPLE steps. Trivially, since we set $\tilde{\mathbf{f}}_t = 0$ in such steps, we have $\mathbf{x}_t = \mathbf{x}_{t+1}$. Thus, we have

$$\tilde{\mathbf{f}}_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1}) \;=\; 0 \;=\; 2\mu_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1}).$$

By adding the non-negative term $64\eta n^2\|\mathbf{f}_t - \mu_t\|^2$, we get that for any SIMPLEXSAMPLE step t, we have

$$\tilde{\mathbf{f}}_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1}) \;\le\; 64\eta n^2\|\mathbf{f}_t - \mu_t\|^2 + 2\mu_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1}). \tag{5}$$

Let T_E be the set of all ELLIPSOIDSAMPLE steps t. Summing up either (4) or (5), as the case may be, over all time periods t we get

$$\sum_{t=1}^{T}\tilde{\mathbf{f}}_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1}) \;\le\; 64\eta n^2\sum_{t=1}^{T}\|\mathbf{f}_t - \mu_t\|^2 + 64\eta n^2\sum_{t\in T_E}\|\mu_t - \tilde{\mu}_t\|^2 + 2\sum_{t=1}^{T}\mu_t(\mathbf{x}_t - \mathbf{x}_{t+1}) \tag{6}$$

We bound each term of the inequality (6) above separately. The first term can be easily bounded by the total variation, even though it is the sum of squared deviations from changing means. Essentially, the means don't change very much as time goes on.

Lemma 10 $\sum_{t=1}^{T}\|\mathbf{f}_t - \mu_t\|^2 \;\le\; Q_T$.

The second term, in expectation, is just the variance of the estimators $\tilde{\mu}_t$ of μ_t, which can be bounded in terms of the size of the reservoir and the total variation (see Lemma 2).

Lemma 11 $\mathbf{E}\left[\sum_{t\in T_E}\|\mu_t - \tilde{\mu}_t\|^2\right] \le \frac{\log T}{k}Q_T$.

The third term can be bounded by the sum of successive differences of the means, which, in turn, can be bounded the logarithm of the total variation.

Lemma 12 $\sum_{t=1}^{T}\mu_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1}) \;\le\; 2\log(Q_T+1) + 4$.

Let $Q \ge Q_T$ be a given upper bound. Plugging the bounds from Lemmas 10, 11, and 12 into (6), and using the value $k = \log(T)$, we obtain

$$\sum_{t=1}^{T}\mathbf{f}_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1}) \;\le\; 128\eta n^2 Q + 4\log(Q_T+1) + 8.$$

where we will choose $\eta \ge \frac{\log(Q_T+1)}{8n^2Q}$ so that $\log(Q_T+1) \le 8\eta n^2 Q$. Hence, via Lemmas 8 and 7, we have for any $\mathbf{u} \in \mathcal{K}$,

$$\mathbf{E}\left[\sum_{t=1}^{T}\mathbf{f}_t^\top(\mathbf{y}_t - \mathbf{u})\right] \;\le\; 128\eta n^2 Q + \frac{2\vartheta}{\eta}\log T + 2n\log^2(T) + 4\log(Q_T+1) + 8.$$

Now, choosing $\eta = \min\left\{\sqrt{\frac{\vartheta\log(T)}{n^2Q}}, \frac{1}{25n}\right\}$, for the upper bound $Q \ge Q_T$, and we get the following regret bound:

$$\mathbf{E}\left[\sum_{t=1}^{T}\mathbf{f}_t^\top(\mathbf{y}_t - \mathbf{u})\right] \;\le\; O\left(n\sqrt{\vartheta Q\log T} + n\vartheta\log(T) + n\log^2(T)\right).$$

Here, we absorb the lower order terms $4\log(Q_T+1)+8$ in the other terms using the $O(\cdot)$ notation. The restriction that $\eta \le \frac{1}{25n}$ arises from the proof of Lemma 13 below. ■

4.1 Proof of Main Lemmas

Proof [Lemma 7]
Let t be an ELLIPSOIDSAMPLE step. We first show that $\mathbf{E}[\tilde{\mathbf{f}}_t] = \mathbf{f}_t$. We condition on all the randomness prior to this step, thus, $\tilde{\mu}_t$ is fixed. In the following, $\mathbf{E}_t$ denotes this conditional expectation. Now, condition on the choice i_t and average over the choice of ε_t:

$$\mathbf{E}_t[\tilde{\mathbf{g}}_t | i_t] \;=\; \sum_{\varepsilon_t \in \{1,-1\}} \frac{1}{2} n \left((\mathbf{f}_t - \tilde{\mu}_t)^\top (\mathbf{x}_t + \varepsilon_t \lambda_{i_t}^{-1/2} \mathbf{v}_{i_t}) \right) \lambda_{i_t}^{1/2} \varepsilon_t \mathbf{v}_{i_t} = n((\mathbf{f}_t - \tilde{\mu}_t)^\top \mathbf{v}_{i_t}) \mathbf{v}_{i_t}.$$

Hence,

$$\mathbf{E}_t[\tilde{\mathbf{g}}_t] = \sum_{i=1}^{n} \frac{1}{n} \cdot n((\mathbf{f}_t - \tilde{\mu}_t)^\top \mathbf{v}_i) \mathbf{v}_i = \mathbf{f}_t - \tilde{\mu}_t,$$

since the $\mathbf{v}_i$ form an orthonormal basis. Thus, $\mathbf{E}_t[\tilde{\mathbf{f}}_t] = \mathbf{E}_t[\tilde{\mathbf{g}}_t] + \tilde{\mu}_t = \mathbf{f}_t$.

Furthermore, it is easy to see that $\mathbf{E}_t[\mathbf{y}_t] = \mathbf{x}_t$, since $\mathbf{y}_t$ is drawn from a symmetric distribution centered at $\mathbf{x}_t$ (namely, the uniform distribution on the endpoints of the principal axes of the Dikin ellipsoid centered at $\mathbf{x}_t$). Thus, we conclude that

$$\mathbf{E}_t[\mathbf{f}_t^\top (\mathbf{y}_t - \mathbf{u})] \;=\; \mathbf{f}_t^\top (\mathbf{x}_t - \mathbf{u}) \;=\; \mathbf{E}_t[\tilde{\mathbf{f}}_t^\top (\mathbf{x}_t - \mathbf{u})],$$

and hence, taking expectation over all the randomness, we have

$$\mathbf{E}[\mathbf{f}_t^\top (\mathbf{y}_t - \mathbf{u})] \;=\; \mathbf{E}[\tilde{\mathbf{f}}_t^\top (\mathbf{x}_t - \mathbf{u})].$$

Now, let t be a SIMPLEXSAMPLE step or alternatively $t \le nk$. In this case, we have $|\mathbf{f}_t^\top (\mathbf{y}_t - \mathbf{u})\| \le \|\mathbf{f}_t\| \|\mathbf{y}_t - \mathbf{u}\| \le 2$, and $\tilde{\mathbf{f}}_t^\top (\mathbf{x}_t - \mathbf{u}) = 0$ since $\tilde{\mathbf{f}}_t = 0$. Thus,

$$\mathbf{E}[\mathbf{f}_t^\top (\mathbf{y}_t - \mathbf{u})] \;\le\; \mathbf{E}[\tilde{\mathbf{f}}_t^\top (\mathbf{x}_t - \mathbf{u})] + 2.$$

Overall, if X is the number of SIMPLEXSAMPLE sampling steps or initialization steps, we have

$$\mathbf{E}[\mathbf{f}_t^\top (\mathbf{y}_t - \mathbf{u})] \;\le\; \mathbf{E}_t[\tilde{\mathbf{f}}_t^\top (\mathbf{x}_t - \mathbf{u})] + 2\mathbf{E}[X].$$

Finally, using the fact that $\mathbf{E}[X] = nk + \sum_{t=nk+1}^{T} \frac{nk}{t} \le nk(\log(T)+1) \le 2n\log^2(T)$, the proof is complete. ■

Proof [Lemma 8]

By Lemma 15 (see Section 4.2) applied to the sequence $\{\mathbf{x}_t\}$ as defined in (19), for any $\mathbf{u} \in \mathcal{K}$

$$\sum_{t=1}^{T} \tilde{\mathbf{f}}_t^\top (\mathbf{x}_t - \mathbf{u}) \;\le\; \sum_{t=1}^{T} \tilde{\mathbf{f}}_t^\top (\mathbf{x}_t - \mathbf{x}_{t+1}) + \frac{1}{\eta} [\mathcal{R}(\mathbf{u}) - \mathcal{R}(\mathbf{x}_1)].$$

By Lemma 3, there exists a vector $\mathbf{u}_1 \in \mathcal{K}_\delta \subseteq \mathcal{K}$ for $\delta = \frac{1}{T}$, such that $\|\mathbf{u}_1 - \mathbf{u}\| \leq \frac{2}{T}$ and in addition, $\mathcal{R}(\mathbf{u}_1) - \mathcal{R}(\mathbf{x}_1) \leq \vartheta \log(T)$. Hence,

$$\begin{aligned}
\sum_{t=1}^{T} \tilde{\mathbf{f}}_t^\top (\mathbf{x}_t - \mathbf{u}) \;&\leq\; \sum_{t=1}^{T} \tilde{\mathbf{f}}_t^\top (\mathbf{x}_t - \mathbf{u}_1) \;+ \sum_{t=1}^{T} \tilde{\mathbf{f}}_t^\top (\mathbf{u}_1 - \mathbf{u}) \\
&\leq\; \sum_{t=1}^{T} \tilde{\mathbf{f}}_t^\top (\mathbf{x}_t - \mathbf{x}_{t+1}) + \frac{1}{\eta}[\mathcal{R}(\mathbf{u}_1) - \mathcal{R}(\mathbf{x}_1)] + \sum_{t=1}^{T} \|\mathbf{f}_t\| \|\mathbf{u}_1 - \mathbf{u}\| \\
&\leq\; \sum_{t=1}^{T} \tilde{\mathbf{f}}_t^\top (\mathbf{x}_t - \mathbf{x}_{t+1}) + \frac{\vartheta}{\eta} \log T + \sum_{t=1}^{T} \frac{2}{T} \\
&\leq\; \sum_{t=1}^{T} \tilde{\mathbf{f}}_t^\top (\mathbf{x}_t - \mathbf{x}_{t+1}) + \frac{2\vartheta}{\eta} \log T.
\end{aligned}$$

In the last step, we upper bound $\sum_{t=1}^{T} \frac{2}{T} \leq \frac{\vartheta}{\eta} \log T$, which is valid for $\eta < 1/4$, say. ■

Now we turn to proving Lemma 9. We first develop some machinery to assist us. Lemmas 13 and 14 are essentially generalizations of similar lemmas from Abernethy et al. (2008) to the case in which we have both sampling and ellipsoidal steps.

Lemma 13 *For any time period $t \geq nk$, the next minimizer $\mathbf{x}_{t+1}$ is "close" to $\mathbf{x}_t$:*

$$\mathbf{x}_{t+1} \in W_{\frac{1}{2}}(\mathbf{x}_t).$$

Proof If t is a SIMPLEXSAMPLE step, then $\mathbf{x}_t = \mathbf{x}_{t+1}$ and the lemma is trivial. So assume that t is an ELLIPSOIDSAMPLE step. Now, recall that

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{K}} \Phi_t(\mathbf{x}) \quad \text{and} \quad \mathbf{x}_t = \arg\min_{\mathbf{x} \in \mathcal{K}} \Phi_{t-1}(\mathbf{x}),$$

where $\Phi_t(\mathbf{x}) = \eta \sum_{s=1}^{t} \tilde{\mathbf{f}}_t^\top \mathbf{x} + \mathcal{R}(\mathbf{x})$. Since the barrier function $\mathcal{R}$ goes to infinity as we get close to the boundary, the points $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$ are both in the interior of $\mathcal{K}$. We now show that all points on the boundary of $W_{\frac{1}{2}}(\mathbf{x}_t)$ have higher Φ_t value than $\Phi_t(\mathbf{x}_t)$, and since $\mathbf{x}_{t+1}$ is the minimizer of the strictly convex function Φ_t, we conclude that $\mathbf{x}_{t+1}$ must lie in the interior of $W_{\frac{1}{2}}(\mathbf{x}_t)$.

First, note that since $\mathbf{x}_t$ is in the interior of $\mathcal{K}$, the first order optimality condition gives $\nabla \Phi_{t-1}(\mathbf{x}_t) = 0$, and we conclude that $\nabla \Phi_t(\mathbf{x}_t) = \eta \tilde{\mathbf{f}}_t$. Now consider any point in $\mathbf{z}$ on the boundary of $W_{\frac{1}{2}}(\mathbf{x}_t)$, that is, $\mathbf{y} = \mathbf{x}_t + \mathbf{h}$ for some vector $\mathbf{h}$ such that $\|\mathbf{h}\|_{\mathbf{x}_t} = \frac{1}{2}$. Using the multi-variate Taylor expansion, we get

$$\Phi_t(\mathbf{y}) \;=\; \Phi_t(\mathbf{x}_t + \mathbf{h}) \;=\; \Phi_t(\mathbf{x}_t) + \nabla \Phi_t(\mathbf{x}_t)^\top \mathbf{h} + \frac{1}{2} \mathbf{h}^\top \nabla^2 \Phi_t(\xi) \mathbf{h} \;=\; \Phi_t(\mathbf{x}_t) + \eta \tilde{\mathbf{f}}_t^\top \mathbf{h} + \frac{1}{2} \mathbf{h}^\top \nabla^2 \Phi_t(\xi) \mathbf{h} \tag{7}$$

for some ξ on the line segment between $\mathbf{x}_t$ and $\mathbf{x}_t + \mathbf{h}$. This latter fact also implies that $\|\xi - \mathbf{x}_t\|_{\mathbf{x}_t} \leq \|\mathbf{h}\|_{\mathbf{x}_t} \leq \frac{1}{2}$. Hence, by (3),

$$\nabla^2 \mathcal{R}(\xi) \;\succeq\; (1 - \|\xi - \mathbf{x}_t\|_{\mathbf{x}_t})^2 \nabla^2 \mathcal{R}(\mathbf{x}_t) \;\succeq\; \frac{1}{4} \nabla^2 \mathcal{R}(\mathbf{x}_t).$$

Thus $\mathbf{h}^\top \nabla^2 \mathcal{R}(\xi)\mathbf{h} \geq \frac{1}{4}\|\mathbf{h}\|_{\mathbf{x}_t} = \frac{1}{8}$. Next, we bound $|\tilde{\mathbf{f}}_t^\top \mathbf{h}|$ as follows:

$$|\tilde{\mathbf{f}}_t^\top \mathbf{h}| \leq \|\tilde{\mathbf{f}}_t\|_{\mathbf{x}_t}^\star \|\mathbf{h}\|_{\mathbf{x}_t} \leq \frac{1}{2}\|\tilde{\mathbf{f}}_t\|_{\mathbf{x}_t}^\star.$$

Claim 2 $\|\tilde{\mathbf{f}}_t\|_{\mathbf{x}_t}^\star \leq 3n$.

Proof We have $\tilde{\mathbf{f}}_t = \tilde{\mu}_t + \tilde{\mathbf{g}}_t$, where $\tilde{\mathbf{g}}_t = n\left((\mathbf{f}_t - \tilde{\mu}_t)^\top \mathbf{y}_t\right)\varepsilon_t \lambda_{i_t}^{1/2} \mathbf{v}_{i_t}$. We have

$$\begin{aligned}\|\tilde{\mathbf{g}}_t\|_{\mathbf{x}_t}^{\star 2} &= \left[n\left((\mathbf{f}_t - \tilde{\mu}_t)^\top \mathbf{y}_t\right)\varepsilon_t \lambda_{i_t}^{1/2}\mathbf{v}_{i_t}\right]^\top [\nabla^2 \mathcal{R}(\mathbf{x}_t)]^{-1}\left[n\left((\mathbf{f}_t - \tilde{\mu}_t)^\top \mathbf{y}_t\right)\varepsilon_t \lambda_{i_t}^{1/2}\mathbf{v}_{i_t}\right] \\ &= n^2\left((\mathbf{f}_t - \tilde{\mu}_t)^\top \mathbf{y}_t\right)^2,\end{aligned}$$

since $\mathbf{v}_{i_t}^\top [\nabla^2 \mathcal{R}(\mathbf{x}_t)]^{-1} \mathbf{v}_{i_t} = 1/\lambda_{i_t}$. Hence,

$$\|\tilde{\mathbf{f}}_t\|_{\mathbf{x}_t}^\star \leq \|\tilde{\mu}_t\|_{\mathbf{x}_t}^\star + \|\tilde{\mathbf{g}}_t\|_{\mathbf{x}_t}^\star \leq \|\tilde{\mu}_t\| + n|(\mathbf{f}_t - \tilde{\mu}_t)^\top \mathbf{y}_t| \leq 3n,$$

since $\|\tilde{\mu}_t\|_{\mathbf{x}_t}^\star \leq \|\tilde{\mu}_t\| \leq 1$. We also used the facts that $\|\mathbf{y}_t\| \leq 1$ and $\|\mathbf{f}_t - \tilde{\mu}_t\| \leq 2$. ■

Hence, from (7) we get

$$\Phi_t(\mathbf{y}) \geq \Phi_t(\mathbf{x}_t) - \eta \cdot \frac{3n}{2} + \frac{1}{16} > \Phi_t(\mathbf{x}_t),$$

since $\eta \leq \frac{1}{25n}$. This concludes the proof that all boundary points of $W_{\frac{1}{2}}(\mathbf{x}_t)$ have higher Φ_t value than $\Phi_t(\mathbf{x}_t)$. ■

Lemma 14 *For any time period $t \geq nk$, we have*

$$\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_{\mathbf{x}_t}^2 \leq 4\eta \tilde{\mathbf{f}}_t^\top (\mathbf{x}_t - \mathbf{x}_{t+1}).$$

Proof Applying the Taylor series expansion to the function Φ_t around the point $\mathbf{x}_t$, we get that for some point $\mathbf{z}_t$ on the line segment joining $\mathbf{x}_t$ to $\mathbf{x}_{t+1}$, we have

$$\Phi_t(\mathbf{x}_t) = \Phi_t(\mathbf{x}_{t+1}) + \nabla\Phi_t(\mathbf{x}_{t+1})^\top(\mathbf{x}_t - \mathbf{x}_{t+1}) + (\mathbf{x}_{t+1} - \mathbf{x}_t)^\top \nabla^2 \Phi_t(\mathbf{z}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t) = \Phi_t(\mathbf{x}_{t+1}) + \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\mathbf{z}_t}^2,$$

because $\nabla\Phi_t(\mathbf{x}_{t+1}) = 0$ since $\mathbf{x}_{t+1}$, the minimizer of Φ_t, is in the interior of $\mathcal{K}$. We also used the fact that $\nabla^2\Phi_t(\mathbf{z}_t) = \nabla^2\mathcal{R}(\mathbf{z}_t)$. Thus, we have

$$\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\mathbf{z}_t}^2 = \Phi_t(\mathbf{x}_t) - \Phi_t(\mathbf{x}_{t+1}) = \Phi_{t-1}(\mathbf{x}_t) - \Phi_{t-1}(\mathbf{x}_{t+1}) + \eta\tilde{\mathbf{f}}_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1}) \leq \eta\tilde{\mathbf{f}}_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1}),$$

since $\mathbf{x}_t$ is the minimizer of Φ_{t-1} in $\mathcal{K}$. It remains to show that $\frac{1}{4}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\mathbf{x}_t}^2 \leq \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\mathbf{z}_t}^2$, for which it suffices to show $\frac{1}{4}\nabla^2\mathcal{R}(\mathbf{x}_t) \preceq \nabla^2\mathcal{R}(\mathbf{z}_t)$.

By Lemma 13 we have $\mathbf{x}_{t+1} \in W_{1/2}(\mathbf{x}_t)$, and hence $\mathbf{z}_t \in W_{1/2}(\mathbf{x}_t)$ (since $\mathbf{z}_t$ is on the line segment between $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$). Therefore, using (3) we have $\frac{1}{4}\nabla^2\mathcal{R}(\mathbf{x}_t) \preceq \nabla^2\mathcal{R}(\mathbf{z}_t)$ as required. ■

Proof [Lemma 9]
First, we have

$$\begin{aligned}(\tilde{\mathbf{f}}_t - \mu_t)^\top(\mathbf{x}_t - \mathbf{x}_{t+1}) &\le \|\tilde{\mathbf{f}}_t - \mu_t\|^\star_{\mathbf{x}_t} \cdot \|\mathbf{x}_t - \mathbf{x}_{t+1}\|_{\mathbf{x}_t} && \text{(by (1))}\\ &\le \|\tilde{\mathbf{f}}_t - \mu_t\|^\star_{\mathbf{x}_t} \cdot \sqrt{4\eta\tilde{\mathbf{f}}_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1})} && \text{(Lemma 14)}\\ &\le 2\eta\|\tilde{\mathbf{f}}_t - \mu_t\|^{\star 2}_{\mathbf{x}_t} + \frac{1}{2}\tilde{\mathbf{f}}_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1}).\end{aligned}$$

The last inequality follows using the fact that $ab \le \frac{1}{2}(a^2 + b^2)$ for real numbers a, b. Simplifying, we get that

$$\begin{aligned}\tilde{\mathbf{f}}_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1}) &\le 4\eta\|\tilde{\mathbf{f}}_t - \mu_t\|^{\star 2}_{\mathbf{x}_t} + 2\mu_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1})\\ &\le 8\eta\left(\|\tilde{\mathbf{f}}_t - \tilde{\mu}_t\|^{\star 2}_{\mathbf{x}_t} + \|\mu_t - \tilde{\mu}_t\|^{\star 2}_{\mathbf{x}_t}\right) + 2\mu_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1})\\ &\le 32\eta\left(\|\tilde{\mathbf{g}}_t\|^{\star 2}_{\mathbf{x}_t} + \|\mu_t - \tilde{\mu}_t\|^2\right) + 2\mu_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1}).\end{aligned}$$

The last inequality is because $\|\cdot\|^\star_{\mathbf{x}} \le 2\|\cdot\|$ from (2) and the assumption that $\mathcal{K}$ is contained inside the unit ball.

Using the definition of $\tilde{\mathbf{g}}_t$ from Algorithm 3, we get that

$$\begin{aligned}\|\tilde{\mathbf{g}}_t\|^{\star 2}_{\mathbf{x}_t} &= n^2\left((\mathbf{f}_t - \tilde{\mu}_t)^\top\mathbf{y}_t\right)^2\lambda_{i_t}\cdot\left(\mathbf{v}_{i_t}^\top[\nabla^2\mathcal{R}(\mathbf{x}_t)]^{-1}\mathbf{v}_{i_t}\right)\\ &= n^2\left((\mathbf{f}_t - \tilde{\mu}_t)^\top\mathbf{y}_t\right)^2\\ &\le n^2\|\mathbf{f}_t - \tilde{\mu}_t\|^2\\ &\le 2n^2[\|\mathbf{f}_t - \mu_t\|^2 + \|\mu_t - \tilde{\mu}_t\|^2].\end{aligned}$$

The first inequality follows by applying Cauchy-Schwarz and using the fact that $\|\mathbf{y}_t\| \le 1$. Plugging this bound into the previous bound we conclude that

$$\tilde{\mathbf{f}}_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1}) \le 64\eta n^2\|\mathbf{f}_t - \mu_t\|^2 + 64\eta n^2\|\mu_t - \tilde{\mu}_t\|^2 + 2\mu_t^\top(\mathbf{x}_t - \mathbf{x}_{t+1}).$$

■

Proof [Lemma 10]

Recall that $\mu_t = \arg\min_\mu \sum_{\tau=1}^t \|\mathbf{f}_\tau - \mu\|^2$. As a first step, we show that

$$\sum_{\tau=1}^T \|\mathbf{f}_t - \mu_t\|^2 \le \sum_{\tau=1}^T \|\mathbf{f}_t - \mu_T\|^2.$$

This is proved by induction on t. For $T = 1$ the inequality is trivial; we actually have equality. Assume correctness for some $T-1$. Moving to T, we have

$$\begin{aligned}
\sum_{t=1}^{T} \|\mathbf{f}_t - \mu_t\|^2 &= \sum_{t=1}^{T-1} \|\mathbf{f}_t - \mu_t\|^2 + \|\mathbf{f}_T - \mu_T\|^2 \\
&\leq \sum_{t=1}^{T-1} \|\mathbf{f}_t - \mu_{T-1}\|^2 + \|\mathbf{f}_T - \mu_T\|^2 && \text{(By inductive hypothesis)} \\
&\leq \sum_{t=1}^{T-1} \|\mathbf{f}_t - \mu_T\|^2 + \|\mathbf{f}_T - \mu_T\|^2 && (\mu_{T-1} = \arg\min_{\mathbf{x}} \textstyle\sum_{t=1}^{T-1} \|\mathbf{f}_t - \mathbf{x}\|^2) \\
&= \sum_{t=1}^{T} \|\mathbf{f}_t - \mu_T\|^2.
\end{aligned}$$

Hence,

$$\sum_{\tau=1}^{T} \|\mathbf{f}_t - \mu_t\|^2 \leq \sum_{\tau=1}^{T} \|\mathbf{f}_t - \mu\|^2 = Q_T.$$

■

Proof [Lemma 11]
Any ELLIPSOIDSAMPLE step t must have $t \geq nk$, so by Claim 1 the algorithm exactly implements reservoir sampling with a reservoir of size k for each of the n coordinates.

Now, for any coordinate i, $\tilde{\mu}_t(i)$ is the average of a k samples chosen *without* replacement from F_t. Thus, we have $\mathbf{E}[\tilde{\mu}_t(i)] = \mu_t(i)$, and hence $\mathbf{E}[(\tilde{\mu}_t(i) - \mu_t(i))^2] = \mathrm{VAR}[\tilde{\mu}_t(i)]$.

Now consider another estimator $\nu_t(i)$, which averages k samples chosen *with* replacement from F_t. It is a well-known statistical fact (see, e.g., Rice, 2001) that $\mathrm{VAR}[\tilde{\mu}_t(i)] \leq \mathrm{VAR}[\nu_t(i)]$. Thus, we bound $\mathrm{VAR}[\nu_t(i)]$ instead.

Suppose $t > nk$. Let $\mu = \frac{1}{T}\sum_{t=1}^{T} \mathbf{f}_t$. Since $\mathbf{E}[\nu_t(i)] = \mu_t(i)$, we have

$$\begin{aligned}
\mathrm{VAR}[\nu_t(i)] = \mathbf{E}[(\nu_t(i) - \mu_t(i))^2] &\leq \mathbf{E}[(\nu_t(i) - \mu(i))^2] \\
&= \frac{1}{k}\sum_{\tau=1}^{t} \frac{1}{t}(\mathbf{f}_\tau(i) - \mu(i))^2
\end{aligned}$$

Summing up over all coordinates i, we get

$$\mathbf{E}[\|\tilde{\mu}_t - \mu_t\|^2] \leq \sum_i \mathrm{VAR}[\nu_t(i)] \leq \frac{1}{kt} Q_t \leq \frac{1}{kt} Q_T.$$

Summing up over all ELLIPSOIDSAMPLE steps t, we get

$$\mathbf{E}\left[\sum_{t \in T_E} \|\tilde{\mu}_t - \mu_t\|^2\right] \leq \sum_{t \in T_E} \frac{1}{kt} Q_T \leq \frac{\log(T)}{k} Q_T.$$

■

Proof [Lemma 12]
We have

$$\sum_{t=1}^{T} \mu_t^\top (\mathbf{x}_t - \mathbf{x}_{t+1}) = \sum_{t=1}^{T} \mathbf{x}_{t+1}(\mu_{t+1} - \mu_t) + \mu_1 \mathbf{x}_1 - \mathbf{x}_{T+1}\mu_{T+1}.$$

Thus, since $\|\mathbf{x}_t\| \leq 1$ and $\|\mu_t\| \leq 1$, we have

$$\sum_{t=1}^{T} \mu_t^\top (\mathbf{x}_t - \mathbf{x}_{t+1}) \;\leq\; \sum_{t=2}^{T} \|\mu_{t+1} - \mu_t\| + 4.$$

To proceed, we appeal to Lemma 16 (see Section 4.2), and apply it for $x_t := \|\mathbf{f}_t - \mu_t\|$. Let $\mu = \frac{1}{T}\sum_{t=1}^{T} \mathbf{f}_t$. Arguing as in Lemma 10, we have

$$\sum_t x_t^2 \;=\; \sum_{t=1}^{T} \|\mathbf{f}_t - \mu_t\|^2 \;\leq\; \sum_{t=1}^{T} \|\mathbf{f}_t - \mu\|^2 \;\leq\; Q_T.$$

Notice that

$$\mu_t - \mu_{t-1} = \frac{1}{t}\sum_{\tau=1}^{t} \mathbf{f}_\tau - \frac{1}{t-1}\sum_{\tau=1}^{t-1} \mathbf{f}_\tau = \frac{1}{t}\mathbf{f}_t + (\frac{1}{t} - \frac{1}{t-1})\sum_{\tau=1}^{t-1} \mathbf{f}_\tau = \frac{1}{t}(\mathbf{f}_t - \mu_{t-1}).$$

Hence,

$$\|\mu_t - \mu_{t-1}\| \;=\; \frac{1}{t}\|\mathbf{f}_t - \mu_{t-1}\| \;\leq\; \frac{1}{t}x_t + \frac{1}{t}\|\mu_t - \mu_{t-1}\|,$$

from which we conclude that for all $t \geq 2$ we have $\|\mu_t - \mu_{t-1}\| \leq \frac{t^{-1}}{1-t^{-1}}x_t \leq \sum_t \frac{2}{t}x_t$. Now, we apply Lemma 16 to conclude that

$$\sum_{t=2}^{T} \|\mu_{t+1} - \mu_t\| + 4 \quad \leq\; 2\log(Q_T + 1) + 4.$$

■

4.2 Auxiliary Lemmas

In this section, we give a number of auxiliary lemmas that are independent of the analysis of the algorithm. These lemmas give useful bounds that are used in the main analysis.

The first lemma gives a general regret bound for any follow-the-regularized-leader style algorithm. The proof of this bound is essentially due to Kalai and Vempala (2005).

Lemma 15 *Consider an online linear optimization instance over a convex set $\mathcal{K}$, with a regularization function $\mathcal{R}$ and a sequence $\{x_t\}$ defined by*

$$\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathcal{K}} \left\{ \sum_{\tau=1}^{t-1} \mathbf{f}_\tau^\top \mathbf{x} + \mathcal{R}(\mathbf{x}) \right\}.$$

For every $\mathbf{u} \in \mathcal{K}$, the sequence $\{\mathbf{x}_t\}$ satisfies the following regret guarantee

$$\sum_{t=1}^{T} \mathbf{f}_t^\top (\mathbf{x}_t - \mathbf{u}) \quad\leq\quad \sum_{t=1}^{T} \mathbf{f}_t^\top (\mathbf{x}_t - \mathbf{x}_{t+1}) + \frac{1}{\eta}[\mathcal{R}(\mathbf{u}) - \mathcal{R}(\mathbf{x}_1)].$$

Proof For convenience, denote by $\mathbf{f}_0 = \frac{1}{\eta}\mathcal{R}$, and assume we start the algorithm from $t = 0$ with an arbitrary $\mathbf{x}_0$. The lemma is now proved by induction on T.

In the base case, for $T = 1$, by definition we have that $\mathbf{x}_1 = \arg\min_{\mathbf{x}}\{\mathcal{R}(\mathbf{x})\}$, and thus $\mathbf{f}_0(\mathbf{x}_1) \leq \mathbf{f}_0(\mathbf{u})$ for all $\mathbf{u}$, thus $\mathbf{f}_0(\mathbf{x}_0) - \mathbf{f}_0(\mathbf{u}) \leq \mathbf{f}_0(\mathbf{x}_0) - \mathbf{f}_0(\mathbf{x}_1)$.

Now assume that that for some $T \geq 1$, we have

$$\sum_{t=0}^{T} \mathbf{f}_t(\mathbf{x}_t) - \mathbf{f}_t(\mathbf{u}) \quad \leq \quad \sum_{t=0}^{T} \mathbf{f}_t(\mathbf{x}_t) - \mathbf{f}_t(\mathbf{x}_{t+1}).$$

We now prove the claimed inequality for $T+1$. Since $\mathbf{x}_{T+2} = \arg\min_{\mathbf{x}}\{\sum_{t=0}^{T+1} \mathbf{f}_t(\mathbf{x})\}$ we have:

$$\begin{aligned}
\sum_{t=0}^{T+1} \mathbf{f}_t(\mathbf{x}_t) - \sum_{t=0}^{T+1} \mathbf{f}_t(\mathbf{u}) \quad &\leq \quad \sum_{t=0}^{T+1} \mathbf{f}_t(\mathbf{x}_t) - \sum_{t=0}^{T+1} \mathbf{f}_t(\mathbf{x}_{T+2}) \\
&= \quad \sum_{t=0}^{T} (\mathbf{f}_t(\mathbf{x}_t) - \mathbf{f}_t(\mathbf{x}_{T+2})) + \mathbf{f}_{T+1}(\mathbf{x}_{T+1}) - \mathbf{f}_{T+1}(\mathbf{x}_{T+2}) \\
&\leq \quad \sum_{t=0}^{T} (\mathbf{f}_t(\mathbf{x}_t) - \mathbf{f}_t(\mathbf{x}_{t+1})) + \mathbf{f}_{T+1}(\mathbf{x}_{T+1}) - \mathbf{f}_{T+1}(\mathbf{x}_{T+2}) \\
&= \quad \sum_{t=0}^{T+1} \mathbf{f}_t(\mathbf{x}_t) - \mathbf{f}_t(\mathbf{x}_{t+1}).
\end{aligned}$$

In the third line we used the induction hypothesis for $\mathbf{u} = \mathbf{x}_{T+2}$. We conclude that

$$\begin{aligned}
\sum_{t=1}^{T} \mathbf{f}_t(\mathbf{x}_t) - \mathbf{f}_t(\mathbf{u}) \quad &\leq \quad \sum_{t=1}^{T} \mathbf{f}_t(\mathbf{x}_t) - \mathbf{f}_t(\mathbf{x}_{t+1}) + [-\mathbf{f}_0(\mathbf{x}_0) + \mathbf{f}_0(\mathbf{u}) + \mathbf{f}_0(\mathbf{x}_0) - \mathbf{f}_0(\mathbf{x}_1)] \\
&= \quad \sum_{t=1}^{T} \mathbf{f}_t(\mathbf{x}_t) - \mathbf{f}_t(\mathbf{x}_{t+1}) + \frac{1}{\eta}\left[\mathcal{R}(\mathbf{u}) - \mathcal{R}(\mathbf{x}_1)\right].
\end{aligned}$$

■

Lemma 16 *Suppose we have real numbers $x_1, x_2, \ldots, x_T$ such that $0 \leq x_t \leq 1$ and $\sum_t x_t^2 \leq Q$. Then*

$$\sum_{t=1}^{T} \frac{1}{t} x_t \;\leq\; \log(Q+1) + 1.$$

Proof By Lemma 17 below, the values of x_t that maximize $\sum_{t=1}^{T} \frac{1}{t} x_t$ must have the following structure: there is a k such that for all $t \leq k$, we have $x_t = 1$, and for any index $t > k$, we have $x_{k+1}/x_t \geq \frac{1}{k}/\frac{1}{t}$, which implies that $x_t \leq k/t$. We first note that $k \leq Q$, since $Q \geq \sum_{t=1}^{k} x_t^2 = k$. Now, we can bound the value as follows:

$$\sum_{t=1}^{T} \frac{1}{t} x_t \;\leq\; \sum_{t=1}^{k} \frac{1}{t} + \sum_{t=k+1}^{T} \frac{k}{t^2} \;\leq\; \log(k+1) + k \cdot \frac{1}{k} \;\leq\; \log(Q+1) + 1.$$

■

Lemma 17 *Let $a_1 \geq a_2 \geq \ldots a_T > 0$. Then the optimal solution of*

$$\max \sum_i a_i x_i \text{ subject to}$$
$$\forall i: \ 0 \leq x_i \leq 1$$
$$\sum_i x_i^2 \leq Q$$

has the following properties: $x_1 \geq x_2 \geq \ldots x_T$, and for any pair of indices i, j, with $i < j$, either $x_i = 1$, $x_i = 0$ or $x_i/x_j \geq a_i/a_j$.

Proof The fact that in the optimal solution $x_1 \geq x_2 \geq \ldots x_T$ is obvious, since otherwise we could permute the x_i's to be in decreasing order and increase the value.

The second fact follows by the Karush-Kuhn-Tucker (KKT) optimality conditions, which imply the existence of constants $\mu, \lambda_1, \ldots, \lambda_T, \rho_1, \ldots, \rho_T$ for which the optimal solution satisfies

$$\forall i: \ -a_i + 2\mu x_i + \lambda_i + \rho_i = 0.$$

Furthermore, the complementary slackness condition says that the constants λ_i, ρ_i are equal to zero for all indices of the solution which satisfy $x_i \notin \{0, 1\}$. For such x_i, the KKT equation is

$$-a_i + 2\mu x_i = 0,$$

which implies the lemma. ■

5. Tuning the Learning Rate: Proof of Theorem 4

Theorem 6 requires *a priori* knowledge of a good bound Q on the total quadratic variation Q_T. This may not be possible in many situations. Typically, in online learning scenarios where a regret bound of $O(\sqrt{A_T})$ for some quantity A_T which grows with T is desired, one first gives an online learning algorithm $L(\eta)$ where $\eta \leq 1$ is a learning rate parameter which obtains a regret bound of

$$\text{Regret}_T \ \leq \ \eta A_T + O(1/\eta).$$

Then, we can obtain a master online learning algorithm whose regret grows like $O(\sqrt{A_T})$ as follows. We start with $\eta = 1$, and run the learning algorithm $L(\eta)$. Then, the master algorithm tracks how A_T grows with T. As soon as A_T quadruples, the algorithm resets η to half its current value, and restarts with $L(\eta)$. This simple trick can be shown to obtain $O(\sqrt{A_T})$ regret.

Unfortunately, this trick doesn't work in our case, where $A_T = Q_T$, since we cannot even compute Q_T accurately in the bandit setting. For this reason, obtaining a regret bound of $\tilde{O}(\sqrt{Q_T})$ becomes quite non-trivial. In this section, we give a method to obtain such a regret bound. At its heart, we still make use of the η-halving trick, but in a subtle way. We assume that we know a good bound on $\log(T)$ in advance. This is not a serious restriction, it can be circumvented by standard tricks, but we make this assumption in order to simplify the exposition.

We design our master algorithm in the following way. Let $L(\eta)$ be Algorithm 1 with the given parameter η and $k = \log(T)$. We initialize $\eta_0 = \frac{1}{25n}$. The master algorithm then runs in phases

indexed by $i = 0, 1, 2, \ldots$. In phase i, the algorithm runs $L(\eta_i)$ where $\eta_i = \eta_0/2^i$. The decision to end a phase i and start phase $i+1$ is taken in the following manner: let t_i be first period of phase i, and let t be the current period. We start phase $i+1$ as soon as

$$\sum_{\tau=t_i}^{t} \tilde{\mathbf{f}}_\tau^\top (\mathbf{x}_\tau - \mathbf{x}_{\tau+1}) \;\geq\; \frac{2}{\eta_i} \vartheta \log(T).$$

Thus, phase i ends at time period $t-1$, and the point $\mathbf{x}_t$ computed by $L(\eta_i)$ is discarded by the master algorithm since $L(\eta_{i+1})$ starts at this point and $\mathbf{x}_t$ is reset to the initial point of $L(\eta_{i+1})$. Note that this sum can be computed by the algorithm, and hence the algorithm is well-defined. This completes the description of the master algorithm.

5.1 Analysis

Define $I_i = \{t_i, t_i+1, \ldots, t_{i+1}-1\}$, that is, the interval of time periods which constitute phase i.

By Lemma 8, for any $\mathbf{u} \in \mathcal{K}$, we have

$$\sum_{t \in I_i} \tilde{\mathbf{f}}_t^\top (\mathbf{x}_t - \mathbf{u}) \;\leq\; \sum_{t \in I_i} \tilde{\mathbf{f}}_t^\top (\mathbf{x}_t - \mathbf{x}_{t+1}) + \frac{2}{\eta_i} \vartheta \log(T) \;\leq\; \frac{4}{\eta_i} \vartheta \log(T).$$

Note that this inequality uses the fact that the sum $\sum_{\tau=t_i}^{t} \tilde{\mathbf{f}}_\tau^\top (\mathbf{x}_\tau - \mathbf{x}_{\tau+1})$ is a monotonically increasing as t increases, since by Lemma 14, we have that $\tilde{\mathbf{f}}_t^\top (\mathbf{x}_t - \mathbf{x}_{t+1}) \geq 0$.

Let $i^\star$ be the index of the final phase. Summing up this bound over all phases, we have

$$\sum_{t=1}^{T} \tilde{\mathbf{f}}_t^\top (\mathbf{x}_t - \mathbf{u}) \;\leq\; \sum_{i=0}^{i^\star} \frac{4}{\eta_i} \vartheta \log(T) \;\leq\; \frac{8}{\eta_{i^\star}} \vartheta \log(T).$$

Then, using Lemma 7 we get that the expected regret of this algorithm is bounded by

$$\mathbf{E}\left[\sum_{t=1}^{T} \mathbf{f}_t^\top (\mathbf{y}_t - \mathbf{u})\right] \;\leq\; \mathbf{E}\left[\frac{1}{\eta_{i^\star}}\right] \cdot (8\vartheta \log(T)) + O(n \log^2(T)). \tag{8}$$

We now need to bound $\mathbf{E}\left[\frac{1}{\eta_{i^\star}}\right]$. If the choice of the randomness in the algorithm is such that $i^\star = 0$, then $\frac{1}{\eta_{i^\star}} \leq 25n$ is an upper bound.

Otherwise, $i^\star > 0$, and so the phase $i^\star - 1$ is well-defined. For brevity, let $J = I_{i^\star-1} \cup \{t_{i^\star}\}$, and let J_E be the EllipsoidSample steps in J. For this interval, we have (here, $\mathbf{x}_{t_{i^\star}}$ is the point computed by $L(\eta_{i^\star-1})$, which is discarded by the master algorithm when phase $i^\star$ starts):

$$\sum_{t \in J} \tilde{\mathbf{f}}_t^\top (\mathbf{x}_t - \mathbf{x}_{t+1}) \;\geq\; \frac{2}{\eta_{i^\star-1}} \vartheta \log(T) \;=\; \frac{1}{\eta_{i^\star}} \vartheta \log(T).$$

Applying the bound (6), and using the fact that $\eta_{i^\star-1} = 2\eta_{i^\star}$, we get

$$\sum_{t \in J} \tilde{\mathbf{f}}_t^\top (\mathbf{x}_t - \mathbf{x}_{t+1}) \;\leq\; 128 \eta_{i^\star} n^2 \sum_{t \in J} \|\mathbf{f}_t - \mu_t\|^2 + 128 \eta_{i^\star} n^2 \sum_{t \in J_E} \|\mu_t - \tilde{\mu}_t\|^2 + 2 \sum_{t \in J} \mu_t^\top (\mathbf{x}_t - \mathbf{x}_{t+1}).$$

Putting these together, and dividing by $\eta_{i^\star}$, we get

$$\frac{1}{\eta_{i^\star}^2} \vartheta \log(T) \;\leq\; 128 n^2 \sum_{t \in J} \|\mathbf{f}_t - \mu_t\|^2 + 128 n^2 \sum_{t \in J_E} \|\mu_t - \tilde{\mu}_t\|^2 + \frac{2}{\eta_{i^\star}} \sum_{t \in J} \mu_t^\top (\mathbf{x}_t - \mathbf{x}_{t+1}). \tag{9}$$

Lemmas 10 and 12 give us the following upper bounds:

$$\sum_{t\in J}\|\mathbf{f}_t-\mu_t\|^2 \;\le\; Q_T \quad \text{and} \quad \sum_{t\in J}\mu_t^\top(\mathbf{x}_t-\mathbf{x}_{t+1}) \;\le\; 2\log(Q_T+1)+4.$$

Denote the expectation of a random variable conditioned on all the randomness before phase $i^\star-1$ by $\mathbf{E}_{i^\star-1}$. By Lemma 11 we have the bound

$$\mathbf{E}_{i^\star-1}\left[\sum_{t\in J_E}\|\mu_t-\tilde{\mu}_t\|^2\right] \;\le\; \frac{\log(T)}{k}Q_T.$$

Taking the expectation conditioned on all the randomness before phase $i^\star-1$ on both sides of inequality (9) and applying the above bounds, and using $k=\log(T)$, we get

$$\frac{1}{\eta_{i^\star}^2}\vartheta\log(T) \;\le\; 256n^2Q_T+\frac{4\log(Q_T+1)+8}{\eta_{i^\star}}.$$

Hence, one of $256n^2Q_T$ or $\frac{2}{\eta_{i^\star}}\log(Q_T)$ must be at least $\frac{1}{2\eta_{i^\star}^2}\vartheta\log(T)$. In the first case, we get the bound $\frac{1}{\eta_{i^\star}}\le 25n\sqrt{\frac{Q_T}{\vartheta\log(T)}}$. In the second case, we get the bound $\frac{1}{\eta_{i^\star}}\le\frac{8\log(Q_T+1)+16}{\vartheta\log(T)}$.

In all cases (including the case when $i^\star=0$), we have $\frac{1}{\eta_{i^\star}}\le O\left(n\sqrt{\frac{Q_T}{\vartheta\log(T)}}+n\right)$, and hence we can bound

$$\mathbf{E}\left[\frac{1}{\eta_{i^\star}}\right]\cdot(\vartheta\log(T)) \;=O\left(n\sqrt{\vartheta Q_T\log T}+n\vartheta\log(T)\right).$$

Plugging this into (8), and for $k=\log(T)$, we get that the expected regret is bounded by

$$\mathbf{E}\left[\sum_{t=1}^T\mathbf{f}_t^\top(\mathbf{y}_t-\mathbf{u})\right] \;=\; O\left(n\sqrt{\vartheta Q_T\log(T)}+n\log^2(T)+n\vartheta\log(T)\right).$$

6. Conclusions and Open Problems

In this paper, we gave the first bandit online linear optimization algorithm whose regret is bounded by the square-root of the total quadratic variation of the cost vectors. These bounds naturally interpolate between the worst-case and stochastic models of the problem.[4]

This algorithm continues a line of work which aims to prove variation-based regret bounds for any online learning framework. So far, such bounds have been obtained for four major online learning scenarios: expert prediction, online linear optimization, portfolio selection (and exp-concave cost functions), and bandit online linear optimization in this paper.

The concept of variational regret bounds in the setting of the ubiquitous multi-armed bandit problem opens many interesting directions for further research and open questions:

1. Improve upon the bounds presented in this paper by removing the dependence on the number of iterations completely - that is, remove the $\mathrm{poly}(\log(T))$ terms in the regret bound.

4. In the stochastic multi-armed bandit setting, the regret is known to be bounded by a logarithm in the number of iterations rather than square root (Auer et al., 2002). However, note that the regret is defined differently in the stochastic case, which makes the logarithmic dependency even possible. In this paper we consider a stronger notion of worst-case regret.

2. For the special case of the classic non-stochastic MAB problem, obtain regret bounds which depend on the variation of the best action in hindsight (vs. the total variation).

3. Is it possible to improve regret for the classic non-stochastic multi-armed bandit problem without using the self-concordance methodology (perhaps by extending the algorithm in Hazan and Kale (2008) to the bandit setting)?

References

J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *The 21st Annual Conference on Learning Theory (COLT)*., 2008.

J. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *J. Mach. Learn. Res.*, 9999:2785–2836, December 2010. ISSN 1532-4435. URL `http://portal.acm.org/citation.cfm?id=1953011.1953023`.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, 2002. ISSN 0885-6125. doi: http://dx.doi.org/10.1023/A:1013689704352.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2003. ISSN 0097-5397.

B. Awerbuch and R. D. Kleinberg. Adaptive routing with end-to-end feedback: distributed learning and geometric approaches. In *STOC*, pages 45–53, 2004. ISBN 1-58113-852-0.

A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, volume 2 of *MPS/SIAM Series on Optimization*. SIAM, Philadelphia, 2001.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Mach. Learn.*, 66(2-3):321–352, 2007. ISSN 0885-6125.

V. Dani and T. P. Hayes. Robbing the bandit: less regret in online geometric optimization against an adaptive adversary. In *SODA*, pages 937–943, 2006. ISBN 0-89871-605-5.

V. Dani, T. Hayes, and S. Kakade. The price of bandit information for online optimization. In *NIPS*, 2008.

A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *SODA*, pages 385–394, 2005. ISBN 0-89871-585-7.

J. Hannan. Approximation to bayes risk in repeated play. *In M. Dresher, A. W. Tucker, and P. Wolfe, editors, Contributions to the Theory of Games, volume III*, pages 97–139, 1957.

E. Hazan and S. Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. In *The 21st Annual Conference on Learning Theory (COLT)*., 2008.

E. Hazan and S. Kale. Better algorithms for benign bandits. In *SODA '09: Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 38–47, Philadelphia, PA, USA, 2009a. Society for Industrial and Applied Mathematics.

E. Hazan and S. Kale. On stochastic and worst-case models for investing. In *Advances in Neural Information Processing Systems (NIPS) 22*, 2009b.

A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, March 1985.

H. B. McMahan and A. Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *COLT*, pages 109–123, 2004.

A.S. Nemirovskii. Interior point polynomial time methods in convex programming, 2004. Lecture Notes.

Y. E. Nesterov and A. S. Nemirovskii. *Interior Point Polynomial Algorithms in Convex Programming*. SIAM, Philadelphia, 1994.

J. A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, April 2001. ISBN 0534399428. URL `http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0534399428`.

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

J. S. Vitter. Random sampling with a reservoir. *ACM Trans. Math. Softw.*, 11(1):37–57, 1985.