Online Geometric Optimization in the Bandit Setting Against an Adaptive Adversary

H. Brendan McMahan and Avrim Blum

Carnegie Mellon University, Pittsburgh, PA, 15213,
{mcmahan,avrim}@cs.cmu.edu

Abstract. We give an algorithm for the bandit version of a very general online optimization problem considered by Kalai and Vempala [1], for the case of an adaptive adversary. In this problem we are given a bounded set $S \subseteq \mathbb{R}^n$ of feasible points. At each time step t, the online algorithm must select a point $\mathbf{x}^t \in S$ while simultaneously an adversary selects a cost vector $\mathbf{c}^t \in \mathbb{R}^n$. The algorithm then incurs cost $\mathbf{c}^t \cdot \mathbf{x}^t$. Kalai and Vempala show that even if S is exponentially large (or infinite), so long as we have an efficient algorithm for the *offline* problem (given $\mathbf{c} \in \mathbb{R}^n$, find $\mathbf{x} \in S$ to minimize $\mathbf{c} \cdot \mathbf{x}$) and so long as the cost vectors are bounded, one can efficiently solve the *online* problem of performing nearly as well as the best fixed $\mathbf{x} \in S$ in hindsight. The Kalai-Vempala algorithm assumes that the cost vectors $\mathbf{c}^t$ are given to the algorithm after each time step. In the "bandit" version of the problem, the algorithm only observes its cost, $\mathbf{c}^t \cdot \mathbf{x}^t$. Awerbuch and Kleinberg [2] give an algorithm for the bandit version for the case of an oblivious adversary, and an algorithm that works against an adaptive adversary for the special case of the shortest path problem. They leave open the problem of handling an adaptive adversary in the general case. In this paper, we solve this open problem, giving a simple online algorithm for the bandit problem in the general case in the presence of an adaptive adversary. Ignoring a (polynomial) dependence on n, we achieve a regret bound of $\mathcal{O}(T^{3/4}\sqrt{\ln(T)})$.

1 Introduction

Kalai and Vempala [1] give an elegant, efficient algorithm for a broad class of online optimization problems. In their setting, we have an arbitrary (bounded) set $S \subseteq \mathbb{R}^n$ of feasible points. At each time step t, an online algorithm $\mathcal{A}$ must select a point $\mathbf{x}^t \in S$ and simultaneously an adversary selects a cost vector $\mathbf{c}^t \in \mathbb{R}^n$ (throughout the paper we use superscripts to index iterations). The algorithm then observes $\mathbf{c}^t$ and incurs cost $\mathbf{c}^t \cdot \mathbf{x}^t$. Kalai and Vempala show that so long as we have an efficient algorithm for the *offline* problem (given $\mathbf{c} \in \mathbb{R}^n$ find $\mathbf{x} \in S$ to minimize $\mathbf{c} \cdot \mathbf{x}$) and so long as the cost vectors are bounded, we can efficiently solve the *online* problem of performing nearly as well as the best fixed $\mathbf{x} \in S$ in hindsight. This generalizes the classic "expert advice" problem, because we do not require the set S to be represented explicitly: we just need an efficient oracle for selecting the best $\mathbf{x} \in S$ in hindsight. Further, it decouples the number of experts from the underlying dimensionality n of the decision set, under the assumption the cost of a decision is a linear function of n features of the decision. The standard experts setting can be recovered by letting $S = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$, the columns of the $n \times n$ identity matrix.

A problem that fits naturally into this framework is an online shortest path problem where we repeatedly travel between two points a and b in some graph whose edge costs change each day (say, due to traffic). In this case, we can view the set of paths as a set S of points in a space of dimension equal to the number of edges in the graph, and $\mathbf{c}^t$ is simply the vector of edge costs on day t. Even though the number of paths in a graph can be exponential in the number of edges (i.e., the set S is of exponential size), since we can solve the shortest path problem for any *given* set of edge lengths, we can apply the Kalai-Vempala algorithm. (Note that a different algorithm for the special case of the online shortest path problem is given by Takimoto and Warmuth [3].)

A natural generalization of the above problem, considered by Awerbuch and Kleinberg [2], is to imagine that rather than being given the entire cost vector $\mathbf{c}^t$, the algorithm is simply told the cost incurred $\mathbf{c}^t \cdot \mathbf{x}^t$. For example, in the case of shortest paths, rather than being told the lengths of all edges at time t, this would correspond to just being told the total time taken to reach the destination. Thus, this is the "bandit version" of the Kalai-Vempala setting. Awerbuch and Kleinberg present two results: an algorithm for the general problem in the presence of an *oblivious* adversary, and an algorithm for the special case of the shortest path problem that works in the presence of an *adaptive* adversary. The difference between the two adversaries is that an oblivious adversary must commit to the entire sequence of cost vectors in advance, whereas an adaptive adversary may determine the next cost vector based on the online algorithm's play (and hence, the information the algorithm received) in the previous time steps. Thus, an adaptive adversary is in essence playing a repeated game. They leave open the question of achieving good regret guarantees for an adaptive adversary in the general setting.

In this paper we solve the open question of [2], giving an algorithm for the general bandit setting in the presence of an adaptive adversary. Moreover, our method is significantly simpler than the special-purpose algorithm of Awerbuch and Kleinberg for shortest paths. Our bounds are somewhat worse: we achieve regret bounds of $\mathcal{O}(T^{3/4}\sqrt{\ln T})$ compared to the $\mathcal{O}(T^{2/3})$ bounds of [2]. We believe improvement in this direction may be possible, and present some discussion of this issue at the end of the paper.

The basic idea of our approach is as follows. We begin by noticing that the only history information used by the Kalai-Vempala algorithm in determining its action at time t is the sum $\mathbf{c}^{1:t-1} = \sum_{\tau=1}^{t-1} \mathbf{c}^\tau$ of all cost vectors received so far (we use this abbreviated notation for sums over iteration indexes throughout the paper). Furthermore, the way this is used in the algorithm is by adding random noise $\boldsymbol{\mu}$ to this vector, and then calling the offline oracle to find the $\mathbf{x}^t \in S$ that minimizes $(\mathbf{c}^{1:t-1} + \boldsymbol{\mu}) \cdot \mathbf{x}^t$. So, if we can design a bandit algorithm that produces an estimate $\hat{\mathbf{c}}^{1:t-1}$ of $\mathbf{c}^{1:t-1}$, and show that with high probability even an adaptive adversary will not cause $\hat{\mathbf{c}}^{1:t-1}$ to differ too substantially from $\mathbf{c}^{1:t-1}$, we can then argue that the distribution $\hat{\mathbf{c}}^{1:t-1} + \boldsymbol{\mu}$ is close enough to $\mathbf{c}^{1:t-1} + \boldsymbol{\mu}$ for the Kalai-Vempala analysis to apply. In fact, to make our analysis a bit more general, so that we could potentially use other algorithms as subroutines, we will argue a little differently. Let $\mathrm{OPT}(\mathbf{c}) = \min_{\mathbf{x}\in S}(\mathbf{c}\cdot\mathbf{x})$. We will show that with high probability, $\mathrm{OPT}(\hat{\mathbf{c}}^{1:T})$ is close to $\mathrm{OPT}(\mathbf{c}^{1:T})$ and $\hat{\mathbf{c}}^{1:T}$ satisfies conditions needed for the subroutine to achieve low regret on $\hat{\mathbf{c}}^{1:T}$. This means that our subroutine, which believes it has seen $\hat{\mathbf{c}}^{1:T}$, will achieve performance on $\hat{\mathbf{c}}^{1:T}$ close to $\mathrm{OPT}(\mathbf{c}^{1:T})$. We then finish off by arguing that our performance on $\mathbf{c}^{1:T}$ is close to its performance on $\hat{\mathbf{c}}^{1:T}$.

The behavior of the bandit algorithm will in fact be fairly simple. We begin by choosing a basis B of (at most) n points in S to use for sampling (we address the issue of how B is chosen when we describe our algorithm in detail). Then, at each time step t, with probability γ we *explore* by playing a random basis element, and otherwise (with probability $1-\gamma$) we *exploit* by playing according to the Kalai-Vempala algorithm. For each basis element $\mathbf{b}_j$, we use our cost incurred while exploring with that basis element, scaled by n/γ, as an estimate of $\mathbf{c}^{1:t-1}\cdot\mathbf{b}_j$. Using martingale tail inequalities, we argue that even an adaptive adversary cannot make our estimate differ too wildly from the true value of $\mathbf{c}^{1:t-1}\cdot\mathbf{b}_j$, and use this to show that after matrix inversion, our estimate $\hat{\mathbf{c}}^{1:t-1}$ is close to its correct value with high probability.

2 Problem Formalization

We can now fully formalize the problem. First, however, we establish a few notational conventions. As mentioned previously, we use superscripts to index iterations (or rounds) of our algorithm, and use the abbreviated summation notation $\mathbf{c}^{1:t}$ when summing variables over iterations. Vectors quantities are indicated in bold, and subscripts index into vectors or sets. Hats (such as $\hat{\mathbf{c}}^t$) denote estimates of the corresponding actual quantities. The variables and constants used in the paper are summarized in Table (1).

As mentioned above, we consider the setting of [1] in which we have an arbitrary (bounded) set $S \subseteq \mathbb{R}^n$ of feasible points. At each time step t, the online algorithm $\mathcal{A}$ must select a point $\mathbf{x}^t \in S$ and simultaneously an adversary selects a cost vector $\mathbf{c}^t \in \mathbb{R}^n$. The algorithm then incurs cost $\mathbf{c}^t\cdot\mathbf{x}^t$. Unlike [1], however, rather than being told $\mathbf{c}^t$, the algorithm simply learns its cost $\mathbf{c}^t\cdot\mathbf{x}^t$.

For simplicity, we assume a fixed adaptive adversary $\mathcal{V}$ and time horizon T for the duration of this paper. Since our choice of algorithm parameters depends on T, we assume[1] T is known to the algorithm. We refer to the sequence of decisions made by the algorithm so far as a decision history, which can be written $h^t = [\mathbf{x}^1,\ldots,\mathbf{x}^t]$. Let H^* be the set of all possible decision histories of length 0 through $T-1$. Without loss of generality (e.g., see [5]), we assume our adaptive adversary is deterministic, as specified by a function $\mathcal{V}: H^* \to \mathbb{R}^n$, a mapping

[1] One can remove this requirement by guessing T, and doubling the guess each time we play longer than expected (see, for example, Theorem 6.4 from [4]).

from decision histories to cost vectors. Thus, $\mathcal{V}(h^{t-1}) = \mathbf{c}^t$ is the cost vector for timestep t.

We can view our online decision problem as a game, where on each iteration t the adversary $\mathcal{V}$ selects a new cost vector $\mathbf{c}^t$ based on h^{t-1}, and the online algorithm $\mathcal{A}$ selects a decision $\mathbf{x} \in S$ based on its past plays and observations, and possibly additional hidden state or randomness. Then, $\mathcal{A}$ pays $\mathbf{c}^t \cdot \mathbf{x}^t$ and observes this cost. For our analysis, we assume a L_1 bound on S, namely $\|\mathbf{x}\|_1 \leq D/2$ for all $\mathbf{x} \in S$, so $\|\mathbf{x} - \mathbf{y}\|_1 \leq D$ for all $\mathbf{x}, \mathbf{y} \in S$. We also assume that $|\mathbf{c} \cdot \mathbf{x}| \leq M$ for all $\mathbf{x} \in S$ and all $\mathbf{c}$ played by $\mathcal{V}$. We also assume S is full rank, if it is not we simply project to a lower-dimensional representation. Some of these assumptions can be lifted or modified, but this set of assumptions simplifies the analysis.

For a fixed decision history h^T and cost history $k^T = (\mathbf{c}^1, \ldots, \mathbf{c}^T)$, we define $\mathrm{loss}(h^T, k^T) = \sum_{t=1}^{T}(\mathbf{c}^t \cdot \mathbf{x}^t)$. For a randomized algorithm $\mathcal{A}$ and adversary $\mathcal{V}$, we define the random variable $\mathrm{loss}(\mathcal{A}, \mathcal{V})$ to be $\mathrm{loss}(h^T, k^T)$, where h^T is drawn from the distribution over histories defined by $\mathcal{A}$ and $\mathcal{V}$, and $k^T = (\mathcal{V}(h^0), \ldots, \mathcal{V}(h^{T-1}))$. When it is clear from context, we will omit the dependence on $\mathcal{V}$, writing only $\mathrm{loss}(\mathcal{A})$.

Our goal is to define an online algorithm with low regret. That is, we want a guarantee that the total loss incurred will, in expectation, not be much larger than the optimal strategy in hindsight against the cost sequence we actually faced. To formalize this, first define an oracle $\mathcal{R} : \mathbb{R}^n \to S$ that solves the offline optimization problem, $\mathcal{R}(\mathbf{c}) = \mathrm{argmin}_{\mathbf{x} \in S}(\mathbf{c} \cdot \mathbf{x})$. We then define $\mathrm{OPT}(k^T) = \mathbf{c}^{1:T} \cdot \mathcal{R}(\mathbf{c}^{1:T})$. Similarly, $\mathrm{OPT}(\mathcal{A}, \mathcal{V})$ is the random variable $\mathrm{OPT}(k^T)$ when k^T is generated by playing $\mathcal{V}$ against $\mathcal{A}$. We again drop the dependence on $\mathcal{V}$ and $\mathcal{A}$ when it is clear from context. Formally, we define expected regret as

$$E\left[\mathrm{loss}(\mathcal{A}, \mathcal{V}) - \mathrm{OPT}(\mathcal{A}, \mathcal{V})\right] = E[\mathrm{loss}(\mathcal{A}, \mathcal{V})] - E\left[\min_{\mathbf{x} \in S} \sum_{t=1}^{T} (\mathbf{c}^t \cdot \mathbf{x})\right]. \qquad (1)$$

Note that the $E[\mathrm{OPT}(\mathcal{A}, \mathcal{V})]$ term corresponds to applying the min operator separately to each possible cost history to find the best fixed decision with respect to that particular cost history, and then taking the expectation with respect to these histories. In [5], an alternative weaker definition of regret is given. We discuss relationships between the definitions in Appendix B.

3 Algorithm

We introduce an algorithm we call BGA, standing for *Bandit-style Geometric decision algorithm against an Adaptive adversary.* The algorithm alternates between playing decisions from a fixed basis to get unbiased estimates of costs, and playing (hopefully) good decisions based on those estimates. In order to determine the good decisions to play, it uses some online geometric optimization algorithm for the full observation problem. We denote this algorithm by GEX (*Geometric Experts algorithm*). The implementation of GEX we analyze is based on the FPL algorithm of Kalai and Vempala [1]; we detail this implementation

```
Choose parameters γ and ε, where ε is a parameter of GEX
t = 1
Fix a basis B = {b_1, ..., b_n} ⊆ S
while playing do
  Let χ^t = 1 with probability γ and χ^t = 0 otherwise
  if χ^t = 0 then
    Select x^t from the distribution GEX(ĉ^1, ..., ĉ^{t-1})
    Incur cost z^t = c^t · x^t
    ĉ^t = 0 ∈ R^n
  else
    Draw j uniformly at random from {1, ..., n}
    x^t = b_j
    Incur cost and observe z^t = c^t · x^t
    Define ℓ̂^t by ℓ̂_i^t = 0 for i ≠ j and ℓ̂_j^t = (n/γ) z^t
    ĉ^t = (B^†)^{-1} ℓ̂^t
  end if
  ĉ^{1:t} = ĉ^{1:t-1} + ĉ^t
  t = t + 1
end while
```

Algorithm 1: BGA

and analysis in Appendix A. However, other algorithms could be used, for example the algorithm of Zinkevich [6] when S is convex. We view GEX as a function from the sequence of previous cost vectors $(\hat{\mathbf{c}}^1, \ldots, \hat{\mathbf{c}}^{t-1})$ to distributions over decisions.

Pseudocode for our algorithm is given in Algorithm (1). On each timestep, we make decision $\mathbf{x}^t$. With probability $(1-\gamma)$, BGA plays a recommendation $\mathbf{x}^t = \tilde{\mathbf{x}}^t \in S$ from GEX. With probability γ, we ignore $\tilde{\mathbf{x}}^t$ and play a basis decision, $\mathbf{x}^t = \mathbf{b}_i$ uniformly at random from a sampling basis $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$. The indicator variable χ^t is 1 on exploration iterations and 0 otherwise.

Our sampling basis B is a $n \times n$ matrix with columns $\mathbf{b}_i \in S$, so we can write $\mathbf{x} = B\mathbf{w}$ for any $\mathbf{x} \in \mathbb{R}^n$ and weights $\mathbf{w} \in \mathbb{R}^n$. For a given cost vector $\mathbf{c}$, let $\boldsymbol{\ell} = B^{\dagger}\mathbf{c}$ (the superscript $\dagger$ indicates transpose). This is the vector of decision costs for the basis decisions, so $\ell_i^t = \mathbf{c}^t \cdot \mathbf{b}_i$. We define $\hat{\boldsymbol{\ell}}^t$, an estimate of $\boldsymbol{\ell}^t$, as follows: Let $\hat{\boldsymbol{\ell}}^t = 0 \in \mathbb{R}^n$ on exploitation iterations. If on an exploration iteration we play $\mathbf{b}_j$, then $\hat{\boldsymbol{\ell}}^t$ is the vector where $\hat{\ell}_i^t = 0$ for $i \neq j$ and $\hat{\ell}_j^t = \frac{n}{\gamma}(\mathbf{c}^t \cdot \mathbf{b}_j)$. Note that $\mathbf{c}^t \cdot \mathbf{b}_j$ is the observed quantity, the cost of basis decision $\mathbf{b}_j$. On each iteration, we estimate $\mathbf{c}^t$ by $\hat{\mathbf{c}}^t = (B^{\dagger})^{-1}\hat{\boldsymbol{\ell}}^t$. It is straightforward to show that $\hat{\boldsymbol{\ell}}^t$ is an unbiased estimate of basis decision costs and that $\hat{\mathbf{c}}^t$ is an unbiased estimate of $\mathbf{c}^t$ on each timestep t.

The choice of the sampling basis plays an important role in the analysis of our algorithm. In particular, we use a baricentric spanner, introduced in [2]. A baricentric spanner $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ is a basis for S such that $\mathbf{b}_i \in S$ and for all $\mathbf{x} \in S$ we can write $\mathbf{x} = B\mathbf{w}$ with coefficients $w_i \in [-1, 1]$. It may not be easy to find exact baricentric spanners in all cases, but [2] proves they always exist and gives an algorithm for finding 2-approximate baricentric spanners (where the weights $w_i \in [-2, 2]$), which is sufficient for our purposes.

Table 1. Summary of notation

$S \subseteq \mathbb{R}^n$	set of decisions, a compact subset of $\mathbb{R}^n$
$D \in \mathbb{R}$	L_1 bound on diameter of S, $\forall \mathbf{x}, \mathbf{y} \in S,\ \lvert\mathbf{x} - \mathbf{y}\rvert_1 \leq D$
$n \in N$	dimension of decision space
h^t	decision history , $h^t = \mathbf{x}^1, \ldots, \mathbf{x}^t$
H^*	set of possible decision histories
$\mathcal{V} : H^* \to \mathbb{R}^n$	adversary, function from decision histories to cost vectors
$\mathcal{A}$	an online optimization algorithm
G^{t-1}	history of BGA randomness for timesteps 1 through $t-1$
$\mathbf{c}^t \in \mathbb{R}^n$	cost vector on time t
$\hat{\mathbf{c}}^t \in \mathbb{R}^n$	BGA's estimate of the cost vector on time t
$M \in \mathbb{R}^+$	bound on single-iteration cost, $\lvert\mathbf{c}^t \cdot \mathbf{x}^t\rvert \leq M$
$B \subseteq S$	sampling basis $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$
$\beta_\infty \in \mathbb{R}$	matrix max norm on $(B^\dagger)^{-1}$
$\boldsymbol{\ell}^t \in [-M, M]^n$	vector, $\ell_i^t = \mathbf{c}^t \cdot \mathbf{b}_i$ for $\mathbf{b}_i \in B$
$\hat{\boldsymbol{\ell}}^t \in \mathbb{R}^n$	BGA's estimate of $\boldsymbol{\ell}^t$
$T \in \mathbb{N}$	end of time, index of final iteration
$\mathbf{x}^t \in S$	BGA's decision on time t
$\tilde{\mathbf{x}}^t \in S$	decision recommended by GEX on time t
$\chi^t \in \{0, 1\}$	indicator, $\chi^t = 1$ if BGA explores on t, 0 otherwise
$\gamma \in [0, 1]$	the probability BGA explores on each timestep
$z^t \in [-M, M]$	BGA's loss on iteration t, $z^t = \mathbf{c}^t \cdot \mathbf{x}^t$,
$\hat{z}^t \in [-R, R]$	loss of GEX, $\hat{z}^t = \hat{\mathbf{c}}^t \cdot \tilde{\mathbf{x}}^t$

4 Analysis

4.1 Preliminaries

At each time step, BGA either (with probability $1-\gamma$) plays the recommendation $\tilde{\mathbf{x}}^t$ from GEX, or else (with probability γ) plays a random basis vector from B. For purposes of analysis, however, it will be convenient to imagine that we request a recommendation $\tilde{\mathbf{x}}^t$ from GEX on every iteration, and also that we randomly pick a basis to explore, $\mathbf{b}^t \in \{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$, on each iteration. We then decide to play either $\tilde{\mathbf{x}}^t$ or $\mathbf{b}^t$ based on the outcome χ^t of a coin of bias γ. Thus, the complete history of the algorithm is specified by the *algorithm history* $G^{t-1} = [\chi^1, \tilde{\mathbf{x}}^1, \mathbf{b}^1, \chi^2, \tilde{\mathbf{x}}^2, \mathbf{b}^2, \ldots, \chi^{t-1}, \tilde{\mathbf{x}}^{t-1}, \mathbf{b}^{t-1}]$, which encodes all previous random choices. The sample space for all probabilities and expectations is the set of all possible algorithm histories of length T. Thus, for a given adversary $\mathcal{V}$, the various random variables and vectors we consider, such as $\mathbf{x}^t, \mathbf{c}^t, \hat{\mathbf{c}}^t, \tilde{\mathbf{x}}^t$, and others, can all be viewed as functions on the set of possible algorithm histories. Unless otherwise stated, our expectations and probabilities are with respect to the distribution over these histories.

A partial history G^{t-1} can be viewed a subset of the sample space (an event) consisting of all complete histories that have G^{t-1} as a prefix. We frequently consider conditional distributions and corresponding expectations with respect to partial algorithm histories. For instance, if we condition on a history G^{t-1}, the random variables $\mathbf{c}^1, \ldots, \mathbf{c}^t, \boldsymbol{\ell}^1, \ldots, \boldsymbol{\ell}^t, \hat{\boldsymbol{\ell}}^1, \ldots, \hat{\boldsymbol{\ell}}^{t-1}, \hat{\mathbf{c}}^1, \ldots \hat{\mathbf{c}}^{t-1}, \mathbf{x}^1, \ldots, \mathbf{x}^{t-1}$, and $\chi^1, \ldots, \chi^{t-1}$ are fully determined.

We now outline the general structure of our argument. Let $\hat{z}^t = \hat{\mathbf{c}}^t \cdot \tilde{\mathbf{x}}^t$ be the loss perceived by the GEX on iteration t. In keeping with earlier definitions, $\text{loss(BGA)} = z^{1:T}$ and $\text{loss(GEX)} = \hat{z}^{1:T}$. We also let $\text{OPT} = \text{OPT}(\text{BGA}, \mathcal{V}) = \mathbf{c}^{1:T} \cdot \mathcal{R}(\mathbf{c}^{1:T})$, the performance of the best post-hoc decision, and similarly $\widehat{\text{OPT}} = \text{OPT}(\hat{\mathbf{c}}^1, \ldots, \hat{\mathbf{c}}^T) = \hat{\mathbf{c}}^{1:t} \cdot \mathcal{R}(\hat{\mathbf{c}}^{1:t})$.

The base of our analysis is a bound on the loss of GEX with respect to the cost vectors $\hat{\mathbf{c}}^t$ of the form

$$E[\text{loss(GEX)}] \leq E[\widehat{\text{OPT}}] + (\text{terms}). \tag{2}$$

Such a result is given in Appendix A, and follows from an adaptation of the analysis from [1]. We then prove statements having the general form

$$E[\text{loss(BGA)}] \leq E[\text{loss(GEX)}] + (\text{terms}) \tag{3}$$

and

$$E[\widehat{\text{OPT}}] \leq E[\text{OPT}] + (\text{terms}). \tag{4}$$

These statements connect our real loss to the "imaginary" loss of GEX, and similarly connect the loss of the best decision in GEX's imagined world with the loss of the best decision in the real world. Combining the results corresponding to Equations (2), (3), and (4) leads to an overall bound on the regret of BGA.

4.2 High Probability Bounds on Estimates

We prove a bound on the accuracy of BGA's estimates $\hat{\boldsymbol{\ell}}^t$, and use this to show a relationship between OPT and $\widehat{\text{OPT}}$ of the form in Equation 4.

Define random variables $\mathbf{e}^0 = \mathbf{0}$ and $\mathbf{e}^t = \boldsymbol{\ell}^t - \hat{\boldsymbol{\ell}}^t$. We are really interested in the corresponding sums $\mathbf{e}^{1:t}$, where $e_i^{1:t}$ is the total error in our estimate of $\mathbf{c}^{1:t} \cdot \mathbf{b}_i$. We now bound $|e_i^{1:t}|$.

Theorem 1. *For $\lambda > 0$,*

$$\Pr\left[|e_i^{1:t}| \geq \lambda \frac{nM}{\gamma}\sqrt{t}\right] \leq 2e^{-\lambda^2/2}.$$

Proof. It is sufficient to show the sequence $\mathbf{e}^0, \mathbf{e}^1, \mathbf{e}^{1:2}, \mathbf{e}^{1:3}, \ldots, \mathbf{e}^{1:T}$ of random variables is a bounded martingale sequence with respect to the filter $G^0, G^1, \ldots, G^T$; that is, $E[e_i^{1:t} \mid G^{t-1}] = e_i^{1:t-1}$. The result then follows from Azuma's Inequality (see, for example,[7]).

First, observe that $e_i^{1:t} = \ell_i^t - \hat{\ell}_i^t + e_i^{1:t-1}$. Further, the cost vector $\mathbf{c}^t$ is determined if we know G^{t-1}, and so ℓ_i^t is also fixed. Thus, accounting for the $\frac{\gamma}{n}$ probability we explore a particular basis decision $\mathbf{b}_i$, we have

$$E\left[e_i^{1:t} \mid G^{t-1}\right] = \frac{\gamma}{n}\left[\ell_i^t - \frac{n}{\gamma}\ell_i^t + e_i^{1:t-1}\right] + \left(1 - \frac{\gamma}{n}\right)\left[\ell_i^t - 0 + e_i^{1:t-1}\right] = e_i^{1:t-1},$$

and so we conclude that the $e_i^{1:t}$ forms a martingale sequence. Notice that $|e_i^{1:t} - e_i^{1:t-1}| = |\ell_i^t - \hat{\ell}_i^t|$. If we don't sample, $\hat{\ell}_i^t = 0$ and so $|e_i^{1:t} - e_i^{1:t-1}| \le M$. If we do sample, we have $\hat{\ell}_i^t = \frac{n}{\gamma}\ell_i^t$, and so $|e_i^{1:t} - e_i^{1:t-1}| \le \frac{nM}{\gamma}$. This bound is worse, so it holds in both cases. The result now follows from Azuma's inequality. □

Let $\beta_\infty = ||(B^\dagger)^{-1}||_\infty$, a matrix L_∞-norm on $(B^\dagger)^{-1}$, so that for any $\mathbf{w}$, $||(B^\dagger)^{-1}\mathbf{w}||_\infty \le \beta_\infty\,||\mathbf{w}||_\infty$.

Corollary 1. *For $\delta \in (0, 1]$, and all t from 1 to T,*

$$\Pr\left[||\hat{\mathbf{c}}^{1:t} - \mathbf{c}^{1:t}||_\infty \ge \beta_\infty J(\delta,\gamma)\sqrt{t}\right] \le \delta.$$

where $J(\delta,\gamma) = \frac{1}{\gamma}nM\sqrt{2\ln(2n/\delta)}$.

Proof. Solving $\delta/n = 2e^{-\lambda^2/2}$ yields $\lambda = \sqrt{2\ln(2n/\delta)}$, and then using this value in Theorem (1) gives

$$\Pr\left[|e_i^{1:t}| \ge J(\delta,\gamma)\sqrt{t}\right] \le \delta/n.$$

for all $i \in \{1, 2, \ldots, n\}$. Then,

$$\begin{aligned}\Pr\left[\ ||\mathbf{e}^{1:t}||_\infty \ge J(\delta,\gamma)\sqrt{t}\right] &\le \sum_{i=1}^{n}\Pr\left[|e_i^{1:t}| \ge J(\delta,\gamma)\sqrt{t}\right]\\ &\le \delta\end{aligned}$$

by the union bound. Now, notice that we can relate $\hat{\boldsymbol{\ell}}^{1:t}$ and $\hat{\mathbf{c}}^{1:t}$ by

$$(B^\dagger)^{-1}\hat{\boldsymbol{\ell}}^{1:t} = (B^\dagger)^{-1}\sum_{\tau=1}^{t}\boldsymbol{\ell}^\tau = \sum_{\tau=1}^{t}(B^\dagger)^{-1}\boldsymbol{\ell}^\tau = \sum_{\tau=1}^{t}\hat{\mathbf{c}}^\tau = \hat{\mathbf{c}}^{1:t}.$$

and similarly for $\boldsymbol{\ell}^{1:t}$ and $\mathbf{c}^{1:t}$. Then

$$\begin{aligned}\Pr\left[||\hat{\mathbf{c}}^{1:t} - \mathbf{c}^{1:t}||_\infty \ge \beta_\infty J(\delta,\gamma)\sqrt{t}\right] &= \Pr\left[||(B^\dagger)^{-1}(\hat{\boldsymbol{\ell}}^{1:t} - \boldsymbol{\ell}^{1:t})||_\infty \ge \beta_\infty J(\delta,\gamma)\sqrt{t}\right]\\ &\le \Pr\left[\beta_\infty||\mathbf{e}^{1:t}||_\infty \ge \beta_\infty J(\delta,\gamma)\sqrt{t}\right]\\ &= \Pr\left[||\mathbf{e}^{1:t}||_\infty \ge J(\delta,\gamma)\sqrt{t}\right]\\ &\le \delta.\end{aligned}$$

□

We can now prove our main result for the section, a statement of the form of Equation (4) relating OPT and $\widehat{\mathrm{OPT}}$:

Theorem 2. *If we play $\mathcal{V}$ against* BGA *for T timesteps,*

$$E[\widehat{\mathrm{OPT}}] \leq E[\mathrm{OPT}] + (1-\delta)\left(\frac{3}{2} D \beta_\infty J(\delta,\gamma)\sqrt{T}\right) + \delta M T.$$

Proof. Let $\Phi = \hat{\mathbf{c}}^{1:T} - \mathbf{c}^{1:T}$. By definition of $\mathcal{R}$, $\mathcal{R}(\hat{\mathbf{c}}^{1:T}) \cdot \hat{\mathbf{c}}^{1:T} \leq \mathcal{R}(\mathbf{c}^{1:T}) \cdot \hat{\mathbf{c}}^{1:T}$ or equivalently $\mathcal{R}(\mathbf{c}^{1:T} + \Phi) \cdot (\mathbf{c}^{1:T} + \Phi) \leq \mathcal{R}(\mathbf{c}^{1:T}) \cdot (\mathbf{c}^{1:T} + \Phi)$, and so by expanding and rearranging we have

$$\begin{aligned}\mathcal{R}(\mathbf{c}^{1:T} + \Phi) \cdot \mathbf{c}^{1:T} - \mathcal{R}(\mathbf{c}^{1:T}) \cdot \mathbf{c}^{1:T} &\leq (\mathcal{R}(\mathbf{c}^{1:T}) - \mathcal{R}(\mathbf{c}^{1:T} + \Phi)) \cdot \Phi \\ &\leq D\|\Phi\|_\infty. \qquad (5)\end{aligned}$$

Then,

$$\begin{aligned}|\,\mathrm{OPT} - \widehat{\mathrm{OPT}}| &= |\mathcal{R}(\mathbf{c}^{1:T}) \cdot \mathbf{c}^{1:T} - \mathcal{R}(\mathbf{c}^{1:T} + \Phi) \cdot (\mathbf{c}^{1:T} + \Phi)| \\ &\leq |(\mathcal{R}(\mathbf{c}^{1:T}) - \mathcal{R}(\mathbf{c}^{1:T} + \Phi)) \cdot \mathbf{c}^{1:T}| + |\mathcal{R}(\mathbf{c}^{1:T} + \Phi) \cdot \Phi| \\ &\leq (D + D/2)\|\Phi\|_\infty,\end{aligned}$$

where we have used Equation (5). Recall from Section (2), we assume $\|\mathbf{x}\|_1 \leq D/2$ for all $\mathbf{x} \in S$, so $\|\mathbf{x} - \mathbf{y}\|_1 \leq D$ for all $\mathbf{x}, \mathbf{y} \in S$. The theorem follows by applying the bound on Φ given by Corollary (1), and then observing that the above relationship holds for at least a $1-\delta$ fraction of the possible algorithm histories. For the other δ fraction, the difference might be as much as δMT. Writing the overall expectation as the sum of two expectations conditioned on whether or not the bound holds gives the result. □

4.3 Relating the Loss of BGA and its GEX Subroutine

Now we prove a statement like Equation (3), relating loss(BGA) to loss(GEX).

Theorem 3. *If we run* BGA *with parameter γ against $\mathcal{V}$ for T timesteps,*

$$E[\mathit{loss}(\mathrm{BGA})] \leq (1-\gamma) E[\mathit{loss}(\mathrm{GEX})] + \gamma M T.$$

Proof. For a given adversary $\mathcal{V}$, G^{t-1} fully determines the sequence of cost vectors given to algorithm GEX. So, we can view GEX as a function from G^{t-1} to probability distributions over S. If we present a cost vector $\hat{\mathbf{c}}$ to GEX, then the expected cost to GEX given history G^{t-1} is $\sum_{\tilde{\mathbf{x}} \in S} \Pr(\tilde{\mathbf{x}} \mid G^{t-1})\ (\hat{\mathbf{c}} \cdot \tilde{\mathbf{x}})$. If we define $\bar{\mathbf{x}}^t = \sum_{\tilde{\mathbf{x}} \in S} \Pr(\tilde{\mathbf{x}} \mid G^{t-1})\, \tilde{\mathbf{x}}$, we can re-write the expected loss of GEX against $\hat{\mathbf{c}}$ as $\hat{\mathbf{c}} \cdot \bar{\mathbf{x}}^t$; that is, we can view GEX as incurring the cost of some convex combination of the possible decisions in expectation. Let $\hat{\boldsymbol{\ell}}^{t,j}$ be $\hat{\boldsymbol{\ell}}^t$ given that we explore by playing basis vector $\mathbf{b}_j$ on time t, and similarly let $\hat{\mathbf{c}}^{t,j} = (B^{\dagger})^{-1}\hat{\boldsymbol{\ell}}^{t,j}$. Observe that $\hat{\ell}_i^{t,j} = \frac{n}{\gamma}\ell_i^t$ for $j = i$ and 0 otherwise, and so

$$\sum_{j=1}^{n} \hat{\boldsymbol{\ell}}^{t,j} = \frac{n}{\gamma}\boldsymbol{\ell}^t = \frac{n}{\gamma}B^{\dagger}\mathbf{c}^t. \tag{6}$$

Now, we can write

$$\begin{aligned}
E[\hat{z}^t \mid G^{t-1}] &= (1-\gamma)\,0 + \gamma \sum_{j=1}^{n} \frac{1}{n} \sum_{\tilde{\mathbf{x}}^t \in S} \Pr(\tilde{\mathbf{x}}^t \mid G^{t-1})\ (\hat{\mathbf{c}}^{t,j} \cdot \tilde{\mathbf{x}}^t) \\
&= \gamma \left[\sum_{j=1}^{n} \frac{1}{n} \hat{\mathbf{c}}^{t,j} \right] \cdot \bar{\mathbf{x}}^t \\
&= \frac{\gamma}{n}(B^{\dagger})^{-1} \left[\sum_{j=1}^{n} \hat{\boldsymbol{\ell}}^{t,j} \right] \cdot \bar{\mathbf{x}}^t, \quad \text{and using Equation (6),} \\
&= \mathbf{c}^t \cdot \bar{\mathbf{x}}^t.
\end{aligned}$$

Now, we consider the conditional expectation of z^t and see that

$$\begin{aligned}
E[z^t \mid G^{t-1}] &= (1-\gamma)(\mathbf{c}^t \cdot \bar{\mathbf{x}}^t) + \gamma \sum_{i=1}^{n} \frac{1}{n}(\mathbf{c}^t \cdot \mathbf{b}_i) \\
&\leq (1-\gamma)E[\hat{z}^t \mid G^{t-1}] + \gamma M, \qquad (7)
\end{aligned}$$

Then we have,

$$\begin{aligned}
E[z^t] &= E\left[E[z^t \mid G^{t-1}]\right] \\
&\leq E\left[(1-\gamma)E[\hat{z}^t \mid G^{t-1}] + \gamma M\right] \\
&= (1-\gamma)E\left[E[\hat{z}^t \mid G^{t-1}]\right] + \gamma M \\
&= (1-\gamma)E[\hat{z}^t] + \gamma M, \qquad (8)
\end{aligned}$$

by using the inequality from Equation (7). The theorem follows by summing the inequality (8) over t from 1 to T and applying linearity of expectation. □

4.4 A Bound on the Expected Regret of BGA

Theorem 4. *If we run* BGA *with parameter γ using subroutine* GEX *with parameter ϵ (as defined in Appendix A), then for all $\delta \in (0,1]$,*

$$\begin{aligned}
&E[loss(\text{BGA})] \\
&\quad \leq E[\text{OPT}] + \mathcal{O}\left(D\frac{1}{\gamma}nM\sqrt{2\ln(2n/\delta)}\sqrt{T} + \delta MT + \frac{\epsilon}{\gamma^2}n^3M^2T + \frac{n}{\epsilon} + \gamma MT \right)
\end{aligned}$$

Proof. In Appendix A, we show an algorithm to plug in for GEX, based on the FPL algorithm of [1] and give bounds on regret against a deterministic adaptive adversary. We first show how to apply that analysis to GEX running as a subroutine to BGA.

First, we need to bound $|\hat{\mathbf{c}}^t \cdot \mathbf{x}|$. By definition, for any $\mathbf{x} \in S$, we can write $\mathbf{x} = B\mathbf{w}$ for weights $\mathbf{w}$ with $\mathbf{w}_i \in [-1,1]$ (or $[-2,2]$ if it is an approximate baricentric spanner). Note that $||\hat{\boldsymbol{\ell}}^t||_1 \leq (\frac{n}{\gamma})M$, and for any $\mathbf{x} \in S$, we can write $\mathbf{x}$ as $B\mathbf{w}$ where $\mathbf{w}_i \in [-2,2]$. Thus,

$$|\hat{\mathbf{c}}^t \cdot \mathbf{x}| = |(B^\dagger)^{-1}\hat{\boldsymbol{\ell}}^t \cdot B\mathbf{w}| = |(\hat{\boldsymbol{\ell}}^t)^\dagger B^{-1} B\mathbf{w}| = |\hat{\boldsymbol{\ell}}^t \cdot \mathbf{w}| \leq ||\hat{\boldsymbol{\ell}}^t||_1 \; ||\mathbf{w}||_\infty \leq \frac{2nM}{\gamma}.$$

Let $R = 2nM/\gamma$. Suppose at the beginning of time we fix the random decisions of BGA that are not made by GEX, that is, we fix a sequence $X = [\chi^1, \mathbf{b}^1, \ldots, \chi^T, \mathbf{b}^T]$. Fixing this randomness together with $\mathcal{V}$ determines a new deterministic adaptive adversary $\hat{\mathcal{V}}$ that GEX is effectively playing against. To see this, let $\tilde{h}^{t-1} = [\tilde{\mathbf{x}}^1, \ldots, \tilde{\mathbf{x}}^{t-1}]$. If we combine $\tilde{h}^{t-1}$ with the information in X, it fully determines a partial history G^{t-1}. If we let $h^{t-1} = [\mathbf{x}^1, \ldots, \mathbf{x}^{t-1}]$ be the partial decision history that can be recovered from G^{t-1}, then $\hat{\mathcal{V}}(\tilde{h}^{t-1}) = \chi^t \frac{d}{\gamma}\mathcal{V}(h^{t-1})$. Thus, when GEX is run as a subroutine of BGA, we can apply Lemma (3) from the Appendix and conclude

$$E[\text{loss(GEX)} \mid X] \leq E[\widehat{\text{OPT}} \mid X] + \epsilon(4n+2)R^2T + \frac{4n}{\epsilon} \tag{9}$$

For the remainder of this proof, we use big-Oh notation to simplify the presentation. Now, taking the expectation of both sides of Equation (9),

$$E[\text{loss(GEX)}] \leq E[\widehat{\text{OPT}}] + \mathcal{O}\left(\epsilon n R^2 T + \frac{n}{\epsilon}\right)$$

Applying Theorem (3),

$$E[\text{loss(BGA)}] \leq (1-\gamma)E[\widehat{\text{OPT}}] + \mathcal{O}\left(\epsilon n R^2 T + \frac{n}{\epsilon} + \gamma MT\right)$$

and then using Theorem (2) we have

$$\begin{aligned}
&E[\text{loss(BGA)}] \\
&\quad \leq (1-\gamma)E[\text{OPT}] + \mathcal{O}\left(J(\delta,\gamma)D\sqrt{T} + \delta MT + \epsilon n R^2 T + \frac{n}{\epsilon} + \gamma MT\right) \\
&\quad \leq E[\text{OPT}] + \mathcal{O}\left(D\frac{1}{\gamma}nM\sqrt{2\ln(2n/\delta)}\sqrt{T} + \delta MT + \frac{\epsilon}{\gamma^2}n^3M^2T + \frac{n}{\epsilon} + \gamma MT\right)
\end{aligned}$$

For the last line, note that while $E[OPT]$ could be negative, it is still bounded by MT, and so this just adds another γMT term, which is captured in the big-Oh term. □

Ignoring the dependence on n, M, and D and simplifying, we see BGA's expected regret is bounded by

$$E[\text{regret}(\text{BGA})] = \mathcal{O}\left(\frac{\sqrt{T}\sqrt{\ln(1/\delta)}}{\gamma} + \delta T + \frac{\epsilon T}{\gamma^2} + \frac{1}{\epsilon} + \gamma T\right).$$

Setting $\gamma = \delta = T^{-1/4}$ and $\epsilon = T^{-3/4}$, we get a bound on our loss of order $\mathcal{O}(T^{3/4}\sqrt{\ln T})$.

5 Conclusions and Open Problems

We have presented a general algorithm for online optimization over an arbitrary set of decisions $S \subseteq \mathbb{R}^n$, and proved regret bounds for our algorithm that hold against an adaptive adversary.

A number of questions are raised by this work. In the "flat" bandits problem, bounds of the form $\mathcal{O}(\sqrt{T})$ are possible against an adaptive adversary [4]. Against a oblivious adversary in the geometric case, a bound of $\mathcal{O}(T^{2/3})$ is achieved in [2]. We achieve a bound of $\mathcal{O}(T^{3/4}\sqrt{\ln T})$ for this problem against an adaptive adversary. In [4], lower bounds are given showing that the $\mathcal{O}(\sqrt{T})$ result is tight, but no such bounds are known for the geometric decision-space problem. Can the $\mathcal{O}(T^{3/4}\sqrt{\ln T})$ and possibly the $\mathcal{O}(T^{2/3})$ bounds be tightened to $\mathcal{O}(\sqrt{T})$? A related issue is the use of information received by the algorithm; our algorithm and the algorithm of [2] only use a γ fraction of the feedback they receive, which is intuitively unappealing. It seems plausible that an algorithm can be found that uses all of the feedback, possibly achieving tighter bounds.

Acknowledgments

The authors wish to thank Adam Kalai, Geoff Gordon, Bobby Kleinberg, Tom Hayes, and Varsha Dani for useful conversations and correspondence. Funding provided by NSF grants CCR-0105488, NSF-ITR CCR-0122581, and NSF-ITR IIS-0312814.

References

1. Kalai, A., Vempala, S.: Efficient algorithms for on-line optimization. In: Proceedings of the The 16th Annual Conference on Learning Theory. (2003)
2. Awerbuch, B., Kleinberg, R.: Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In: Proceedings of the 36th ACM Symposium on Theory of Computing. (2004) To appear.
3. Takimoto, E., Warmuth, M.K.: Path kernels and multiplicative updates. In: Proceedings of the 15th Annual Conference on Computational Learning Theory. Lecture Notes in Artificial Intelligence, Springer (2002)
4. Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E.: The nonstochastic multi-armed bandit problem. SIAM Journal on Computing **32** (2002) 48–77

5. Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E.: Gambling in a rigged casino: the adversarial multi-armed bandit problem. In: Proceedings of the 36th Annual Symposium on Foundations of Computer Science, IEEE Computer Society Press, Los Alamitos, CA (1995) 322–331
6. Zinkevich, M.: Online convex programming and generalized infinitesimal gradient ascent. In: Proceedings of the Twentieth International Conference on Machine Learning. (2003)
7. Motwani, R., Raghavan, P.: Randomized algorithms. Cambridge University Press (1995)
8. Zinkevich, M.: Online convex programming and generalized infinitesimal gradient ascent. Technical Report CMU-CS-03-110, Carnegie Mellon University (2003)

A Specification of a Geometric Experts Algorithm

In this section we point out how the FPL algorithm and analysis of [1] can be adapted to our setting to use as the GEX subroutine, and prove the corresponding bound needed for Theorem (4). In particular, we need a bound for an arbitrary $S \subseteq \mathbb{R}^n$ and arbitrary cost vectors, requiring only that on each timestep, $|\mathbf{c} \cdot \mathbf{x}| \leq R$. Further, the bound must hold against an adaptive adversary.

FPL solves the online optimization problem when the entire cost vector $\mathbf{c}^t$ is observed at each timestep. It maintains the sum $\mathbf{c}^{1:t-1}$, and on each timestep plays decision $\mathbf{x}^t = \mathcal{R}(\mathbf{c}^{1:t-1} + \boldsymbol{\mu})$, where $\boldsymbol{\mu}$ is chosen uniformly at random from $[0, 1/\epsilon]^n$, given ϵ, a parameter of the algorithm. The analysis of FPL in [1] assumes positive cost vectors $\mathbf{c}$ satisfying $||\mathbf{c}||_1 \leq A$, and positive decision vectors from $S \subseteq \mathbb{R}^n_+$ with $||\mathbf{x} - \mathbf{y}||_1 \leq D$ for all $\mathbf{x}, \mathbf{y} \in S$ and $|\mathbf{c} \cdot \mathbf{x} - \mathbf{c} \cdot \mathbf{y}| \leq R$ for all cost vectors $\mathbf{c}$ and $\mathbf{x}, \mathbf{y} \in S$. Further, the bounds proved are with respect to a fixed series of cost vectors, not an adaptive adversary. We now show how to bridge the gap from these assumptions to our assumptions.

First, we adapt an argument from [2], showing that by using our baricentric spanner basis, we can transform our problem into one where the assumptions of FPL are met. We then argue that a corresponding bound holds against an adaptive adversary.

Lemma 1. *Let $S \subseteq \mathbb{R}^n$ be a set of (not necessarily positive) decisions, and $k^t = [\mathbf{c}^1, \ldots, \mathbf{c}^T]$ a set of cost vectors on those decisions, such that $|\mathbf{c}^t \cdot \mathbf{x}| \leq R$ for all $\mathbf{x} \in S$ and $\mathbf{c}^t \in k^t$. Then, there is an algorithm $\mathcal{A}(\epsilon)$ that achieves*

$$E[\text{loss}(\mathcal{A}(\epsilon), k^t)] \leq \text{OPT}(k^t) + \epsilon(4n+2)R^2T + \frac{4n}{\epsilon}$$

Proof. This an adaptation of the arguments of Appendix A of [2]. Fix a baricentric spanner $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ for S. Then, for each $\mathbf{x} \in S$, let $\mathbf{x} = B\mathbf{w}$ and define $f(\mathbf{x}) = [-\sum_{i=1}^n \mathbf{w}_i, \mathbf{w}_1, \ldots, \mathbf{w}_n]$. Let $f(S) = S'$. For each cost vector $\mathbf{c}^t$ define $g(\mathbf{c}^t) = [R,\ R + \mathbf{c}^t \cdot bs_1,\ \ldots,\ R + \mathbf{c}^t \cdot \mathbf{b}_n]$. It is straightforward to verify that $\mathbf{c}^t \cdot \mathbf{x} = g(\mathbf{c}^t) \cdot f(\mathbf{x})$, and further $g(\mathbf{c}^t) \geq 0$, $||g(\mathbf{c}^t)||_1 \leq (2n+1)R$, and the difference in cost of any two decisions against a fixed $g(\mathbf{c}^t)$ is at most $2R$. By definition of a baricentric spanner, $\mathbf{w}_i \in [-1, 1]$ and so the L_1 diameter of S' is at most $4n$. Note the assumption of positive decision vectors in Theorem 1 of [1] can easily be lifted by additively shifting the space of decision vectors until it is positive. This changes the loss of the algorithm and of the best decision by the same amount, so additive regret bounds are unchanged. The result of this lemma then follows from the bound of Theorem 1 from [1]. □

We now need to extend the above bound to adaptive adversaries. The key point here is that the algorithm is *self-oblivious*. A self-oblivious algorithm always plays a decision from some distribution that depends only on the cost history so far and not the outcome of its previous probabilistic choices. Thus, a self-oblivious algorithm can be viewed as a function from cost histories to distributions over decisions. For such algorithms, for any (possibly adaptive) adversary

$\mathcal{V}$ there always exists an oblivious adversary that causes at least as much regret. The idea for the proof below is due to Adam Kalai.[2]

Lemma 2. *Fix T, let H^* be the set of decision histories of length 0 to $T-1$, and let K^* be the set of all cost histories of length 0 to $T-1$. Then, fix a decision algorithm $\mathcal{A} : K^* \to \Delta(S)$, where $\Delta(S)$ is the set of probability distributions on the set S of possible decisions. Define*

$$R(\mathcal{A}, \mathcal{V}) = E_{\mathcal{A},\mathcal{V}} \left[\sum_{t=1}^{T} \mathbf{c}^t \mathbf{x}^t - \min_{\mathbf{x} \in S} \sum_{t=1}^{T} \mathbf{c}^t \mathbf{x} \right]$$

Let $\mathcal{V}$ be an arbitrary adversary. Then, there exists an oblivious adversary $\mathcal{V}'$ such that

$$R(\mathcal{A}, \mathcal{V}') \geq R(\mathcal{A}, \mathcal{V})$$

Proof. An adversary is *t-oblivious* if its first t costs are chosen obliviously; note all adversaries are 1-oblivious. Let $\mathcal{V}$ be an arbitrary adversary, and suppose it is k-oblivious. If $k = T$, we are done. Otherwise, let $\mathbf{c}_o^1, \ldots, \mathbf{c}_o^k$ be the first k (obliviously chosen) costs selected by $\mathcal{V}$. Expectations are over the random variables $\mathbf{x}^1, \ldots, \mathbf{x}^T$ and $\mathbf{c}^1, \ldots, \mathbf{c}^T$ when $\mathcal{V}$ plays against $\mathcal{A}$, though in this case $\mathbf{c}^1, \ldots, \mathbf{c}^k$ are fully determined. Let $K^T = \mathbf{c}^1, \ldots, \mathbf{c}^T$, the random vector corresponding to the cost history.

Let $g(K^T) = \min_{\mathbf{x} \in S} \sum_{t=1}^{T} \mathbf{c}^t \mathbf{x}$. Using linearity of expectation, we can split the expected regret $R(\mathcal{A}, \mathcal{V})$ into 3 terms:

$$E[\sum_{t=1}^{k} \mathbf{c}_o^t \mathbf{x}^t] + E[\mathbf{c}^{k+1} \mathbf{x}^{k+1}] + E[\sum_{t=k+2}^{T} \mathbf{c}^t \mathbf{x}^t - g(K^T)]$$

Since $\mathcal{A}$ and $\mathbf{c}^1, \ldots, \mathbf{c}^k$ are fixed, $E[\mathbf{x}^{k+1}] = E[\mathcal{A}(\mathbf{c}_o^1, \ldots, \mathbf{c}_o^k)] = \bar{x}$ is also known. Since $\mathcal{V}$ is only k-oblivious, it gets to pick $\mathbf{c}^{k+1}$ with knowledge of $\mathbf{x}^1, \ldots, \mathbf{x}^k$. We have

$$\Pr(\mathbf{c}^{k+1}) = \int_{\mathbf{x}^1, \ldots, \mathbf{x}^k} \Pr(\mathbf{x}^1, \ldots, \mathbf{x}^k) \mathcal{I}[\mathcal{V}(\mathbf{x}^1, \ldots, \mathbf{x}^k) = \mathbf{c}^{k+1}],$$

where $\mathcal{I}$ is an indicator function, returning 1 if $\mathcal{V}(\mathbf{x}^1, \ldots, \mathbf{x}^k) = \mathbf{c}^{k+1}$ and zero otherwise. The probability $\Pr(\mathbf{x}^1, \ldots, \mathbf{x}^k)$ is well defined because $\mathcal{V}$ and $\mathcal{A}$ are fixed. Importantly, note that the distribution over $\mathbf{c}^{k+1}$ is independent of the distribution over $\mathbf{x}^{k+1}$; this follows from the assumption that $\mathcal{A}$ is self-oblivious, that is, it picks its distributions based only on the past cost vectors, not on its

[2] We thank Tom Hayes and Varsha Dani for pointing out a bug in the proof we had in the original version of this paper.

own actions. Thus, letting $L^k = E[\sum_{t=1}^k \mathbf{c}_o^t \mathbf{x}^t]$ we can write

$$R(\mathcal{A}, \mathcal{V}) = L^k + \bar{x}E[\mathbf{c}^{k+1}] + E[\sum_{t=k+2}^{T} \mathbf{c}^t \mathbf{x}^t - g(K^T)] \tag{10}$$

$$= L^k + \int_{\mathbf{c}^{k+1}} \Pr(\mathbf{c}^{k+1}) \left[\mathbf{c}^{k+1}\bar{x} + E[\sum_{t=k+2}^{T} \mathbf{c}^t \mathbf{x}^t - g(K^T) \mid \mathbf{c}^{k+1}] \right] d\mathbf{c}^{k+1} \tag{11}$$

$$\leq L^k + \sup_{\mathbf{c}^{k+1}} \left[\mathbf{c}^{k+1}\bar{x} + E[\sum_{t=k+2}^{T} \mathbf{c}^t \mathbf{x}^t - g(K^T) \mid \mathbf{c}^{k+1}], \right] \tag{12}$$

where the sup is over all $\mathbf{c}^{k+1}$ with $\Pr(\mathbf{c}^{k+1}) > 0$. Observe that the quantity inside the supremum is well defined before any costs or decisions are selected, and so $\mathcal{V}$ could do at least as well by selecting $\mathbf{c}^{k+1}$ obliviously to be some c that achieves the supremum. Thus, there is a $(k+1)$-oblivious adversary that causes at least as much regret as $\mathcal{V}$. Extending this result inductively, we conclude there is a fully oblivious (T-oblivious) adversary $\mathcal{V}'$ such that $R(\mathcal{A}, \mathcal{V}') \geq R(\mathcal{A}, \mathcal{V})$. □

Lemma 3. *The regret bound from Lemma 1 applies even if the adversary is adaptive.*

Proof. First, observe that as long as FPL re-randomizes at each timestep, it is self-oblivious, and so Lemma 2 applies. Suppose some adaptive adversary $\mathcal{V}$ causes regret that exceeds the bound in Lemma 1. We can apply Lemma 2 to $\mathcal{V}$ and construct an oblivious $\mathcal{V}'$ that also exceeds the bound, a contradiction.

Thus, we can use $\mathcal{A}(\epsilon)$ as our GEX subroutine for full-observation online geometric optimization.

B Notions of Regret

In [5], an alternative definition of regret is given, namely,

$$E[\text{loss}_{\mathcal{V},\mathcal{A}}(h^T)] - \min_{x \in S} E\left[\sum_{t=1}^{T} \mathbf{c}^t \cdot \mathbf{x}\right]. \tag{13}$$

This definition is equivalent to ours in the case of an *oblivious* adversary, but against an adaptive adversary the "best decision" for this definition is not the best decision for a *particular* decision history, but the best decision if the decision must be chosen before a cost history is selected according to the distribution over such histories. In particular,

$$E\left[\min_{x \in S} \sum_{t=1}^{T} \mathbf{c}^t \cdot \mathbf{x}\right] \leq \min_{x \in S} E\left[\sum_{t=1}^{T} \mathbf{c}^t \cdot \mathbf{x}\right]$$

and so a bound on Equation (1) is at least as strong as a bound on Equation (13). In fact, bounds on Equation (13) can be very poor when the adversary is adaptive. There are natural examples where the stronger definition (1) gives regret $\mathcal{O}(T)$ while the weaker definition (13) indicates no regret. Adapting an example from [5], let $S = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ (the "flat" bandit setting) and consider the algorithm $\mathcal{A}$ that plays uniformly at random from S. The adversary $\mathcal{V}$ gives $\mathbf{c}^1 = \mathbf{0}$, and if $\mathcal{A}$ then plays $\mathbf{e}_i$ on the first iteration, thereafter the adversary plays the cost vector $\mathbf{c}^t$ where $c_i^t = 0$ and $c_j^t = 1$ for $j \neq i$. The expected loss of $\mathcal{A}$ is $\frac{n-1}{n}T$. For regret as defined by Equation (13), $\min_{\mathbf{x} \in S} E[\mathbf{c}^{1:T} \cdot \mathbf{x}] = \frac{n-1}{n}T$, indicating no regret, while $E[\min_{\mathbf{x} \in S}(\mathbf{c}^{1:T} \cdot \mathbf{x})] = 0$, and so the stronger definition indicates $\mathcal{O}(T)$ regret.

Unfortunately, this implies like the proof techniques for bounds on expected weak regret like those in [4] and [2] cannot be used to get bounds on regret as defined by Equation (1). The problem is that even if we have unbiased estimates of the costs, these cannot be used to evaluate the term $E[\min_{x \in S} \sum_{t=1}^{T} (\mathbf{c}^t \cdot \mathbf{x})]$ in (1) because min is a non-linear operator. We surmount this problem by proving high-probability bounds on our estimates of $\mathbf{c}^t$, which allows us to use a union bound to evaluate the expectation over the min operator. Note that the high probability bounds proved in [4] and [2] can be seen as corresponding to our definition of expected regret.