On Following the Perturbed Leader in the Bandit Setting

Jussi Kujala and Tapio Elomaa

Institute of Software Systems
Tampere University of Technology
P. O. Box 553, FI-33101 Tampere, Finland
jussi.kujala@tut.fi, elomaa@cs.tut.fi

Abstract. In an online decision problem an algorithm is at each time step required to choose one of the feasible points without knowing the cost associated with it. An adversary assigns the cost to possible decisions either obliviously or adaptively. The online algorithm, naturally, attempts to collect as little cost as possible. The cost difference of the online algorithm and the best static decision in hindsight is called the *regret* of the algorithm.

Kalai and Vempala [1] showed that it is possible to have efficient solutions to some problems with a linear cost function by following the perturbed leader. Their solution requires the costs of all decisions to be known. Recently there has also been some progress in the *bandit setting*, where only the cost of the selected decision is observed. A bound of $O(T^{2/3})$ on T rounds was first shown by Awerbuch and Kleinberg [2] for the regret against an oblivious adversary and later McMahan and Blum [3] showed that a bound of $O(\sqrt{\ln T}T^{3/4})$ is obtainable against an adaptive adversary.

In this paper we study Kalai and Vempala's model from the viewpoint of bandit algorithms. We show that the algorithm of McMahan and Blum attains a regret of $O(T^{2/3})$ against an oblivious adversary. Moreover, we show a tighter $O(\sqrt{m \ln m}\sqrt{T})$ bound for the expert setting using m experts.

1 Introduction

An *online algorithm* performs a sequence of tasks without any information on the future conditions, while an *offline algorithm* has the full information on them at its disposal. Online algorithms are often analyzed against an adversary, which gets to pick the next task (with the intention to hurt the algorithm as much as possible) after observing the preceding actions of the algorithm. Moreover, their performance is compared to that of offline algorithms. There are also differences in how much information does the online algorithm receive. Typically a cost is assigned to the decisions taken by the algorithm.

One method of ranking online algorithms is to compare their cost to the best static object, i.e., to a situation where the offline algorithm consistently chooses a single static output. Of course, the applicability of this model depends on the specific problem. For example, a binary search tree algorithm could compare its cost to that of the best static tree for the access sequence [4]. It can be argued that a static distribution models well some real world situations.

Kalai and Vempala [1] showed that many online decision problems have efficient solutions, given an oracle for the offline version of the problem. The problems they consider are those where the decisions can be thought of as vectors $x \in \mathbb{R}^n$ and the cost is a linear function on decision vectors. We will adopt their model and notation, which are as follows.

Throughout this article we denote the dimension of the decision and cost vector space by n. At each time step $t \in \{1, \dots, T\}$ a decision vector x_t is chosen from a set $S_D \subset \mathbb{R}^n$. Then a cost vector c_t from a set $S_C \subset \mathbb{R}^n$ is revealed, and the algorithm suffers a cost $c_t \cdot x_t$. The cost is an Euclidean dot product of the cost vector and the decision vector. Thus, the cost function captures redundancy in problems where the costs for different decisions can be linearly related.

Kalai and Vempala [1] (re)introduced the *Follow the Perturbed Leader* (FPL) algorithm, which selects a decision based on past cost vectors. Originally this method was proposed by Hannan [5] already in 1957. In short, the FPL selects the best decision for the past cost vectors which are perturbed by a random vector (hence the name). Note that if the set of decisions is discrete then a randomized algorithm must be used, because a deterministic algorithm can always be misguided. For example, if there are two experts A and B that incur a cost in sequential turns $A, B, A, B, A, B, \dots$ then choosing the expert with the lower cost results in a high regret. In a convex decision space deterministic algorithms can be competitive. In deed, the algorithm of Zinkevich [6] deterministically solves a convex optimization problem, where both the decision space and cost function are convex.

A substantial amount of research has been devoted to *expert* algorithms [7], where no assumption on the linear nature of the cost function is made. One can rather assume the experts (and their costs) to be independent of each other. In the expert setting one needs to have an explicit representation for the decision vector space S_D, while in online decision problems one only requires access to an efficient oracle for choosing the best decision in hindsight. For example, the *weighted majority* algorithm of Littlestone and Warmuth [8] can be used to solve the same problems as FPL if the set of decision vectors is finite.

However, a more specialized algorithm has advantages for example in the *shortest online routing* problem. In this problem there is a directed acyclic graph with a start node and an end node. At each turn an algorithm selects a route from the start node to the end node based on past observations of the costs on edges. The weighted majority algorithm may be applied to the shortest online routing by considering each route as an individual expert. However, in the worst case the number of routes is exponential in the number of edges and, thus, the computational complexity becomes intractable. Furthermore, one can expect that the internal structure of the shortest online routing results in a lower regret, because different routes may share edges and their costs are related. In deed, the weighted majority algorithm can be applied in an efficient manner to a problem that is very similar to the shortest online routing [9].

The shortest online routing problem can be solved by the FPL algorithm by considering each edge as an element in a vector. An element is either zero or one depending on whether the edge was used on the route or not. Thus, the cost vector consists of the edge weights and only these need to be saved for future use, in contrast to the weighted majority algorithm, which stores a weight for each route.

Let us introduce some further notation. There is an oracle $M\colon S_C \to S_D$ such that $M(c)$ is the optimal decision for a cost vector c. We abbreviate the sum of costs up to time step t as $c_{1:t} = \sum_{i=1}^{t} c_i$. Hence, if T is the total number of rounds, then $M(c_{1:T})$ is the best static decision and $M(c_{1:T}) \cdot c_{1:T}$ is its cost. The *regret* of an algorithm is the cost difference between the best static decision $M(c_{1:T})$ and the algorithm. The sets of decision and cost vectors have to be restricted, because otherwise it is possible to suffer an unbounded amount of regret during a single turn. There is a bound on the diameter of the decision set $\|x_t - x'_t\|_1 \leq D$ for all $x_t, x'_t \in S_D$, the cost set $\|c_t\|_1 \leq A$ for any $c_t \in S_C$, and the maximum cost $|c_t \cdot x_t| \leq R$.

With the above restrictions the FPL algorithm achieves a cost of

$$(1+\epsilon)M(c_{1:T}) \cdot c_{1:T} + \frac{D \ln n}{\epsilon}$$

for a parameter ϵ. Thus, the regret of the FPL algorithm is $O(\sqrt{T})$ when a suitable ϵ is plugged into the above formula. This means that the decisions selected by the FPL algorithm are in the long run almost as good as the best decision chosen consistently.

Recall that the FPL requires all past information to be known, i.e., all previous cost vectors. In a natural generalization of the problem the whole cost vector is not observed. For example, in the shortest online routing with end-to-end feedback [2] only the cost on the chosen route is given. Similar restrictions have been studied in the expert setting, where one expert is selected at each turn from a pool of experts and the cost $0 \leq c \leq 1$ of the selected expert is suffered. In a restricted setting only the cost of the selected expert is observed. In this situation the algorithm of Auer et al. [10] obtains a regret of $O(\sqrt{Tm \ln m})$ for m experts. Note that we use m to indicate the number of the experts (or decision vectors) and n to indicate the dimension of some particular problem as described before. We are especially interested in the case where $n \ll m$.

Awerbuch and Kleinberg [2] consider the online decision situation, where only the cost $c_t \cdot x_t$ is observed after a decision x_t. Their algorithm uses the FPL algorithm as a black-box and estimates the cost vector from independent samples. This setting can be split into two separate problems, first how to estimate the cost vector, and second how to select and analyze the decision based on the estimates of the cost vectors. This situation is the *bandit FPL*. The algorithm of Awerbuch and Kleinberg has a regret of the order $O(Rn^{5/3} \ln^{1/3} n\, T^{2/3})$ against an *oblivious* adversary, which decides the cost vectors in advance without knowledge of the decisions made by the algorithm. McMahan and Blum [3] obtained

Table 1. Summary of earlier bounds for the regret in terms of the length of the sequence T and our results.

Paper	Adversary Model	Setting	Regret Bound
Auer et al. [10]	Adaptive	Bandit, Expert	$O(\sqrt{Tm\ln m})$
Kalai and Vempala [1]	Adaptive	Transparent	$O(\sqrt{DRAT})$
Awerbuch and Kleinberg [2]	Oblivious	Bandit	$O(Rn^{5/3}\ln^{1/3} n\, T^{2/3})$
McMahan and Blum [3]	Adaptive	Bandit	$O(\sqrt{\ln T}\; T^{3/4})^a$
This paper	Oblivious	Bandit, Expert	$O(\sqrt{Tm\ln m})$
	Oblivious	Bandit	$O(Rn(DA\ln n)^{1/3}T^{2/3})$

[a] Polynomial dependence on parameters is not shown

a regret of $O(\sqrt{\ln T}\; T^{3/4})$ for the bandit FPL against an *adaptive* adversary, which may use the past decisions $x_1, \ldots, x_{t-1}$ of the algorithm when choosing the cost vector c_t.

Flaxman, Kalai, and McMahan [11] analyze a bandit algorithm for the Zinkevich's [6] convex optimization algorithm, proving a regret of $O(T^{5/6})$ against an oblivious adversary. Hutter and Poland [12] study an *adaptive FPL*, i.e., a version of FPL in which the total number of rounds T is not known. Additionally, they show that many of the results obtained in the expert setting can be elegantly proved also in the FPL setting, albeit with worse constants.

Table 1 summarizes the relevant results from previous work and puts the results derived in this paper into the map.

In this paper we analyze one way of implementing a bandit algorithm for the FPL. Section 2 gives a general condition for unbiased estimates for the cost vector c_t when a decision x_t has been made and the cost $c_t \cdot x_t$ has been observed. In Sections 3–4 we analyze this estimate. Our results apply against an oblivious adversary. In particular, Section 3 shows how to use the framework of Kalai and Vempala to derive an upper bound for the bandit FPL. In Section 3.1 we proceed to derive a $O(\sqrt{Tm\ln m})$ regret in the expert setting. This is to the best of our knowledge the first time that a $O(\sqrt{T})$ regret has been demonstrated in a bandit FPL. Finally, in Section 4 we show an upper bound $O(T^{2/3})$ that holds in the general case. This upper bound can be applied to the algorithm analyzed by McMahan and Blum [3], thus improving their bound $O(\sqrt{\ln T}\; T^{3/4})$ against an adaptive adversary to $O(T^{2/3})$ against an oblivious adversary (which, of course, is not necessarily an improvement).

2 An Unbiased Estimate for a Cost Vector

Awerbuch and Kleinberg [2] as well as McMahan and Blum [3] mention one of problems of the bandit FPL to be a lack of a good way of estimating the cost vector. In this section we will study a one kind of unbiased estimate for the cost vector in a restricted setting. We consider the case where the number of decision vectors m is finite. Let $x_1, \dots, x_m$ be the decision vectors of the problem and define a matrix $X = [x_1, \dots, x_m]$ to contain the decision vectors as columns.

At time t we choose a decision x_t from a distribution given by probabilities $p(x_t) > 0$ (recall that randomization is necessary) and then observe the cost $c_t \cdot x_t$. We limit ourselves to the situation in which the cost vector c_t is estimated independently at each turn with some unbiased $\hat{c}_t$ and the cumulative estimate is simply the additive $\hat{c}_{1:t}$. Thus an estimate $\hat{c}_t$ of c_t can be based only on the selected decision x_t, the probability distribution on decisions, and the cost $c_t \cdot x_t$. We assume that $\hat{c} = (c \cdot x / p(x))\, v_x$, where $v_x \in \mathbb{R}^n$ is the direction vector of the estimate. Let the directions v_x make up a matrix $V = [v_1, \dots, v_m]$. Then

$$\mathbb{E}\left(\hat{c}\right) = \sum_{x \in X} \left(c \cdot x\right) v_x = \sum_{x \in X} v_x x^T c = V X^T c. \tag{1}$$

This implies that ideally we should choose V so that $V X^T = I_{n \times n}$. However, the matrix X does not necessarily have a rank of n. Then

$$V X^T = P_{R(X^T)}, \tag{2}$$

where $P_{R(X^T)}$ is a projection matrix to the column space of X, is a sufficient condition to obtain an unbiased $\hat{c}$. Such an estimate is unbiased in the sense that the part of the cost vector c, which is in the null space of the matrix X^T, does not matter when calculating the actual costs. Another possibility is to first make the matrix X full rank by projecting it to a suitable subspace [3]. Both approaches require $R(X^T)$ to be known.

Kalai and Vempala's [1] approach resulted in an efficient solution given $M(c)$. To have an efficient procedure for estimation, we need to assume that the probabilities $p(x_t)$ can be obtained, and that an solution for the matrix V is found. Previous work [3, 2] has estimated the cost vector from independent samples on a certain subset of decision vectors, so the probabilities have been uniform, and hence there has not been a need for more complex estimation of the probabilities on decisions. Our estimates may be applied similarly, or in any other way as long as it is possible to obtain the values $p(x_t)$.

Choosing the matrix V is a separate problem from estimating the probabilities. We show one method for selecting a matrix V efficiently in the routing case. Although $X \in R^{n \times m}$, it is possible to calculate the matrix $XX^T \in R^{n \times n}$ in time polynomial in n. This is because $(XX^T)_{ij}$ is the number of routes that share both edges i and j. Now if X has a rank of n then $V = (XX^T)^{-1} X$ is one possible solution to equation (2). In this case $v_x = (XX^T)^{-1} x$. On the other hand, if X is not full rank, we may select $V = \left(XX^T\right)^+ X$, where $(XX^T)^+$ is *the Moore-Penrose matrix inverse* of the matrix XX^T, which can be calculated efficiently using the singular value decomposition. The Moore-Penrose matrix inverse [13, 14] for a $n \times m$ matrix A is a unique matrix pseudoinverse A^+ such that:

1. $AA^+A = A$,
2. $A^+AA^+ = A^+$,

Table 2. The bandit FPL algorithm considered in this paper.

Set v_x to be the column corresponding to the decision x in the matrix V satisfying a formula $VX^T = P_{R(X^T)}$ (bounds will depend on the particular choice). For example, $v_x = (XX^T)^+ x$ or a choice based on barycentric spanner [2].
On time step t

1. **Sampling step:** With probability γ obtain a decision independent of $\hat{c}_{1:t-1}$ such that the probability of sampling a decision x_t is proportional to $\|v_{x_t}\|_1$.
   Exploitation step: With probability $1-\gamma$ the decision $M(\hat{c}_{1:t-1} + r_t)$ is used.
2. Set
$$\hat{c}_t = \frac{c_t \cdot x_t}{p(x_t)} v_{x_t}$$
   to be the estimate of the c_t, where $p(x_t)$ is the probability of choosing the decision x_t (both sampling and exploitation steps could be taken into consideration).

---

3. $(AA^+)^T = AA^+$, and
4. $(A^+A)^T = A^+A$.

It is well known that these conditions imply that AA^+ is the projection matrix to the column space of a matrix A. Thus, because the column spaces of X and XX^T are the same, and XX^T is a symmetrix matrix,

$$V^T X = \left(XX^T\right)^+ XX^T = XX^T \left(XX^T\right)^+ = P_{R(X^T)}$$

as was required. Actually, $(XX^T)^{-1} = (XX^T)^+$ if X is full rank, so we would not have needed to handle separately the case when X is full rank.

In general, equation (2) gives us some flexibility on choosing the estimate $\hat{c}_t$. In the following chapters we give conditions when an estimate based on a matrix V results in a low regret.

3 An Upper Bound for a Bandit FPL Algorithm

Let us first describe the algorithm that we use. At time t do a *sampling* step with probability γ and an *exploitation* step with probability $1-\gamma$. On a sampling step an independent estimation of the cost vector should be obtained, in case of independent experts this could be drawn from a uniform distribution. On an exploitation step choose a decision $M(\hat{c}_{1:t-1} + r_t)$, where r_t is a random perturbation vector. The probabilities for decisions must be such that the possible decisions span the column space of X, otherwise it is not possible to estimate an unbiased cost vector c_t. Table 2 summarizes the algorithm that will be analyzed in the following.

We derive an upper bound for the regret of the above algorithm against an oblivious adversary. The bound holds under expected value. Recall that R is the maximum cost. For the sampling steps the trivial γRT bound is enough. For the exploitation steps we use the following two lemmas that Kalai and Vempala [1] proved.

Lemma 1. *For all $c_1, \ldots, c_T$ and for all $p_0 = 0, p_1, \ldots, p_T$ it holds that*

$$\sum_{t=1}^{T} M(c_{1:t} + p_t) \cdot c_t \leq M(c_{1:T}) \cdot c_{1:T} + \sum_{t=1}^{T} \left(M\left(c_{1:T}\right) - M\left(c_{1:t} + p_t\right)\right) \cdot (p_t - p_{t-1}).$$

Lemma 2. *Recall A is the maximum one norm of any cost vector. The following holds when items of a perturbation vector are distributed according to the standard double exponential distribution $\epsilon/2\ \exp(-\epsilon|x|)$:*

$$\sum_{t=1}^{T} M(c_{1:t-1} + p_t) \cdot c_t \leq \sum_{t=1}^{T} M(c_{1:t} + p_t) \cdot c_t + \epsilon RAT.$$

The intuitive picture is that Lemma 1 can be used to estimate how much the difference $\hat{c}_{1:t} - c_{1:t}$ matters. Thus, by setting p_t to be $\hat{c}_{1:t-1} - c_{1:t-1} + r_t$, which is the estimation error plus the random perturbation, we obtain by Lemma 1 the following inequality.

$$\begin{aligned}
&\sum_{t=1}^{T} M(\hat{c}_{1:t-1} + c_t + r_t) \cdot c_t \\
&\leq M(c_{1:T}) \cdot c_{1:T} \\
&\quad + \sum_{t=1}^{T} \left(M\left(c_{1:T}\right) - M\left(\hat{c}_{1:t-1} + c_t + r_t\right)\right) \cdot \left(\hat{c}_{t-1} - c_{t-1} + r_t - r_{t-1}\right). \quad (3)
\end{aligned}$$

There are random variables r_t and $\hat{c}_{1:t-1}$ in the equation. Note that $\hat{c}_{1:t-1}$ depends on the variables $r_1, \ldots, r_{t-1}$. As we are going to take expected value over the equation we can replace the variable r_t with a variable s_t that is similarly distributed. We do not replace the variables $r_1, \ldots, r_{t-1}$ within the variable $\hat{c}_{1:t-1}$. This is possible by linearity of expectation. We can now apply the trick used by Kalai and Vempala [1]: as expected value is linear, we can correlate the values s_t by setting $s_t = s_1$ for all t, which eliminates some terms.

Now we proceed to the following chain of equations for the cost of the exploitation steps. Recall that D is the diameter of the decision space.

$$
\begin{aligned}
&\mathbb{E}\left(\sum_{t=1}^{T} M\left(\hat{c}_{1:t-1}+s_t\right)\cdot c_t\right)\\
&\leq \mathbb{E}\left(\sum_{t=1}^{T} M\left(\hat{c}_{1:t-1}+c_t+s_t\right)\cdot c_t\right)+\epsilon RAT && (4)\\
&\leq M(c_{1:T})\cdot c_{1:T}+\epsilon RAT\\
&\quad +\mathbb{E}\left(\sum_{t=1}^{T}\left(M\left(c_{1:T}\right)-M\left(\hat{c}_{1:t-1}+c_t+s_t\right)\right)\cdot\left(\hat{c}_{t-1}-c_{t-1}+s_t-s_{t-1}\right)\right) && (5)\\
&\leq M(c_{1:T})\cdot c_{1:T}+\frac{D(1+\ln n)}{\epsilon}+\epsilon RAT\\
&\quad +\mathbb{E}\left(\sum_{t=1}^{T}\left(M\left(c_{1:T}\right)-M\left(\hat{c}_{1:t-1}+c_t+s_t\right)\right)\cdot\left(\hat{c}_{t-1}-c_{t-1}\right)\right). && (6)
\end{aligned}
$$

Equation (4) follows from Lemma 2 and equation (5) follows by inequality (3). Equation (6) again follows from the arguments of Kalai and Vempala [1]: as $s_t = s_1$, we only need to worry of the first term of the sum, which they upper bounded by $\mathbb{E}\left(D\|s_1\|_\infty\right) \leq D(1+\ln n)/\epsilon$.

Now the problem has reduced to solving the expected value on the right-hand side of equation (6). As we are using the oblivious adversary model, the vectors c_t are not random variables and thus we simply calculate

$$\mathbb{E}\left(M\left(c_{1:T}\right)\cdot\left(\hat{c}_{t-1}-c_{t-1}\right)\right)=M(c_{1:T})\cdot\left(\mathbb{E}\left(\hat{c}_{t-1}\right)-c_{t-1}\right)=0.$$

We are finally left with

$$\sum_{t=1}^{T} M(\hat{c}_{1:t-1}+c_t+s_t)\cdot(c_{t-1}-\hat{c}_{t-1}). \tag{7}$$

This term cannot be bounded elementarily, because both sides of the dot product are dependent random variables.

3.1 Restriction to the Experts Setting

The simplest application of the bandit FPL is the expert setting, where there are m independent experts. Let us examine this framework for a while. Similar techniques as applied here are subsequently used to prove the more general Theorem 2. The expert situation can be described in the general setting by selecting the possible decisions to be the standard basis, i.e., vectors e_i which form the identity matrix I. Naturally, the number of dimensions n equals the number of experts m in this case. Thus, $\hat{c}_t$ now reduces to $(c_t/p(x_t))\, M\left(\hat{c}_{1:t-1}+s_t\right)$, as can be expected if one is familiar with prior work [10].

Theorem 1. *In the expert setting the estimation regret of the exploitation steps is upper bounded by*

$$\mathbb{E}\left(M\left(\hat{c}_{1:t-1} + c_{t-1} + s_t\right) \cdot \left(c_{t-1} - \hat{c}_{t-1}\right)\right) \leq \epsilon \left(2\left(\|c_{t-1}\|_1 + 1\right) + \|c_{t-1}\|_2^2\right).$$

Proof. Let $p_t(i)$ be the probability that the ith expert is chosen at time t and let $p_t(i \mid j)$ be the probability that it is chosen at time t given that the jth expert was chosen at time $t-1$. Let c_t^i be the ith element of the cost vector c_t. Under expectation one term of the sum (7) simplifies in the expert setting to

$$\begin{aligned}
&\mathbb{E}\left(M\left(\hat{c}_{1:t-1} + c_{t-1} + s_t\right) \cdot \left(c_{t-1} - \hat{c}_{t-1}\right)\right) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{m} c_{t-1}^{j} p_t(i \mid j) p_{t-1}(j) - \sum_{i=1}^{m} c_{t-1}^{i} p_t(i \mid i). \qquad (8)
\end{aligned}$$

In order to upper bound equation (8) we use a technique similar to the one employed by Kalai and Vempala [1]. The probability of selecting expert i during an exploitation step is the probability that the used cost vector plus random perturbation vector will lie in the area of space where the ith coordinate is smaller than other coordinates. Let us denote this area by $S_i \subseteq \mathbb{R}^n$. For example, in case of two experts there are two areas S_1 and S_2 which are separated in $\mathbb{R}^2$ by the diagonal line $x = y$. The probability $p_t(i \mid j)$ in terms of the change $\delta = \hat{c}_{t-1}$ in the cost vector $b = \hat{c}_{1:t-2}$ is

$$\begin{aligned}
p_t(i \mid j) &= \frac{\gamma}{m} + (1-\gamma) \int_{S_i} \left(\frac{\epsilon}{2}\right)^n \exp(-\epsilon \|x - (b - \delta)\|_1) \mathrm{d}x \\
&\leq \frac{\gamma}{m} + (1-\gamma) \exp(\epsilon \|\delta\|_1) \int_{S_i} \left(\frac{\epsilon}{2}\right)^n \exp(-\epsilon \|x - b\|_1) \mathrm{d}x \\
&\leq \exp(\epsilon \|\delta\|_1) p_{t-1}(i), \qquad (9)
\end{aligned}$$

where the inequality follows by the triangle inequality. Similarly $p_t(i \mid j)$ may be lower bounded:

$$p_t(i \mid j) \geq \exp(-\epsilon \|\delta\|_1) p_{t-1}(i).$$

We assume that $\epsilon c_i / p_i \leq 1$, which implies that the sampling probability γ should be chosen such that $\epsilon n c_i \leq \gamma$. Then

$$\begin{aligned}
p_t(i \mid j) &\leq \exp\left(\epsilon \left(\frac{c_{t-1}^i}{p_{t-1}(i)} + \|c_{t-1}\|\right)\right) p_{t-1}(i) \\
&\leq 1 + 2\epsilon \left(\frac{c_{t-1}^i}{p_{t-1}(i)} + \|c_{t-1}\|\right) p_{t-1}(i). \qquad (10)
\end{aligned}$$

Now, the left-hand term of equation (8) can be upper bounded as

$$\begin{aligned}
&\sum_{i=1}^{m}\sum_{j=1}^{m} c_{t-1}^{i} p_t(i \mid j) p_{t-1}(j) \\
&\leq \sum_{i=1}^{m}\sum_{j=1}^{m} c_{t-1}^{i} \exp\left(\epsilon c_{t-1}^{j}\left(1+\frac{1}{p_{t-1}(j)}\right)\right) p_{t-1}(i) p_{t-1}(j) \\
&\leq \sum_{i=1}^{m} c_{t-1}^{i} p_{t-1}(i)\left(\sum_{j=1}^{m} p_{t-1}(j) + 2\epsilon c_{t-1}^{j}\left(1+p_{t-1}(j)\right)\right) \\
&= \sum_{i=1}^{m} c_{t-1}^{i} p_{t-1}(i) + 2\epsilon\left(\|c\|_{t-1}+1\right).
\end{aligned}$$

Similarly, the right-hand term of equation (8) can be lower bounded as

$$\sum_{i=1}^{n} c_{t-1}^{i} p_t(i \mid i) \geq \sum_{i=1}^{n} c_{t-1}^{i} p_{t-1}(i) - \epsilon \|c_{t-1}\|_2^2.$$

Hence, the estimation regret obtained is $\epsilon\left(2\left(\|c_{t-1}\|_1+1\right)+\|c_{t-1}\|_2^2\right)$ as claimed, thus, completing the proof.

Theorem 1 can be interpreted to give the regret we suffer from not knowing the true c_t at each turn. All in all, the total regret in the expert setting can be upper bounded by

$$\frac{D(1+\ln n)}{\epsilon} + \epsilon RAT + \gamma RT + \epsilon\left(2\left(\|c_t\|_1+1\right)+\sum_{t=1}^{T}\|c_t\|_2^2\right).$$

We were analyzing the expert setting so $A = m$, $R = 1$, $D = 2$, $\|c_t\|_1 \leq m$, and $\|c_t\|_2^2 \leq m$. Plugging in a good ϵ and γ we get an upper bound of

$$4\sqrt{5}\sqrt{(1+\ln m)mT} + \frac{2\sqrt{1+\ln m}}{\sqrt{5mT}},$$

which is $O(\sqrt{T})$. As far as we know, this is the first time that a $O(\sqrt{T})$ bound has been demonstrated in a bandit FPL. As a side note, we failed to analyze the expert setting when the perturbation distribution was uniform, because if the probabilities are in the order of $\sqrt{\epsilon}$, a high regret occurs when this kind of analysis is used.

4 The General Case

Now we assume that the decision vectors may be arbitrary and prove a weaker bound for the regret than in the expert setting. Equation (2) gave a necessary condition for an unbiased estimate for c_t. However, an unbiased estimate as such does not guarantee a low regret, because the obtained estimates may be far from the true cost vector; i.e., the variance of the estimate is high. To analyze the regret we set a bound for the column vectors v_x of the matrix V that is $\sum_{y\in X}\|v_y\|_1 \leq E$. Intuitively, this restriction ensures that $\mathbb{E}\left(\|\hat{c}_t\|_1\right)$ is at most proportional to E. In the next paragraphs it is argued that we can assume E to be polynomial in n (in fact linear, if we do a certain change in the basis).

Recall the definition of the matrix V: $VX^T = P_{R(X^T)}$. If v_i is the ith *row* of the matrix V, then $Xv_i = P^i_{R(X^T)}$, where $P^i_{R(X^T)}$ is the ith column of the symmetric projection matrix $P_{R(X^T)}$. Now, if the vector $P^i_{R(X^T)}$ belongs to the convex closure spanned by the column vectors of X (which are the decision vectors), then there are solutions in which $\|v_i\|_1 = 1$. Of course, this does not necessarily happen, but we can conclude that the norm $\|v_i\|_1$ is a property of the relative locations of the decision vectors, not a function on the number of decisions m (if m is large relative to n). Even better, the decision vectors can be transformed to a base where we can select $E = n$.

Recall that the *barycentric spanner* introduced by Awerbuch and Kleinberg [2] demonstrates that any decision vector can be presented as a linear combination of n decision vectors, where the coefficients are in $[-1, 1]$. The n decision vectors are called the barycentric spanner, and we choose columns of a matrix X_B to be these decision vectors. If the decision vectors are transformed to a base, where e_i corresponds to the ith decision vector in the barycentric spanner, then we can choose v_i so that $\|v_i\|_1 = 1$, because we can choose v_i to be one for the ith decision in the barycentric spanner and zero otherwise. Thus, $E = n$ in this case. If we are forced to use this trick with the barycentric spanner to bound E, then we estimate only using the vectors in the barycentric spanner. Additionally, we need to be sure that the bounds D, A, and R do not change too much when moving to the base formed with the barycentric spanner. If we denote by D', R' and A' the new bounds when the decision vectors are transformed by X_B, then $D' \leq 2n$, $R' = R$, and $A' = \max_{c'}\|c'\|_1 = \max_c\|X_Bc\|_1 \leq DA$.

Finally, note that even though the vectors are projected to a base spanned by the barycentric spanner, the problem can be solved without projecting the decision vectors to this space by transforming the perturbation vector r_t to $X_B^{-1}r_t$. Then the oracle $M(c)$ selects always the same decisions and thus the regret of both the projected and the original algorithm is the same. Thus, we can analyze an algorithm in the barycentric spanner, but in practice we do not need to project all decision vectors to the space spanned by the barycentric spanner.

The bound in the following theorem depends on the fact that $\epsilon\|v_x\|_1 \leq p(x)$, because then $\exp(\epsilon\|\hat{c}_t\|_1)$ can be upper bounded additively. The assumption $\epsilon\|v_x\|_1 \leq p(x)$ can be satisfied by lower bounding $p(x)$ by choosing our sampling probability for a vector x to be $\gamma\|v_x\|_1/\sum_{y\in X}\|v_y\|_1$. The following theorem gives the bound that results.

Theorem 2. *If a decision $M\left(\hat{c}_{1:t-1} + r_t\right)$ is used on exploitation steps with an estimation method $\hat{c}_t(x_t) = \left(c_t \cdot x_t/p(x_t)\right)v_{x_t}$ and sampling probability for a de-*

cision x is $\gamma\|v_x\|_1/\sum_{y\in X}\|v_y\|_1$, then the estimation regret of an exploitation step is upper bounded by:

$$\mathbb{E}\left(M\left(\hat{c}_{1:t-1}+c_{t-1}+s_t\right)\cdot\left(c_{t-1}-\hat{c}_{t-1}\right)\right)\leq 2\epsilon(R^2E+RA)+\frac{2\epsilon}{\gamma}(R^2AE^2+RAE).$$

The proof of this theorem is given in Appendix A. The idea behind the proof is similar to the one given for the expert setting. The total regret over T time steps is:

$$\frac{D(1+\ln n)}{\epsilon}+\epsilon RAT+\gamma RT+2\epsilon R^2ET+\frac{2\epsilon}{\gamma}R^2AE^2T.$$

Selecting

$$\epsilon=2^{-1/3}(1+\ln n)^{2/3}D^{2/3}A^{-1/3}R^{-1}E^{-2/3}T^{-2/3}$$

and

$$\gamma=2^{1/3}(1+\ln n)^{1/3}D^{1/3}A^{-1/3}R^{-1}E^{2/3}T^{-1/3}$$

leads to a regret in which the highest power of T is in the term

$$3\;2^{1/3}R\left(D\left(1+\ln n\right)AE^2\right)^{1/3}T^{2/3}.$$

If the decision vectors are transformed to a basis spanned by a barycentric spanner, then the regret is

$$3\;4^{1/3}R\left(D\left(1+\ln n\right)A\right)^{1/3}nT^{2/3}. \tag{11}$$

This bound can be applied to the algorithm of McMahan and Blum [3], which they analyzed to have a regret of $O(\sqrt{\ln T}\;T^{3/4})$ against an adaptive adversary. However, some caution is required. Their estimation is based on sampling the decisions in the barycentric spanner and setting the estimate $\hat{c}_t$ to $n(\text{loss occurred})/\gamma X_B^{-1}e_i$ if the ith decision in the barycentric spanner was chosen on a sampling step. On exploitation steps they set the estimate $\hat{c}_t$ to zero. Theorem 2 applies if the denominator in the estimate is the actual probability $p(x_t)$ of choosing the decision in the barycentric spanner, not γ/n. Additionally, the perturbation vector r_t has to be transformed to $X_B^{-1}r_t$, as argued before. Now, a closer investigation to the proof of Theorem 2 reveals that if the first term $2\epsilon(R^2E+RA)$ in the bound is divided by γ/n, then the bound applies. This modification to the bound has a small effect of increasing the constant in bound (11). Hence, the regret is $O(T^{2/3})$ against an oblivious adversary.

5 Conclusions and Discussion

We have studied the performance of the bandit FPL algorithm, which can be used to solve an interesting class of problems, for example the shortest online routing with end-to-end feedback. Although this paper did not give a full solution to this problem, we were able to make some progress. In the standard multi-armed bandits problem $O(\sqrt{T})$ regret bounds are attainable, and they are also asymptotically tight [10]. Our work implies that this is also true when the FPL algorithm is applied to the expert setting, although this is not very surprising. The general upper bound we proved implies that the algorithm of McMahan and Blum [3] has a regret of the order $O(T^{2/3})$ against an oblivious adversary. These results can be seen as progress towards better understanding of the relationship between linear dependence of the cost function on decisions and the regret that can be achieved using the bandit FPL algorithms.

Several questions remain to be solved in future work. First, can our work be applied against an adaptive adversary? Second, what kinds of regrets are actually achievable in the setting considered in this paper?

Acknowledgments

This work was supported by Academy of Finland project "INTENTS: Intelligent Online Data Structures". Moreover, the work of J. Kujala is financially supported by Tampere Graduate School in Information Science and Engineering (TISE).

References

1. Kalai, A., Vempala, S.: Efficient algorithms for online decision problems. In Schölkopf, B., Warmuth, M.K., eds.: Proceeding of the Sixteenth Annual Conference on Computational Learning Theory. Volume 2777 of Lecture Notes in Computer Science., Berlin, Heidelberg, Springer (2003) 26–40
2. Awerbuch, B., Kleinberg, R.D.: Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In: Proceeding of the Thirty-Sixth Annual ACM Symposium on Theory of Computing, New York, NY, ACM Press (2004) 45–53
3. McMahan, H.B., Blum, A.: Geometric optimization in the bandit setting against an adaptive adversary. In Shawe-Taylor, J., Singer, Y., eds.: Proceeding of the Seventeenth Annual Conference on Learning Theory. Volume 3120 of Lecture Notes in Computer Science., Berlin, Heidelberg, Springer (2004) 109–123
4. Sleator, D., Tarjan, R.: Self-adjusting binary search trees. Journal of the ACM **32**(3) (1985) 652–686
5. Hannan, J.: Approximation to Bayes risk in repeated plays. In Dresher, M., Tucker, A., Wolfe, P., eds.: Contributions to the Theory of Games. Volume 3. Princeton University Press, Princeton, NJ (1957) 97–139
6. Zinkevich, M.: Online convex programming and generalized infinitesimal gradient ascent. In Fawcett, T., Mishra, N., eds.: Proceeding of the Twentieth International Conference on Machine Learning, Menlo Park, CA, AAAI Press (2003) 928–936
7. Cesa-Bianchi, N., Freund, Y., Haussler, D., Helmbold, D., Schapire, R.E., Warmuth, M.K.: How to use expert advice. Journal of the ACM **44**(3) (1997) 427–485
8. Littlestone, N., Warmuth, M.K.: The weighted majority algorithm. Information and Computation **108**(2) (1994) 212–261
9. Takimoto, E., Warmuth, M.K.: Path kernels and multiplicative updates. Journal of Machine Learning Research **4** (2003) 773–818
10. Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E.: The non-stochastic multiarmed bandit problem. SIAM Journal on Computing **32**(1) (2002) 48–77

11. Flaxman, A.D., Kalai, A.T., McMahan, H.B.: Online convex optimization in the bandit setting: Gradient descent without a gradient. In: Proceeding of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, New York, NY, ACM Press (2005) 385–394
12. Hutter, M., Poland, J.: Adaptive online prediction by following the perturbed leader. Journal of Machine Learning Research **6** (2005) 639–660
13. Moore, E.H.: On the reciprocal of the general algebraic matrix (abstract). Bulletin of the American Mathematical Society **26** (1920) 394–395
14. Penrose, R.A.: A generalized inverse for matrices. Proceedings of the Cambridge Philosophical Society **51** (1955) 406–413

A Proof of Theorem 2

Proof. We proceed with a straightforward but technically complex evaluation of the expected value.

Similarly as in the expert setting, the probability $p_t(x_t \mid x_{t-1})$ can be upper bounded by

$$\exp\left(\epsilon \|\hat{c}_{t-1}(x_{t-1}) + c_{t-1}\|_1\right) p_{t-1}(x_t)$$

and, respectively, lower bounded by

$$\exp\left(-\epsilon \|\hat{c}_{t-1}(x_{t-1}) + c_{t-1}\|_1\right) p_{t-1}(x_t).$$

With the assumption that $\epsilon\|\hat{c}_{t-1}(x_{t-1}) + c_{t-1}\|_1 \le 1$ and using the properties of the function exp, the bounds above imply the following additive bounds:

$$\begin{aligned}
&p_{t-1}(x_t)(1 - \epsilon\|\hat{c}_{t-1}(x_{t-1}) + c_{t-1}\|_1) \\
&\le p_t(x_t \mid x_{t-1}) \\
&\le p_{t-1}(x_t)(1 + 2\epsilon\|\hat{c}_{t-1}(x_{t-1}) + c_{t-1}\|_1). \qquad (12)
\end{aligned}$$

Now we bound the estimation regret $\mathbb{E}\left(M\left(\hat{c}_{1:t-1} + c_{t-1} + s_t\right) \cdot (c_{t-1} - \hat{c}_{t-1})\right)$. First we upper bound $\mathbb{E}\left(M(\hat{c}_{1:t-1} + c_{t-1} + s_t) \cdot c_{t-1}\right)$.

$$\begin{aligned}
&\mathbb{E}\left(M\left(\hat{c}_{1:t-1} + c_{t-1} + s_t\right) \cdot c_{t-1}\right) \\
&= \sum_{x_{t-1}\in X} \mathbb{E}\left(M\left(\hat{c}_{1:t-2} + \hat{c}_{t-1}(x_{t-1}) + c_{t-1} + s_t\right) \cdot c_{t-1}\right) p_{t-1}(x_{t-1}) \qquad (13) \\
&= \sum_{x_{t-1}\in X} p_{t-1}\left(x_{t-1}\right) \sum_{x_t\in X} \left(\left(c_{t-1} \cdot x_t\right) p_t\left(x_t \mid x_{t-1}\right)\right) \qquad (14) \\
&\le \sum_{x_{t-1}\in X} p_{t-1}\left(x_{t-1}\right) \sum_{x_t\in X} \left(\left(x_t \cdot c_{t-1}\right) p_{t-1}\left(x_t\right)\right) \\
&\quad + 2\epsilon \sum_{x_{t-1}\in X} p_{t-1}\left(x_{t-1}\right) \|\hat{c}_{t-1}(x_{t-1}) + c_{t-1}\|_1 \sum_{x_t\in X} \left(|c_{t-1} \cdot x_t| p_{t-1}\left(x_t\right)\right) \qquad (15) \\
&\le c_{t-1} \cdot \mathbb{E}\left(M\left(\hat{c}_{1:t-2} + s_{t-1}\right)\right) + 2\epsilon\left(R^2 E + RA\right). \qquad (16)
\end{aligned}$$

Equations (13) and (14) result from the definition of the expected value that we are taking, inequality (15) follows from inequalities (12), and inequality (16) follows from $|c_{t-1} \cdot x_t| \leq R$, the definition of the bound E, and the triangle inequality.

Then we lower bound $\mathbb{E}\left(M\left(\hat{c}_{1:t-1} + c_{t-1} + s_t\right) \cdot \hat{c}_{t-1}\right)$.

$$
\begin{aligned}
&\mathbb{E}\left(M\left(\hat{c}_{1:t-1} + c_{t-1} + s_t\right) \cdot \hat{c}_{t-1}\right) \\
&= \sum_{x_{t-1} \in X} \left(p_{t-1}\left(x_{t-1}\right) \hat{c}_{t-1}\left(x_{t-1}\right) \cdot \left(\sum_{x_t \in X} x_t \; p_t\left(x_t \mid x_{t-1}\right) \right) \right) && (17) \\
&\geq \sum_{x_{t-1} \in X} \left(c_{t-1} \cdot x_{t-1}\right) v_{x_{t-1}} \\
&\quad \cdot \sum_{x_t \in X} \left(x_t \left(p_{t-1}\left(x_t\right) \pm 2\epsilon \|\hat{c}_{t-1}\left(x_{t-1}\right) + c_{t-1}\|_1 \; p_{t-1}\left(x_t\right) \right) \right) && (18) \\
&= c_{t-1} \cdot \mathbb{E}\left(M\left(\hat{c}_{1:t-1} + s_{t-1}\right)\right) \\
&\quad - 2\epsilon \sum_{x_{t-1} \in X} \left| \left(c_{t-1} \cdot x_{t-1}\right) v_{x_{t-1}} \cdot \mathbb{E}\left(M(\hat{c}_{1:t-2} + s_{t-1})\right) \right| \\
&\quad \|\hat{c}_{t-1}\left(x_{t-1}\right) + c_{t-1}\|_1 && (19) \\
&\geq c_{t-1} \cdot \mathbb{E}\left(M\left(\hat{c}_{1:t-2} + s_{t-1}\right)\right) - \frac{2\epsilon}{\gamma} \left(R^2 \; AE^2 + RAE\right). && (20)
\end{aligned}
$$

Equation (17) is a definition of the expected value taken, inequality (18) follows from inequalities (12) if the $\pm$ is interpreted as always choosing the sign that results in the worst regret. Equality (19) has its terms calculated open and $\pm$ is transformed to a minus by taking an absolute value of the terms within the sum. Finally, inequality (20) follows from the definitions of different bounds R, A, E, and the fact that $p(x)$ is lower bounded by a sampling probability.

Thus,

$$
\begin{aligned}
&\mathbb{E}\left(M(\hat{c}_{1:t-1} + c_{t-1} + s_t) \cdot (c_{t-1} - \hat{c}_{t-1})\right) \\
&\leq 2\epsilon \left(R^2 E + RA\right) + \frac{2\epsilon}{\gamma} \left(R^2 AE^2 + RAE\right)
\end{aligned}
$$

as claimed.