Following the Perturbed Leader to Gamble at Multi-Armed Bandits

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. Following the perturbed leader (FPL) is a powerful technique for solving online decision problems. Kalai and Vempala [1] rediscovered this algorithm recently. A traditional model for online decision problems is the multi-armed bandit. In it a gambler has to choose at each round one of the k levers to pull with the intention to minimize the cumulated cost. There are four versions of the nonstochastic optimization setting out of which the most demanding one is a game played against an adaptive adversary in the bandit setting. An adaptive adversary may alter its game strategy of assigning costs to decisions depending on the decisions chosen by the gambler in the past. In the bandit setting the gambler only gets to know the cost of the choice he made, rather than the costs of all available alternatives. In this work we show that the very straightforward and easy to implement algorithm Adaptive Bandit FPL can attain a regret of $O(\sqrt{T \ln T})$ against an adaptive adversary. This regret holds with respect to the best lever in hindsight and matches the previous best regret bounds of $O(\sqrt{T \ln T})$.

1 Introduction

Following the perturbed leader (FPL) is a natural and straightforward approach that can be applied in a wide variety of online decision problems, where costs associated with future decisions are not known [1, 2]. In FPL one simply chooses the decision that has fared the best in the earlier rounds (the leader). In order to cope with an adversary, the necessary randomization is implemented by adding a perturbation to the total costs prior to selecting the leader. We have previously shown that FPL will attain an expected regret $O(\sqrt{T})$ in T rounds against an oblivious (non-adaptive) adversary in a bandit version of the expert setting [3].[1]

In this paper we study the performance of the FPL algorithm when faced with an adaptive adversary in the multi-armed bandit model. Dani and Hayes [4] give an example where the EXP3 algorithm of Auer et al. [5] can attain, with a constant probability, a regret of $\Omega(T^{2/3})$ against an adaptive adversary.

[1] In this paper, instead of the bandit version of the expert setting, we talk about the multi-armed bandit model. However, we do not distinguish the two settings in any way.

Table 1. The four variations of the nonstochastic optimization problem.

	Oblivious Adversary	**Adaptive Adversary**
Full Information	OFI	AFI
Bandit Setting	OBS	ABS

Their argument applies as such to our earlier FPL algorithm [3]. However, a small modification to the algorithm lets us circumnavigate this lower bound.

In this paper we introduce a new FPL algorithm that matches the best regret bound $O(\sqrt{T \ln T})$ of the best algorithms for this problem [6, 4]. The new algorithm is the same as our earlier one, except that instead of using an unbiased estimate of the true cost vector, it uses one which guarantees that we almost never estimate the cost of a lever too high. Thus, it is difficult to hide the actual cost under the variance of the estimate of the cost. This method does not produce too much perturbation to the estimated costs, leading to a good regret bound. This technique resembles that of Auer [6], who modified EXP3 algorithm to have $O(\sqrt{T \ln T})$ regret against an adaptive adversary.

In short, the contribution of our algorithm is that it is the first FPL algorithm in the multi-armed bandit model that is $O(\sqrt{T \ln T})$ competitive against an adaptive adversary. In addition, the terms in the regret bound are somewhat lower than the previous expected regrets of the algorithms by Auer [6] and Dani and Hayes [4].

In the following section we briefly recapitulate the multi-armed bandit model. In Section 3 we describe the FPL framework and the online geometric optimization problem in more detail. We also review related work. Section 4, then, introduces the Adaptive Bandit FPL algorithm. Its regret bound is analyzed in Section 5. Some proofs are deferred to Appendix A. Finally, Section 6 concludes this paper.

2 Multi-Armed Bandit Model

Multi-armed bandit [7] is a formulation for the exploration-exploitation dilemma: Whether a gambler should stick with the lever of a slot machine that has paid off best so far or should he go looking for a a better one among those that were not tried yet. The gambler's trade-off is in wasting effort on exploring unprofitable alternatives versus risking to miss a superior lever. This online optimization problem has been under scrutiny for over fifty years [7, 2] and is finally starting to be well understood [8, 5, 4, 9]. The early research on the multi-armed bandit model concentrated on studying stochastic and probabilistic slot machines, leaving the gambler to learn a good play strategy (for a review and references see [5]). The recent research, on the other hand, has deemed such machines inadequate to model all situations of interest [5], turning instead to study adversarial slot machines and randomized algorithms to play against them.

There are four different variations of this iterated zero-sum two-player game between the gambler and the k-armed slot machine (see Table 1). The gambler may be told the payoff of each lever in each round or, more realistically, he may know only the result of the lever that he chose to play. The former model is the *full information* game, while the latter one is known as the *bandit setting*. The slot machine's payoff may be insensitive to the gambler's actions (i.e., it can as well be chosen in advance), in which case we have an *oblivious (non-adaptive) adversary*, or it may depend on the gambler's earlier choices and then the slot machine is an *adaptive adversary*. In this paper we are concerned with the most demanding setting out of these alternatives — ABS.

Gambling at a multi-armed bandit is a zero-sum game because the gambler's losses equal the winnings of the casino. If the duration of the game is not restricted, the gambler's total loss clearly has no bounds. Therefore, his performance is usually measured as the *regret*; the difference between his actual loss on the T rounds played and the single best lever in hindsight. With respect to this performance measure the gambler can do almost as well as the best static decision. To formulate, let $\ell_t \in \{1, \ldots, k\}$ be the lever chosen by the gambler at round (or turn) $t \in \{1, \ldots, T\}$. Simultaneously with the gambler's decision the adversary assigns costs $\boldsymbol{c}_t = \langle c_{1,t}, \ldots, c_{k,t} \rangle$ to the levers. Of course, the costs need to be restricted somehow, because otherwise it is possible to suffer an unbounded loss on one lever pull. Let us assume that the costs are normalized to the range $[0, 1]$.

After both parties are ready, the gambler's choice is communicated to the adversary and (in the bandit setting) the gambler learns the cost $c_{\ell_t,t}$ of the lever he played. Eventually, when all T rounds have been played, the gambler has suffered a total *loss* $\sum_{t=1}^{T} c_{\ell_t,t}$. The best lever $j \in \{1, \ldots, k\}$ in hindsight is the one that minimizes $\sum_{t=1}^{T} c_{j,t}$. Hence, the performance measure regret is

$$\sum_{t=1}^{T} c_{\ell_t,t} - \min_{j \in \{1,\ldots,k\}} \sum_{t=1}^{T} c_{j,t}.$$

In the OFI case the optimal expected regret bound $\Theta(\sqrt{T \ln k})$ can be obtained by Freund and Schapire's [8] HEDGE algorithm, a variant of the WEIGHTED MAJORITY algorithm [10]. The same bound directly also holds in the AFI case [11, 12]. In the bandit setting the situation is different; an adaptive adversary is strictly more powerful than an oblivious one [11].

For the oblivious bandit setting Auer et al. [5] put forward the EXP3 algorithm which is similar to the HEDGE algorithm. Dani and Hayes [4] note that while EXP3 guarantees that

$$\mathbf{E}(\text{regret vs. a lever } i) = O(\sqrt{kT \ln k}),$$

for a certain adaptive adversary it unfortunately holds that

$$\mathbf{E}(\text{regret vs. the best lever}) = \Omega(T^{2/3}).$$

Auer [6] has demonstrated how to modify the EXP3 algorithm to obtain a regret of $O(\sqrt{kT \ln T})$ against an adaptive adversary. Dani and Hayes also give an algorithm called ACCOUNTS that has a regret of $O(\sqrt{kT \ln T})$ against an adaptive adversary, although, with high constant terms.

3 Online Geometric Optimization and Following the Perturbed Leader

In the generalization of the problem — online geometric optimization [1] — an algorithm has to choose, in each round $t \in \{1, \ldots, T\}$, a vector $\boldsymbol{x}_t$ from a fixed decision set $S_D \subset \mathbb{R}^d$. The adversary simultaneously chooses a cost vector $\boldsymbol{c}_t$ from a fixed cost set $S_C \subset \mathbb{R}^d$. The loss of the algorithm at round t is the inner product $\boldsymbol{x}_t \cdot \boldsymbol{c}_t$ and its goal is to minimize regret

$$\sum_{t=1}^{T} \boldsymbol{x}_t \cdot \boldsymbol{c}_t - \min_{\boldsymbol{x}_* \in S_D} \sum_{t=1}^{T} \boldsymbol{x}_* \cdot \boldsymbol{c}_t. \tag{1}$$

In order to formulate the multi-armed bandit model in the same inner product loss setting, we can present the gambler's choice of a lever as an indicator vector; one in which all elements are zero except the one corresponding to the chosen lever.

There are a number of online problems that fit the geometric setting; e.g., *online routing* [1], *online set cover* [13], and *online traveling salesman problem* [13]. The usefulness of this model stems from the fact that there are algorithms with regret bounds depending on the dimension of the vector space of the decisions. In the above-mentioned problems the costs between different decisions can be linearly related and, hence, the dimension of the decision space is lower than the number of decisions.

Kalai and Vempala [1] showed that online geometric decision problems have efficient solutions, given an oracle for the offline version of the problem. They reintroduced the FPL algorithm, which selects a decision based on past cost vectors. Originally this method was proposed by Hannan [2] already fifty years ago. In short, the algorithm selects the best decision for the past cost vectors which are perturbed by a random vector. Zinkevich [14] has proposed another approach for solving the online geometric setting, but FPL is more relevant when the space of the decision vectors is not convex.

The regret bound of FPL in the full information setting depends on the structure of the decision vector space S_D and the cost vector space S_C. The following restrictions apply. There is a bound on

- the diameter of the decision set: $\|\boldsymbol{x}_t - \boldsymbol{x}'_t\|_1 \leq D$ for all $\boldsymbol{x}_t, \boldsymbol{x}'_t \in S_D$,
- the cost set: $\|\boldsymbol{c}_t\|_1 \leq A$ for any $\boldsymbol{c}_t \in S_C$, and
- the maximum cost: $|\boldsymbol{c}_t \cdot \boldsymbol{x}_t| \leq R$.

Let $\text{opt}(\boldsymbol{c})$ be the optimal decision for a cost vector $\boldsymbol{c}$. We abbreviate the sum of cost vectors up to the time step t by $\boldsymbol{c}_{1:t} = \sum_{i=1}^{t} \boldsymbol{c}_i$. Hence, if T is the total number of rounds, then $\text{opt}(\boldsymbol{c}_{1:T})$ is the best static decision and $\text{opt-cost}(\boldsymbol{c}_{1:T}) =$

Table 2. Multiplicative FPL* [1].

On turn t:

1. Choose a random perturbation vector $\boldsymbol{\mu}_t \propto \exp(-\epsilon_t \|x\|_1)$ drawn from the exponential distribution, in which $\epsilon_t = O(1/\sqrt{t})$.
2. Pick decision $\mathrm{opt}(\boldsymbol{c}_{1:t-1} + \boldsymbol{\mu}_t)$.

$\mathrm{opt}(\boldsymbol{c}_{1:T}) \cdot \boldsymbol{c}_{1:T}$ is its cost. With these restrictions the FPL algorithm (see Table 2) achieves a cost of

$$(1 + 2\epsilon A)\,\text{opt-cost}(\boldsymbol{c}_{1:T}) + \frac{D \ln d}{\epsilon}$$

for a parameter $\epsilon < 1$ [1]. Thus, the regret of the FPL algorithm is $O(\sqrt{DRAT})$ when a suitable ϵ is plugged into the above formula.

The regret bound of FPL does not depend on the number of different decisions, whereas the regret bound of WEIGHTED MAJORITY depends on the number of experts. If the decisions are independent — i.e., experts of WEIGHTED MAJORITY — the difference between FPL and WEIGHTED MAJORITY is of implementation. Although the regret bound of FPL [1] is a constant factor $\sqrt{2}$ worse than that of HEDGE, Kalai has observed that we can choose the perturbation vector so that the decisions of FPL and HEDGE are identical [4]. Unfortunately, to the best of our knowledge there is no proof that this method works in the more general online geometric optimization.

Hutter and Poland have shown that many of the results obtained in the expert setting can be elegantly proved also in the FPL setting, albeit with worse constants [12].

3.1 Bandit FPL

FPL solves a full information problem (OFI or AFI). For the bandit setting, where only the cost of one decision per time step is revealed, there are several variations of FPL [15, 16]. Bandit FPL algorithms typically choose at each time step between exploration and exploitation, like algorithms based on WEIGHTED MAJORITY. The exploration rate $\gamma_t \in [0, 1]$ determines which option to choose at step t. With probability γ_t one explores a representative sample from the decisions and with probability $1 - \gamma_t$ one exploits by choosing the decision that appears best in light of the (estimated) past costs.

Awerbuch and Kleinberg [15] considered the oblivious bandit setting and their algorithm has a regret of the order $O(T^{2/3})$. They also give an algorithm with the same regret bound designed specifically for the shortest online routing problem in the ABS case. McMahan and Blum [16] obtained a regret of $O(T^{3/4}\sqrt{\ln T})$ for the bandit FPL in the general ABS case. Dani and Hayes [11] improved the regret of McMahan and Blum's algorithm in the ABS case to $O(T^{2/3})$. Note that the above bounds omit dependences on other factors than T, which are typically quite complicated.

These algorithms for the geometric bandit setting, in general, estimate the cost vector only during exploration turns and ignore the costs on exploitation turns. If we want to get below the $O(T^{2/3})$ regret, it is necessary to use also the information obtained during exploitation turns [11]. Currently no general algorithm for the online geometric bandit setting exists with $O(\sqrt{T})$ regret, but there are customized algorithms for particular problems, like for the online shortest routing, which achieve a regret close to $O(\sqrt{T})$ [17].

In [3] we showed that a FPL variant — Bandit FPL (B-FPL) — can attain $O(\sqrt{T})$ expected regret in the OBS setting when restricted to the expert framework. In this paper we will turn to the ABS case, this time restricting our considerations to the multi-armed bandit model. We will prove that a FPL algorithm — Adaptive Bandit FPL, AB-FPL — attains an expected regret $O(\sqrt{kT \ln T})$. Thus, we match the asymptotic bound $O(\sqrt{k\,T \ln T})$ of the ACCOUNTS algorithm of Dani and Hayes [4] and of the algorithm given by Auer [6], but our hidden terms are somewhat lower. Although AB-FPL is currently restricted to multi-armed bandits, we hope that the idea in AB-FPL is general enough to be implemented in a B-FPL algorithm that has $O(\sqrt{T})$ regret in the online geometric bandit setting against a non-adaptive adversary.

4 Adaptive Bandit FPL

In this and the following section we show how FPL can be applied against an adaptive adversary in the multi-armed bandit setting. An argument by Dani and Hayes [4] shows that the B-FPL [3] algorithm cannot attain a regret of the order $O(\sqrt{T})$ in the ABS case and that its worst-case regret is $\Omega(T^{2/3})$. We now demonstrate that it is possible to modify the B-FPL algorithm so that it is competitive against an adaptive adversary. We call the resulting algorithm Adaptive Bandit FPL (AB-FPL). The modified algorithm AB-FPL is given in Table 3.

In short, AB-FPL is the same as B-FPL, except that instead of using an unbiased estimate $\hat{\boldsymbol{c}}_{1:t}$ of the real vector $\boldsymbol{c}_{1:t}$ it uses $\hat{\boldsymbol{c}}'_{1:t} = \hat{\boldsymbol{c}}_{1:t} - \lambda\,\boldsymbol{\sigma}_t \sqrt{\ln t}$, where λ is a small constant and $\boldsymbol{\sigma}_t$ is a vector that consists of upper bounds to conditional standard deviations of items in $\hat{\boldsymbol{c}}_{1:t}$. Intuitively, this guarantees that we almost never estimate the cost of a lever too high and, thus, it is difficult to hide the actual cost under the variance of the estimate of the cost. Fortunately this method does not produce a too great perturbation to the estimated costs, making it possible to have a reasonable regret bound as we will show below. Similar work has been done by Auer [6] with WEIGHTED MAJORITY style algorithms. Let us first state the regret bound explicitly.

Theorem 1. *The expected regret of* AB-FPL *against the best lever in hindsight is at most $O(\sqrt{T \ln T})$ against an adaptive adversary. Asymptotically the most significant term is $5\sqrt{k\,T \ln T}$.*

Table 3. Adaptive Bandit FPL (AB-FPL).

- Let there be k levers $\{1, \dots k\}$ and let $\boldsymbol{p}_{i,t}$ be the probability of choosing lever i at turn t. We will describe how to compute these probabilities in Section 5.3.
- Let $c_{i,t}$ be the cost of lever i at turn t and let vector $\boldsymbol{c}_t$ consists of the costs for all levers.
- Moreover, let $\boldsymbol{c}_{1:t}$ be the sum of cost vectors up to turn t.
- Let $\mathrm{opt}(\boldsymbol{c})$ be the indicator vector of some minimal value in $\boldsymbol{c}$; i.e., $\mathrm{opt}(\boldsymbol{c}) = i$ if i is a minimal value in $\boldsymbol{c}$.
- Define two parameters:
  1. ϵ_t is intuitively the width of the perturbation vector $\boldsymbol{\mu}_t$ used at an exploitation turn. We use $\epsilon_t = \epsilon_T = \sqrt{\ln T}/(3\sqrt{k\,T})$.
  2. $\gamma_t = \min(1, k\epsilon_t)$ is the probability of sampling a lever uniformly at random.
- Maintain two variables:
  1. Vector $\hat{\boldsymbol{c}}_{1:t}$ that is an estimate of $\boldsymbol{c}_{1:t}$. Estimated in the usual way as in EXP3 and B-FPL
  $$\hat{\boldsymbol{c}}_{i,t} = \begin{cases} \boldsymbol{c}_{i,t}/\boldsymbol{p}_{i,t} & \text{if } i \text{ was chosen on turn } t; \\ 0 & \text{otherwise.} \end{cases}$$
  2. Vector $\boldsymbol{\sigma}_t^2$ that contains upper bounds of the conditional variances of random variables $\hat{\boldsymbol{c}}_{i,1:t}$. It is defined as
  $$\boldsymbol{\sigma}_{i,t}^2 = \sum_{\tau=1}^{t} \frac{1}{\boldsymbol{p}_{i,\tau}} \geq \sum_{\tau=1}^{t} \frac{\boldsymbol{c}_{i,\tau}^2}{\boldsymbol{p}_{i,\tau}} - \boldsymbol{c}_{i,\tau}^2 = \sum_{\tau=1}^{t} \mathbf{Var}\left(\hat{\boldsymbol{c}}_{i,\tau} \mid \tau-1, \dots, 1\right).$$
- Define $\hat{\boldsymbol{c}}'_t = \hat{\boldsymbol{c}}_t - \lambda \boldsymbol{\sigma}_t \sqrt{\ln(t+1)}$, where $\lambda = \sqrt{1 + \sqrt{2}/\sqrt{k}}$.

On turn t:

Exploration step: With probability γ_t choose a lever uniformly at random.
Exploitation step: Otherwise choose a lever given by

$$\mathrm{opt}\left(\hat{\boldsymbol{c}}'_{1:t-1} + \boldsymbol{\mu}_t\right),$$

where $\boldsymbol{\mu}_t \propto \exp(-\epsilon_t \|x\|_1)$ is a random perturbation vector in which the elements are drawn from the two-sided exponential distribution.

---

5 Proof for the Regret Bound

We first outline the proof of Theorem 1. First we observe that the cumulated cost of AB-FPL during the *exploitation steps* is upper bounded by

$$\mathbf{E}\left(\sum_{t=1}^{T} \mathrm{opt}\left(\hat{\boldsymbol{c}}'_{1:t-1} + \boldsymbol{\mu}_t\right) \cdot \boldsymbol{c}_t\right).$$

We further bound this with

$$\text{opt-cost}(\boldsymbol{c}_{1:T}) + O(\sqrt{T \ln T}).$$

The expected cumulative cost during the *exploration steps* in AB-FPL is upper bounded by $\sum_{t=1}^{T} \gamma_t = O(\sqrt{T \ln T})$, and hence for all steps we get a regret bound that is of the order $O(\sqrt{T \ln T})$.

We go through the following steps:

1. FPL run on *unbiased* estimates $\hat{\boldsymbol{c}}_t$ of AB-FPL has the same expected regret as AB-FPL on actual costs. Several authors have made this observation [12, 11]. It can be proved using a simple calculation of expected value.

$$\mathbf{E}_{\text{act}} \left(\sum_{t=1}^{T} \text{opt}\big(\hat{\boldsymbol{c}}'_{1:t-1} + \boldsymbol{\mu}_t\big) \cdot \boldsymbol{c}_t \right) = \mathbf{E}_{\text{act}} \mathbf{E}_{\text{sim}} \left(\sum_{t=1}^{T} \text{opt}\big(\hat{\boldsymbol{c}}'_{1:t-1} + \boldsymbol{r}_t\big) \cdot \hat{\boldsymbol{c}}_t \right)$$

From now on we will refer by simulated FPL to a run of the FPL algorithm with cost vectors $\hat{\boldsymbol{c}}'_t$ as the input. These are available after an actual run of AB-FPL. Vectors $\boldsymbol{r}_t$ are perturbations chosen by a simulated run of FPL on vectors $\hat{\boldsymbol{c}}'_t$; thus, they are independently chosen from the same distribution as perturbations $\boldsymbol{\mu}_t$. By $\mathbf{E}_{\text{act}}$ we denote expectation with regards to the actual choices made by AB-FPL and the adaptive adversary. $\mathbf{E}_{\text{sim}}$, on the other hand, refers to expectation with regards to choices made by a simulated FPL using cost vectors $\hat{\boldsymbol{c}}'_t$. Hence, $\mathbf{E}_{\text{act}}$ is taken over the randomness of the joint distribution on vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_t$ and $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_t$, while $\mathbf{E}_{\text{sim}}$ is taken over the randomness of vectors $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_t$ given random choices that happen under $\mathbf{E}_{\text{act}}$. Plain $\mathbf{E}$ always refers to all randomness within the scope of the expectation operator.
When compared to the right-hand side of the above equation, in AB-FPL we have an additional term that results from it having extra vectors $\boldsymbol{\sigma}_t \sqrt{\ln t} - \boldsymbol{\sigma}_{t-1} \sqrt{\ln(t-1)}$ in the estimate $\hat{\boldsymbol{c}}'_{1:t}$ which is used. In equations we get

$$\begin{aligned} &\mathbf{E} \left(\sum_{t=1}^{T} \text{opt}\big(\hat{\boldsymbol{c}}'_{1:t-1} + \boldsymbol{\mu}_t\big) \cdot \boldsymbol{c}_t \right) \\ &= \underbrace{\mathbf{E}_{\text{act}} \mathbf{E}_{\text{sim}} \left(\sum_{t=1}^{T} \text{opt}\big(\hat{\boldsymbol{c}}'_{1:t-1} + \boldsymbol{r}_t\big) \cdot \hat{\boldsymbol{c}}'_t \right)}_{A} \\ &\quad + \lambda \underbrace{\mathbf{E}_{\text{act}} \mathbf{E}_{\text{sim}} \left(\sum_{t=1}^{T} \text{opt}\big(\hat{\boldsymbol{c}}'_{1:t-1} + \boldsymbol{r}_t\big) \cdot \Big(\boldsymbol{\sigma}_t \sqrt{\ln t} - \boldsymbol{\sigma}_{t-1} \sqrt{\ln\,(t-1)} \Big) \right)}_{B}. \end{aligned}$$

2. In Section 5.1 we show that $B \leq 2\sqrt{k\,T \ln T}$.
3. In Appendix A we bound the term A with

$$\mathbf{E}_{\text{act}} \big(\text{opt-cost}\big(\hat{\boldsymbol{c}}'_{1:T}\big)\big) + e\,(e-1) \left(k \sum_{t=1}^{T} \epsilon_t + \frac{2\lambda\sqrt{T \ln T}}{\sqrt{k}} \right) + \frac{2(1 + \ln k)}{\epsilon_T}.$$

This bound is quite standard and uses techniques introduced earlier [1, 3].

4. Finally, by definition it holds that

$$\text{opt-cost}\left(\hat{\boldsymbol{c}}'_{1:t}\right) \leq \text{opt}(\boldsymbol{c}_{1:t}) \cdot \hat{\boldsymbol{c}}'_{1:t}.$$

Using a martingale concentration inequality we can show that this is, with a high probability, less than the cost of the best static decision opt-cost$(\boldsymbol{c}_{1:t})$. We do this in Section 5.2.

These steps bound the cost on the exploitation steps, as argued above. In Table 3 we set the parameters $\gamma_t/k = \epsilon_t = \epsilon_T = \sqrt{\ln T}/(3\sqrt{k\,T})$. Counting the cost during the sampling turns we obtain a final regret bound where asymptotically the most significant term is

$$5\sqrt{k\,T\ln T}.$$

The remaining terms are bounded by

$$2\sqrt{2}\frac{\sqrt{T\ln T}}{\sqrt{k}} + \frac{6(1+\ln k)\sqrt{k\,T}}{\sqrt{\ln T}}.$$

Using the concentration inequality yields an additional term, but it is of the order $o(\sqrt{T})$ with a small constant factor.

5.1 Bound on the Term B

We now proceed to bound the term

$$B = \mathbf{E}_{\text{act}}\mathbf{E}_{\text{sim}}\left(\sum_{t=1}^{T} \text{opt}\left(\hat{\boldsymbol{c}}'_{1:t-1} + \boldsymbol{r}_t\right) \cdot \left(\boldsymbol{\sigma}_t\sqrt{\ln t} - \boldsymbol{\sigma}_{t-1}\sqrt{\ln(t-1)}\right)\right).$$

First, because $\boldsymbol{\sigma}_{i,t}$ grows faster than $\sqrt{\ln t}$, the inequality

$$\boldsymbol{\sigma}_{i,t}\sqrt{\ln t} - \boldsymbol{\sigma}_{i,t-1}\sqrt{\ln(t-1)} \leq (\boldsymbol{\sigma}_{i,t} - \boldsymbol{\sigma}_{i,t-1})\left(\sqrt{\ln t} + \sqrt{\ln(t-1)}\right)$$

holds. Recall that $\sqrt{x}$ is concave and hence, by linearization, $\sqrt{x+h} \leq \sqrt{x} + h/(2\sqrt{x})$ for any $x, h > 0$. Direct application of this inequality shows that for each $\boldsymbol{\sigma}_{i,t}$

$$\boldsymbol{\sigma}_{i,t} - \boldsymbol{\sigma}_{i,t-1} \leq \frac{1}{2\boldsymbol{p}_{i,t}\,\boldsymbol{\sigma}_{i,t-1}}.$$

Hence, the term B is upper bounded by $\mathbf{E}_{\text{act}}\left(\sqrt{\ln T}\sum_{t=1}^{T}\sum_{i=1}^{k} 1/\boldsymbol{\sigma}_{i,t}\right)$ when the expectation $\mathbf{E}_{\text{sim}}$ is calculated. Since the function $1/\sqrt{x}$ is both convex and decreasing, it is not difficult to see that to maximize individually each $\sum_{i=1}^{k} 1/\boldsymbol{\sigma}_{i,t}$ all probabilities $\boldsymbol{p}_{i,\tau}$ in $\boldsymbol{\sigma}_{i,t} = \sqrt{\sum_{\tau=1}^{t} 1/\boldsymbol{p}_{i,\tau}}$ should be set to $1/k$. We can formally prove this using an involved manipulation of Lagrange multipliers to obtain the inequality:

$$\mathbf{E}_{\mathrm{act}}\left(\sum_{t=1}^{T}\sum_{i=1}^{k} 1/\boldsymbol{\sigma}_{i,t}\right) \le \sum_{t=1}^{T}\sum_{i=1}^{k}\frac{1}{\sqrt{k\,t}} \le 2\sqrt{k\,T}. \tag{2}$$

Thus, we obtain the desired bound

$$B \le 2\sqrt{k\,T\ln T}.$$

5.2 Bounding the Estimated Cost

We claim that with a high enough probability it holds that

$$\mathrm{opt}(\boldsymbol{c}_{1:T}) \cdot \hat{\boldsymbol{c}}'_{1:T} \le \text{opt-cost}(\boldsymbol{c}_{1:T}) \,.$$

Intuitively, $\hat{\boldsymbol{c}}_{1:T}$ is a random variable that is concentrated to a certain degree around its expected value and we take advantage of this in lower bounding the vector $\hat{\boldsymbol{c}}_t$ by $\hat{\boldsymbol{c}}'_t$ with high probability. Formally we use the following McDiarmid's [18] martingale concentration inequality. Note that this version of the inequality allows us to use the conditional variances.

Theorem 2 (Theorem 3.15, p. 224, in [18]). *Let $X_1, \ldots, X_T$ be a martingale difference sequence and V the sum of conditional variances*

$$V = \sum_{t=1}^{T} \mathbf{Var}\left(X_t \mid X_1, \ldots, X_{t-1}\right),$$

then for every $x, v \ge 0$

$$\mathbf{P}\left(\sum_t X_t \ge x \;\; \text{and} \;\; V \le v\right) \le \exp\left(\frac{-x^2}{2v + 2xu/3}\right),$$

where u is a uniform upper bound for X_t.

We apply this inequality and set

- $X_{i,t} = \hat{\boldsymbol{c}}_{i,t} - \boldsymbol{c}_{i,t}$,
- $x = \lambda\,\boldsymbol{\sigma}_{i,T}\ln T$,
- $u = 1/\gamma_T$, and
- $v = \boldsymbol{\sigma}^2_{i,T}$.

Note that $V \leq \boldsymbol{\sigma}_{i,T}^2$ holds always because $\boldsymbol{\sigma}_{i,T}^2$ is defined to be a upper bound of the variance of $\sum_{t=1}^T X_{i,t}$. Now

$$\begin{aligned}
\mathbf{P}\left(\sum_{t=1}^{T} \hat{\boldsymbol{c}}_{i,t} - \boldsymbol{c}_{i,t} \geq \lambda\, \boldsymbol{\sigma}_{i,T}\sqrt{\ln T}\right) &\leq \exp\left(\frac{-\lambda^2\, \boldsymbol{\sigma}_{i,T}^2 \ln T}{2\boldsymbol{\sigma}_{i,T}^2 + 2\lambda\, \boldsymbol{\sigma}_{i,T}\sqrt{\ln T}/(3\gamma_T)}\right) \\
&\leq \exp\left(\frac{-\lambda^2\, \ln T}{2 + 2\lambda\sqrt{\ln T}/(3k\epsilon_T\sqrt{T})}\right) && (3) \\
&\leq \exp\left(\frac{-\lambda^2\, \ln T}{2 + 2\lambda/\sqrt{k}}\right) && (4) \\
&\leq \exp\left(\frac{-\lambda^2\, \ln \sqrt{T}}{1 + \sqrt{2}/\sqrt{k}}\right) \\
&= \frac{1}{\sqrt{T}}. && (5)
\end{aligned}$$

In Inequality 3 we use the definition $\gamma_T = k\epsilon_T$ and the fact that $\boldsymbol{\sigma}_{i,T} \geq \sqrt{T}$. In Inequality 4 we write out $\epsilon_T = \sqrt{\ln T}/(3\sqrt{kT})$. Inequality 5 follows from $\lambda^2 = 1 + \sqrt{2}/\sqrt{k}$. When this bounding is done for each of the k levers separately, the bad case happens with probability at most $k/\sqrt{T}$. Even the bad case is of the order $o(T)$ with extremely high probability, because $\boldsymbol{\sigma}_{i,T} \leq T^{2/3}$, and we can again use the same concentration inequality. Hence,

$$\mathbf{E}\big(\mathrm{opt}(\boldsymbol{c}_{1:T}) \cdot \hat{\boldsymbol{c}}'_{1:T} - \text{opt-cost}(\boldsymbol{c}_{1:T})\big) \leq o(\sqrt{T}).$$

We do not include this term in our regret bound, because asymptotically it is not significant.

5.3 Computing the Probabilities in AB-FPL

We need to obtain the probability $\boldsymbol{p}_{i,t}$ of selecting lever i at turn t. On an exploitation turn the probability of selecting lever i is the probability that $\hat{\boldsymbol{c}}'_{i,1:t-1}$ plus perturbation is smaller than the corresponding value for the other levers. Each of these values is distributed according to a two-sided exponential distribution with the expected value at $\hat{\boldsymbol{c}}'_{j,1:t-1}$ for lever j. The probability that the lever i has the lowest value is $\int_{-\infty}^{\infty} p_1(x)p_2(x)dx$, where $p_1(x)$ is the density of the value of lever i, and $p_2(x)$ is the probability that the values for the other levers are higher. $p_2(x)$ is a product of cumulative distributions for the exponential distribution. Hence the integral can be calculated piecewise, where the pieces are all intervals between successive expectations $\hat{\boldsymbol{c}}'_{j,1:t-1}$. In AB-FPL we must also account for the sampling that is done.

6 Conclusion and Discussion

We have demonstrated that the FPL algorithm can be successfully applied in the multi-armed bandit model against an adaptive adversary with a regret of order

$O(\sqrt{T \ln T})$. In the general bandit geometric setting, however, the best known regret is $O(T^{2/3})$ even against an oblivious adversary [16]. The method we have applied is plausibly general enough to be applied to a hypothetical geometric bandit FPL algorithm that has been designed against an oblivious adversary with a regret of $O(\sqrt{T})$.

Interesting future direction is to examine more closely what are the exact constant factors that are achievable in the regret bounds. Constant factors are important, because even halving the regret in an application where the cost is expensive is a significant improvement. In Vermorel and Mohri's [19] experiments — on non-adversarial stochastically generated data — simple algorithms outperformed those that were designed against an oblivious adversary. It is not surprising that theoretical worst-case regret bounds are higher for more powerful adversaries. It would, nevertheless, be interesting to know if a better regret could be simultaneously obtained in a case that is not worst-case.

Acknowledgments

We thank the anonymous reviewers for valuable comments that helped us to improve the final version of this paper. This work was supported by Academy of Finland projects "INTENTS: Intelligent Online Data Structures" and "ALEA: Approximation and Learning Algorithms." Moreover, the work of J. Kujala is financially supported by Tampere Graduate School in Information Science and Engineering (TISE) and Nokia Foundation.

References

1. Kalai, A.T., Vempala, S.: Efficient algorithms for online decision problems. Journal of Computer and System Sciences **71** (2005) 26–40
2. 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
3. Kujala, J., Elomaa, T.: On following the perturbed leader in the bandit setting. In Jain, S., Simon, H.U., Tomita, E., eds.: Algorithmic Learning Theory, 16th International Conference, ALT 2005. Volume 3734 of Lecture Notes in Artificial Intelligence., Berlin, Heidelberg, New York, Springer-Verlag (2005) 371–385
4. Dani, V., Hayes, T.P.: How to beat the adaptive multi-armed bandit. Technical report, Cornell University (2006) http://arxiv.org/cs.DS/602053.
5. Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E.: The non-stochastic multi-armed bandit problem. SIAM Journal on Computing **32** (2002) 48–77
6. Auer, P.: Using upper confidence bounds for online learning. In: Proceedings of the 41st Annual Symposium on Foundations of Computer Science, Los Alamitos, CA, IEEE Computer Society Press (2000) 270–279
7. Robbins, H.: Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society **58** (1952) 527–535
8. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences **55** (1997) 119–139

9. Cesa-Bianchi, N., Lugosi, G.: Prediction, Learning, and Games. Cambridge University Press, Cambridge (2006)
10. Littlestone, N., Warmuth, M.K.: The weighted majority algorithm. Information and Computation **108** (1994) 212–261
11. Dani, V., Hayes, T.P.: Robbing the bandit: Less regret in online geometric optimization against an adaptive adversary. In: Proceeding of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms, New York, NY, ACM Press (2006) 937–943
12. Hutter, M., Poland, J.: Adaptive online prediction by following the perturbed leader. Journal of Machine Learning Research **6** (2005) 639–660
13. Kakade, S., Kalai, A.T., Ligett, K.: Approximation algorithms going online. Technical Report CMU-CS-07-102, Carnegie Mellon University (2007) reports-archive.adm.cs.cmu.edu/anon/2007/CMU-CS-07-102.pdf.
14. Zinkevich, M.: Online convex programming and generalized infinitesimal gradient ascent. In Fawcett, T., Mishra, N., eds.: Proceeding of the 20th International Conference on Machine Learning, Menlo Park, CA, AAAI Press (2003) 928–936
15. Awerbuch, B., Kleinberg, R.: Near-optimal adaptive routing: Shortest paths and geometric generalizations. In: Proceeding of the 36th Annual ACM Symposium on Theory of Computing, New York, NY, ACM Press (2004) 45–53
16. McMahan, H.B., Blum, A.: Geometric optimization in the bandit setting against an adaptive adversary. In Shawe-Taylor, J., Singer, Y., eds.: Learning Theory, 17th Annual Conference on Learning Theory, COLT 2004. Volume 3120 of Lecture Notes in Computer Science., Berlin, Heidelberg, New York, Springer (2004) 109–123
17. György, A., Linder, T., Lugosi, G.: Tracking the best of many experts. In Auer, P., Meir, R., eds.: Learning Theory, 18th Annual Conference on Learning Theory, COLT 2005. Volume 3559 of Lecture Notes in Computer Science., Berlin, Heidelberg, New York, Springer-Verlag (2005) 204–216
18. McDiarmid, C.: Concentration. In Habib, M., McDiarmid, C., Ramirez-Alfonsin, J., Reed, B., eds.: Probabilistic Methods for Algorithmic Discrete Mathematics. Springer, Berlin, Heidelberg, New York (1998) 195–248
19. Vermorel, J., Mohri, M.: Multi-armed bandit algorithms and empirical evaluation. In Gama, J., Camacho, R., Brazdil, P., Jorge, A., Torgo, L., eds.: Machine Learning: ECML 2005, 16th European Conference on Machine Learning. Volume 3720 of Lecture Notes in Computer Science., Berlin, Heidelberg, New York, Springer (2005) 437–448

A Bound on the Term A

In this appendix we bound the term

$$A = \mathbf{E}_{\mathrm{act}}\mathbf{E}_{\mathrm{sim}}\left(\sum_{t=1}^{T}\mathrm{opt}\big(\hat{\boldsymbol{c}}'_{1:t-1} + \boldsymbol{r}_t\big)\cdot\hat{\boldsymbol{c}}'_t\right).$$

Recall that $\mathbf{E}_{\mathrm{sim}}$ refers to the expectation on randomness originating from vectors $\boldsymbol{r}_t$, i.e., a simulated run of FPL on vectors $\hat{\boldsymbol{c}}'_{1:t}$, and the adversary for the *simulated run* is an oblivious one. The bound for the simulated run of FPL is similar to what Kalai and Vempala [1] have derived against an oblivious adversary. The difference is that in the simulated run of FPL on cost vectors $\hat{\boldsymbol{c}}'_t$ the worst-case cost on a single step is too large, although under expectation the worst-case cost is small enough. The following step is different from their proof.

Lemma 1.

$$\mathbf{E}\left(\sum_{t=1}^{T}\mathrm{opt}\big(\hat{\boldsymbol{c}}'_{1:t-1}+\boldsymbol{r}_t\big)\cdot\hat{\boldsymbol{c}}'_t\right)\le\mathbf{E}\left(\sum_{t=1}^{T}\mathrm{opt}\big(\hat{\boldsymbol{c}}'_{1:t}+\boldsymbol{r}_t\big)\cdot\hat{\boldsymbol{c}}'_t\right)$$
$$+\,e\,(e-1)\left(k\sum_{t=1}^{T}\epsilon_t+\frac{2\lambda\sqrt{T\ln T}}{\sqrt{k}}\right).$$

We also need the following lemma to complete the desired bound on A.

Lemma 2 (Lemmas 1 and 3 in [1]).

$$\mathbf{E}_{\mathrm{sim}}\left(\sum_{t=1}^{T}\mathrm{opt}\big(\hat{\boldsymbol{c}}'_{1:t}+\boldsymbol{r}_t\big)\cdot\hat{\boldsymbol{c}}'_t\right)\le\text{opt-cost}\big(\hat{\boldsymbol{c}}'_{1:T}\big)+\frac{2\,(1+\ln k)}{\epsilon_T}.$$

Applying both of these lemmas gives us the final bound on A:

$$\mathbf{E}_{\mathrm{act}}\big(\text{opt-cost}\big(\hat{\boldsymbol{c}}'_{1:T}\big)\big)+e\,(e-1)\left(k\sum_{t=1}^{T}\epsilon_t+\frac{2\lambda\sqrt{T\ln T}}{\sqrt{k}}\right)+\frac{2(1+\ln k)}{\epsilon_T}.$$

Proof (of Lemma 1). As observed in [1, 3] when the perturbation vector is chosen from the exponential distribution, the probability of selecting a lever in AB-FPL changes in a well defined way when the cost vector $\hat{\boldsymbol{c}}'_{1:t}$ changes:

$$e^{-\epsilon_t\delta}\le p_{\mathrm{new}}/p_{\mathrm{old}}\le e^{\epsilon_t\delta},\tag{6}$$

where δ is the change in the 1-norm of the cost vector, p_{old} was the probability of selecting the lever before the cost vector changed, and p_{new} is the probability after the cost vector changed.

Hence, if $\boldsymbol{p}_{i,t}$ is the probability of selecting lever i at step t,

$$\mathbf{E}_{\mathrm{sim}}\big(\mathrm{opt}\big(\hat{\boldsymbol{c}}'_{1:t-1}+\boldsymbol{r}_t\big)\cdot\hat{\boldsymbol{c}}'_t\big)$$
$$=\sum_{i=1}^{k}\hat{\boldsymbol{c}}'_{i,t}\,\boldsymbol{p}_{i,t}\tag{7}$$
$$\le\sum_{i=1}^{k}\hat{\boldsymbol{c}}'_{i,t}\,\boldsymbol{p}_{i,t+1}\,e^{\epsilon_t\left\|\hat{\boldsymbol{c}}'_t\right\|_1}\tag{8}$$
$$\le\sum_{i=1}^{k}\hat{\boldsymbol{c}}'_{i,t}\,\boldsymbol{p}_{i,t+1}\,\left((1+(e-1)\,\epsilon_t\left\|\hat{\boldsymbol{c}}'_t\right\|_1\right)\tag{9}$$
$$\le\mathbf{E}_{\mathrm{sim}}\big(\mathrm{opt}\big(\hat{\boldsymbol{c}}'_{1:t}+\boldsymbol{r}_t\big)\cdot\hat{\boldsymbol{c}}'_t\big)+(e-1)\epsilon_t\sum_{i=1}^{k}\hat{\boldsymbol{c}}_{i,t}\,\left\|\hat{\boldsymbol{c}}'_t\right\|_1\,\boldsymbol{p}_{i,t+1}.\tag{10}$$

Equality 7 follows by definition, Inequality 8 is an application of Inequality 6, and Inequality 9 follows from the behavior of the exponent function e^x when $x < 1$. We can assume that $x < 1$ because for all $i \in \{ 1, \ldots, k \}$ and $t \in \{ 1, \ldots, T \}$ the probability $\boldsymbol{p}_{i,t} > \gamma_t / k = \epsilon_t$.

Next we bound the remaining regret term on the right in Inequality 10 under $\mathbf{E}_{\text{act}}$:

$$
\begin{aligned}
&\mathbf{E}_{\text{act}} \left(\sum_{i=1}^{k} \hat{\boldsymbol{c}}_{i,t} \, \left\| \hat{\boldsymbol{c}}'_t \right\|_1 \, \boldsymbol{p}_{i,t+1} \right) \\
&= \mathbf{E}_{\text{act}} \left(\boldsymbol{c}_{j,t} \, \frac{\boldsymbol{p}_{j,t+1}}{\boldsymbol{p}_{j,t}} \, \left\| \hat{\boldsymbol{c}}'_t \right\|_1 \right) && (11) \\
&\le e \, \mathbf{E}_{\text{act}} \left(\left\| \hat{\boldsymbol{c}}_t \right\|_1 + \lambda \left\| \boldsymbol{\sigma}_t \sqrt{\ln t} - \boldsymbol{\sigma}_{t-1} \sqrt{\ln (t-1)} \right\|_1 \right) && (12) \\
&\le e \left(k + \frac{\lambda \sqrt{\ln t}}{k \epsilon_t} \mathbf{E}_{\text{act}} \left(\sum_{i=1}^{k} \frac{1}{\boldsymbol{\sigma}_{i,t}} \right) \right). && (13)
\end{aligned}
$$

In Equality 11 we write out the sum and j refers to the lever that was chosen by AB-FPL. In Inequality 12 we use Inequality 6 and the fact that $\delta < 1$ as well as the triangle inequality $\left\| \hat{\boldsymbol{c}}'_t \right\|_1 \le \left\| \hat{\boldsymbol{c}}_t \right\|_1 + \lambda \left\| \boldsymbol{\sigma}_t \sqrt{\ln t} - \boldsymbol{\sigma}_{t-1} \sqrt{\ln (t-1)} \right\|_1$. Inequality 13 follows from calculation of the expectation $\mathbf{E}_{\text{act}}(\left\| \hat{\boldsymbol{c}}_t \right\|_1 \mid t-1, \ldots, 1)$ and from the following inequality, which was justified in Section 5.1,

$$
\left\| \boldsymbol{\sigma}_t \sqrt{\ln t} - \boldsymbol{\sigma}_{t-1} \sqrt{\ln (t-1)} \right\|_1 \le \sqrt{\ln t} \sum_{i=1}^{k} \frac{1}{\boldsymbol{p}_{i,t} \boldsymbol{\sigma}_{i,t-1}}.
$$

Putting Inequalities 10 and 13 together and summing them over all turns t yields the final regret

$$
\begin{aligned}
&e \, (e-1) \left(k \sum_{t=1}^{T} \epsilon_t + \frac{\lambda \sqrt{\ln T}}{k} \mathbf{E}_{\text{act}} \left(\sum_{t=1}^{T} \sum_{i=1}^{k} \frac{1}{\boldsymbol{\sigma}_{i,t}} \right) \right) \\
&\le e \, (e-1) \left(k \sum_{t=1}^{T} \epsilon_t + \frac{2 \lambda \sqrt{T \ln T}}{\sqrt{k}} \right).
\end{aligned}
$$

The term under $\mathbf{E}_{\text{act}}$ was bounded in Section 5.1 by $2\sqrt{k\,T}$ in Inequality 2.