TCS-TR-A-**05-7**

TCS Technical Report

FPL Analysis for Adaptive Bandits

by

J. Poland

Division of Computer Science

Report Series A

July 26, 2005

Hokkaido University
Graduate School of
Information Science and Technology

Email: jan@ist.hokudai.ac.jp

Phone: +81-011-706-7675
Fax: +81-011-706-7675

TCS-TR-A-05-4 July 26, 2005

FPL Analysis for Adaptive Bandits

Jan Poland*
Grad. School of Inf. Sci. and Tech.
Hokkaido University, Japan
jan@ist.hokudai.ac.jp
www-alg.ist.hokudai.ac.jp/~jan

Abstract

A main problem of "Follow the Perturbed Leader" strategies for online decision problems is that regret bounds are typically proven against oblivious adversary. In partial observation cases, it was not clear how to obtain performance guarantees against adaptive adversary, without worsening the bounds. We propose a conceptually simple argument to resolve this problem. Using this, a regret bound of $O(t^{\frac{2}{3}})$ for FPL in the adversarial multi-armed bandit problem is shown. This bound holds for the common FPL variant using only the observations from designated exploration rounds. Using all observations allows for the stronger bound of $O(\sqrt{t})$, matching the best bound known so far (and essentially the known lower bound) for adversarial bandits. Surprisingly, this variant does not even need explicit exploration, it is self-stabilizing. However the sampling probabilities have to be either externally provided or approximated to sufficient accuracy, using $O(t^2 \log t)$ samples in each step.

Keywords: expert advice, online algorithms, partial observations, adaptive adversary, bandit problems, FPL

1 Introduction

"Expert Advice" stands for an active research area which studies online algorithms. In each time step $t = 1, 2, 3, \ldots$ the master algorithm, henceforth called *master* for brevity, is required to commit to a decision, which results in some cost. The master has access to a class of *experts*, each of which suggests a decision at each time step. The goal is to design master algorithms such that the *cumulative regret* (which is just the cumulative excess cost) with respect to any expert is guaranteed to be small. Bounds on the regret are typically proven *in the worst case*, i.e. without any statistical

*This work was supported by JSPS 21st century COE program C01.

assumption on the process assigning the experts' costs. In particular, this might be an *adaptive adversary* which aims at maximizing the master's regret and also knows the master's internal algorithm. This implies that (unless the decision space is continuous and the cost function is convex) the master must *randomize* in order to protect against this danger.

In the recent past, a growing number of different but related online problems have been considered. Prediction of a binary sequence with expert advice has been popular since the work of Littlestone and Warmuth in the early 1990's. Freund and Schapire [FS97] removed the structural assumption on the decision space and gave a very general algorithm called Hedge which in each time step randomly picks one expert and follows its recommendation. We will refer to this setup as the *online decision problem*. Auer et al. [ACBFS95, ACBFS03] considered the first *partial observation* case, namely the bandit setup, where in each time step the master algorithm only learns its own cost, i.e. the cost of the selected expert. All these and many other papers are based on *weighted forecasting* algorithms.

A different approach, *Follow the Perturbed Leader* (FPL), was pioneered as early as 1957 by Hannan [Han57] and rediscovered recently by Kalai and Vempala [KV03]. Compared to weighted forecasters, FPL has two main advantages and one major drawback. First, it applies to the online decision problem and admits a much more elegant analysis for *adaptive learning rate* [HP05]. Even infinite expert classes do not cause much complication. (However, the leading constant of the regret bound is generically a factor of $\sqrt{2}$ worse than that for weighted forcasters.) Adaptive learning rate is necessary unless the total number of time steps to be played is known in advance.

As a second advantage, FPL also admits efficient treatment of cases where the expert class is potentially huge but has a linear structure [MB04, AK04]. We will refer to such problems as *geometric online optimization*. An example is the online shortest path problem on a graph, where the set of admissible paths = experts is exponential in the number of vertices, but the cost of each path is just the sum of the costs of the vertices.

FPL's main drawback is that its general analysis only applies against an *oblivious* adversary, that is an adversary that has to decide on *all* cost vectors before the game starts – as opposed to an adaptive one that before each time step t just needs to commit to the current cost vector. For the full information game, one can show that a regret bound against oblivious adversary implies the *same* bound against an adaptive one [HP05]. The intuition is that FPL's current decision at time t does not depend on its past decisions. Therefore, the adversary may well decide on the current cost vector before knowing FPL's previous decisions. This argument does not apply in partial observation cases, as there FPL's behavior does depend on its past decisions (because the observations do so). As a consequence, authors started to explicitly distinguish between oblivious and adaptive adversary, sometimes restricting to the former, sometimes obtaining bounds of lower quality for the latter. E.g. McMahan

and Blum [MB04] suggest a workaround, proving sublinear regret bounds against an adaptive bandit, however of worse order ($t^{\frac{3}{4}}\sqrt{\log t}$ instead of $t^{\frac{2}{3}}$, for both, geometric online optimization and online decision problem). This is not satisfactory, since in case of the bandit online decision problem for a suitable weighted forecaster, even a $O(\sqrt{t})$ bound against adaptive adversary is known [ACBFS03].

In this work, we remove FPL's major drawback. We give a simple argument (Section 2) which shows that also in case of partial observation, a bound for FPL against an oblivious adversary implies the same bound for adaptive adversary. This will allow in particular to prove a $O\big((tn\sqrt{\log n})^{\frac{2}{3}}\big)$ bound for the bandit online decision problem (Section 3). This bound is shown for the common construction where only the observations of designated exploration rounds are used. As this master algorithm is label efficient, the bound is essentially sharp. In contrast, using all informations will enable us to prove a stronger $O(\sqrt{tn\log n})$ bound (Section 4). This matches the best bound known so far for the adversarial bandit problem [ACBFS03], which is sharp within $\sqrt{\log n}$. The downside of this algorithm is that either the sampling probabilities have to be given by an oracle, or they have to be approximated with to sufficient accuracy, using $O(t^2\log t)$ samples. The case of an infinite expert class is briefly discussed in Section 5.

2 FPL: oblivious ⇒ adaptive

Assume that $c_1, c_2, \ldots \in [0,1]^n$ is a sequence of cost vectors. There are $n \geq 1$ experts. (We will give an example with infinitely many experts Section 5, but for simplicity of presentation, we restrict our main exposition to finite expert classes). That is, c_t^i is expert i's cost at time t, and the costs are bounded (w.l.o.g. in $[0,1]$). In the full observation game, at time t the master would know the past cumulative costs $c_{<t} = c_{1:t-1} = \sum_{s=1}^{t-1} c_s$ (observe that we have introduced some notation here). However, our focus are *partial observations* where this is not the case. Hence, assume that there are *estimates* $\hat{c}_t$ (to be specified later) for the cost vectors c_t. Then at time t, FPL(t) samples a perturbation vector $q_t \in [0,\infty)^n$ the components of which are independently exponentially distributed, that is, $\mathbf{P}(q_t^i \geq x) = e^{-x}$. Afterwards, the expert with the best (minimum) score $\hat{c}_{<t} - \frac{q_t}{\eta_t}$ is selected, where $\eta_t > 0$ is the *learning rate*:

$$\mathrm{FPL}(t, \hat{c}_{<t}) = \arg\min_{1\leq i\leq n}\left\{\hat{c}^i_{<t} - \tfrac{q_t^i}{\eta_t}\right\} \text{ where } q_t^i \overset{d.}{\sim} \mathrm{Exp} \text{ independently.} \qquad (1)$$

Denote the expert FPL chooses at time t by $I_t = \mathrm{FPL}(t, \hat{c}_{<t})$. Then an adaptive adversary is a function $A : [0,1]^{n\times t-1} \times \{1\ldots n\}^{t-1} \to [0,1]^n$. (We assume A to be deterministic but remark that all our results and proofs hold for randomized A without major modification.) The complete game between FPL and A is specified by $c_t = A(c_1c_2\ldots c_{t-1}, I_1I_2\ldots I_{t-1})$ and $I_t = \mathrm{FPL}(t, \hat{c}_{<t})$ for $t = 1, 2, \ldots$ The estimated cost vector $\hat{c}_t$ is revealed to FPL after time t and specified by a mechanism "outside" this game which is defined later (this is the exploration).

After the game has proceeded for a number of time steps T, we want to evaluate FPL's performance. Actually, the *expected* performance is the right quantity to address. If we are rather interested in high probability bounds on the actual performance, then they are easily obtained by observing that the difference of actual to expected performance is a martingale with bounded differences (all instantaneous costs c_t^i are in $[0,1]$). Thus, high probability bounds follow by Azuma's inequality, as we will demonstrate in Proposition 3.

How can we compute FPL's expected costs $\mathbf{E}c_{1:T}^{\text{FPL}} = \mathbf{E}\sum_{t=1}^{T} c_t^{I_t}$? The key observation is that – on the cost vectors generated by FPL and A and with the given estimated costs $\hat{c}_t$ – FPL's expected costs at time t are the same as another algorithm $\widetilde{\text{FPL}}$'s expected costs. $\widetilde{\text{FPL}}$ is defined by

$$\widetilde{\text{FPL}}(t, \hat{c}_{<t}) = \arg\min_{1\le i\le n} \left\{ \hat{c}_{<t}^i - \frac{q_*^i}{\eta_t} \right\}, \tag{2}$$

where q_* is a *single fixed* vector with independently exponentially distributed components. Since we have to be careful to take expectations w.r.t. the appropriate randomness, we explicitely refer to the randomness in the notation by writing e.g. $\mathbf{E}c_t^{\text{FPL}} = \mathbf{E}_{q_t} c_t^{\text{FPL}}$. Then the following statement trivially holds, as q_t and q_* have the same distribution.

Proposition 1 *At each time $t \le T$, we have $\mathbf{E}_{q_t} c_t^{\text{FPL}} = \mathbf{E}_{q_*} c_t^{\widetilde{\text{FPL}}}$.*

This means that in order to analyze FPL, we may now proceed by considering the expected costs of $\widetilde{\text{FPL}}$ instead. We can use the standard analysis based on the tools by Kalai and Vempala [KV03], which requires that $\widetilde{\text{FPL}}$ is executed on a sequence of cost vectors that is fixed and not known in advance. Actually, in contrast to the full observation game analysis, the bandit analysis will *never* require the true cost vectors to be revealed, but rather the estimated cost vectors. For the cost vectors generated by A in response to FPL, the prerequisite for $\widetilde{\text{FPL}}$ is satisfied – just consider $\widetilde{\text{FPL}}$ as a *virtual* or *hypothetic* algorithm which is not actually executed. Therefore it does not make any decisions or cause any response from the adversary. Just for the sake of analysis we *pretend* that it runs and evaluate the expected cost it incurs, which is the same as FPL.

Since our key argument and the way it is used in the analysis appears quite subtle at the first glance, we encourage the reader to thoroughly verify each of the subsequent formal steps.

3 The standard strategy against adversarial bandits

The first algorithm we consider, *bandit-FPL* (bFPL), is specified in Figure 1 and proceeds as follows. At time t, it decides if to perform an exploration or an exploitation

For $t = 1, 2, 3, \ldots$
set $\hat{c}_t^i = 0$ for all i
sample $r_t \in \{0, 1\}$ independently s.t. $P[r_t = 1] = \gamma_t$
If $r_t = 0$ Then set $I_t^b = \mathrm{FPL}(t, \hat{c}_{<t})$ according to (1)
If $r_t = 1$ Then sample I_t^b from $\{1 \ldots n\}$ uniformly ($I_t^b = u_t$)
play decision I_t^b and observe cost $c_t^{I_t^b}$
If $r_t = 1$ Then set $\hat{c}_t^{I_t^b} = n \cdot c_t^{I_t^b} / \gamma_t$

Figure 1: The algorithm bFPL. The exploration rate γ_t and the learning rate η_t (used by subroutine FPL) will be specified in Theorem 2.

step according to some *exploration probability* $\gamma_t \in (0, 1)$. This is realized by sampling $r_t \in \{0, 1\}$ independently from all other randomness with $P[r_t = 1] = \gamma_t$. In case of exploration ($r_t = 1$), the decision I_t^b is uniformly sampled from $\{1 \ldots n\}$, independently from all other randomness. We denote this choice by u_t. (For notational convenience, we will also refer to the irrelevant u_t's in the exploitations steps later.) In case of exploitation ($r_t = 0$), bFPL obtains its decision I_t^b by invoking FPL according to (1). After bFPL has played its decision, it observes its own costs $c_t^{I_t^b}$. Finally, only in case of exploration ($r_t = 1$), the estimated cost vector is set to something different from 0. This is the *standard way* of constructing an FPL variant against an adversarial bandit [MB04, AK04]. We will discuss how to make use of *all* observations in the next section. Here is the formal specification of the algorithm again.

$$I_t^b = \mathrm{bFPL}(t, \hat{c}_{<t}) = \begin{cases} u_t & \text{if } r_t = 1 \\ \mathrm{FPL}(t, \hat{c}_{<t}) & \text{otherwise,} \end{cases} \qquad \hat{c}_t^i = \begin{cases} \frac{n c_t^i}{\gamma_t} & \text{if } r_t = 1 \wedge i = I_t^b \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, the estimated cost vector is chosen *unbiasedly*, i.e. $\mathbf{E}_{r_t, u_t} \hat{c}_t^i = c_t^i$. This technique was introduced in [ACBFS95].

Theorem 2 *Let* $\gamma_t = \min\left\{1, t^{-\frac{1}{3}} \left(n\sqrt{\log n}\right)^{\frac{2}{3}}\right\}$ *and* $\eta_t = \frac{\gamma_t}{n^2} t^{-\frac{1}{3}} \left(n\sqrt{\log n}\right)^{\frac{2}{3}}$. *Then, for any* $T \geq (n \log n)^2$, *each expert* $i \in \{1 \ldots n\}$, *and* any *adaptive assignment of the costs* $c_1, c_2, \ldots$, *bFPL satisfies the regret bound*

$$\mathbf{E} c_{1:T}^{\mathrm{bFPL}} - c_{1:T}^i \leq 4\left(Tn\sqrt{\log n}\right)^{\frac{2}{3}}. \tag{3}$$

(For $T < (n \log n)^2$, *the regret is clearly at most* $(n \log n)^2$.*)*

Proof. All computations we use in the subsequent proof have been taken or adapted from other work. Our point is to bring them into the right order and to carefully check that in this context, against an adaptive adversary, all operations are legitimate. In particular we have to take care that all expectations are w.r.t. the appropriate randomness. Again, we make this explicit in the notation and write e.g.

$\mathbf{E}c_t^{\text{bFPL}} = \mathbf{E}_{q_t,r_{1:t},u_{1:t}} c_t^{\text{bFPL}}$. Note that according to the definition of bFPL, $\mathbf{E}c_t^{\text{bFPL}}$ in fact does *not* depend on $q_{<t}$. During the proof, we will avoid the use of unspecified expectation (without subscripts). Let's introduce abbreviation $h_{<t} = (r_{<t}, u_{<t}, q_{<t})$ for the randomization history, i.e. the tuple containing all past random variables.

Moreover, we will use *conditional expectation*. For instance, $\mathbf{E}_{q_t}[c_t^{\text{FPL}}|h_{<t}]$ denotes a random variable depending on the randomization history $h_{<t}$, where for each possible history the expectation is taken w.r.t. q_t. Since we admit adaptive assignments, we must be aware that they may depend on bFPL's past randomness. To make this explicit, we use the notation $\mathbf{E}[c_t^i|h_{<t}]$ for the adversary's decisions and rewrite our bound to show (3) as

$$\sum_{t=1}^{T} \mathbf{E}_{q_t,r_t,u_t}[c_t^{\text{bFPL}}\,|h_{<t}] - \sum_{t=1}^{T} \mathbf{E}[c_t^i|h_{<t}] \leq 4\Big(Tn\sqrt{\log n}\Big)^{\frac{2}{3}}. \tag{4}$$

In order to keep the presentation simple, we assume the adversary to be deterministic. Then for given randomization history, c_t^i is constant. The same proof (and hence the theorem) remains valid if we admit randomized adversaries.

First note that $\mathbf{E}_{q_t,r_t,u_t}[c_t^{\text{bFPL}}\,|h_{<t}] \leq \mathbf{E}_{q_t}[c_t^{\text{FPL}}|h_{<t}] + \gamma_t$ holds in each time step t by definition of bFPL and $c_t^{I_t^b} \leq 1$. Since $\gamma_t \leq t^{-\frac{1}{3}}\big(n\sqrt{\log n}\big)^{\frac{2}{3}}$, we have

$$\sum_{t=1}^{T} \gamma_t \leq \sum_{t=1}^{T} t^{-\frac{1}{3}}\big(n\sqrt{\log n}\big)^{\frac{2}{3}} \leq \tfrac{3}{2}\Big(Tn\sqrt{\log n}\Big)^{\frac{2}{3}}. \tag{5}$$

Therefore, (4) follows from

$$\sum_{t=1}^{T} \mathbf{E}_{q_t}[c_t^{\text{FPL}}|h_{<t}] - \sum_{t=1}^{T} \mathbf{E}[c_t^i|h_{<t}] \leq \tfrac{5}{2}\Big(Tn\sqrt{\log n}\Big)^{\frac{2}{3}}. \tag{6}$$

Consider this form of FPL (i.e. FPL executed in each time step) as a *virtual* algorithm: It does not run in that way on the inputs. Rather, for the sake of analysis, we pretend that it runs with the $\hat{c}_t$ obtained from bFPL and try to evaluate its (virtual) performance.

We then use Proposition 1 to bring into the play another virtual algorithm, namely $\widetilde{\text{FPL}}$. Since for given randomization history, the expected performance of FPL and $\widetilde{\text{FPL}}$ coincide, (6) is proven if we can show

$$\sum_{t=1}^{T} \mathbf{E}_{q_*}[c_t^{\widetilde{\text{FPL}}}|h_{<t}] - \sum_{t=1}^{T} \mathbf{E}[c_t^i|h_{<t}] \leq \tfrac{5}{2}\Big(Tn\sqrt{\log n}\Big)^{\frac{2}{3}}. \tag{7}$$

Next, we perform the transition from real to estimated costs. Since the estimate $\hat{c}$ was defined to be unbiased, we have $\mathbf{E}[c_t^i|h_{<t}] = \mathbf{E}_{r_t,u_t}[\hat{c}_t^i|h_{<t}]$. By the same argument, since the choice of $\widetilde{\text{FPL}}$ actually does not depend on r_t and u_t, $\mathbf{E}_{q_*}[c_t^{\widetilde{\text{FPL}}}|h_{<t}] = \mathbf{E}_{q_*,r_t,u_t}[\hat{c}_t^{\widetilde{\text{FPL}}}|h_{<t}]$ holds. Hence, (7) follows from

$$\sum_{t=1}^{T} \mathbf{E}_{q_*,r_t,u_t}[\hat{c}_t^{\widetilde{\text{FPL}}}|h_{<t}] - \sum_{t=1}^{T} \mathbf{E}_{r_t,u_t}[\hat{c}_t^i|h_{<t}] \leq \tfrac{5}{2}\Big(Tn\sqrt{\log n}\Big)^{\frac{2}{3}}. \tag{8}$$

Note that, somewhat curiously, $\widetilde{\text{FPL}}$ (like FPL) only incurs estimated costs in case of exploration, i.e. where it actually did not decide the action. We need yet another virtual algorithm, *infeasible* $\widetilde{\text{FPL}}$ or $\widetilde{\text{IFPL}}$, defined as

$$\widetilde{\text{IFPL}}(t, \hat{c}_{1:t}) = \arg\min_{1\le i\le n} \Big\{ \hat{c}^i_{1:t} - \tfrac{q^i_*}{\eta_t} \Big\}, \tag{9}$$

which uses the same perturbation q_* as $\widetilde{\text{FPL}}$. It is not feasible because at time t it makes use of the information $\hat{c}_t$, which is only available afterwards. As it is a virtual algorithm, this does not cause any problems. By [HP05, Theorem 4], which is proven by an argument very similar to (13) below, in case of exploration (i.e. $r_t = 1$) it holds that $\mathbf{E}_{q_*}[\hat{c}_t^{\widetilde{\text{FPL}}} | h_{<t}, r_t = 1] \le \mathbf{E}_{q_*}[\hat{c}_t^{\widetilde{\text{IFPL}}} | h_{<t}, r_t = 1] + \eta_t \big(\frac{n}{\gamma_t}\big)^2$. We remark that this step is valid also for independently sampled perturbations q_t. Clearly, $\mathbf{E}_{q_*}[\hat{c}_t^{\widetilde{\text{FPL}}} | h_{<t}, r_t = 0] = \mathbf{E}_{q_*}[\hat{c}_t^{\widetilde{\text{IFPL}}} | h_{<t}, r_t = 0]$ in case of exploitation ($r_t = 0$). Thus in expectation w.r.t. q_* and r_t, and for any u_t,

$$\mathbf{E}_{q_*}[\hat{c}_t^{\widetilde{\text{FPL}}} | h_{1:T}] = \mathbf{E}_{q_*, r_t}[\hat{c}_t^{\widetilde{\text{FPL}}} | h_{<t}] \le \mathbf{E}_{q_*, r_t}[\hat{c}_t^{\widetilde{\text{IFPL}}} | h_{<t}] + \tfrac{\eta_t n^2}{\gamma_t}.$$

The sum over $\frac{\eta_t n^2}{\gamma_t} \le t^{-\frac{1}{3}} \big(n\sqrt{\log n}\big)^{\frac{2}{3}}$ is bounded as in (5), and we see that (8) holds if we can show

$$\sum_{t=1}^{T} \mathbf{E}_{q_*, r_t, u_t}[\hat{c}_t^{\widetilde{\text{IFPL}}} | h_{<t}] - \sum_{t=1}^{T} \mathbf{E}_{r_t, u_t}[\hat{c}^i_t | h_{<t}] \le \Big(Tn\sqrt{\log n} \Big)^{\frac{2}{3}}. \tag{10}$$

The rest of the proof now follows as in [KV03] or [HP05]. In order to maintain self-containedness, we give it here. Actually we verify (10) for *any* choice of $r_{1:T}, u_{1:T}$, then it also holds in expectation.

In the following, we suppress the dependency on $r_{1:T}, u_{1:T}$ in the notation. Then all expectations are w.r.t. q_*. We use the following convenient notation from [KV03]: For a vector $x \in \mathbb{R}^n$, let $M(x)$ be the unit vector which has a 1 at the index $\arg\min_i\{x^i\}$ and 0's at all other places. Then the process of selecting a minimum can be written as scalar product: $\min_i\{x^i\} = M(x) \circ x$. For convenience, let $\eta_0 = \infty$ and $\tilde{c}_{1:t} = \hat{c}_{1:t} - \frac{q_*}{\eta_t}$. Then it is easy to prove by induction [KV03, HP05] that

$$\hat{c}_{1:t}^{\widetilde{\text{IFPL}}} - \sum_{t=1}^{T} M(\tilde{c}_{1:t}) \circ q_* \Big(\tfrac{1}{\eta_t} - \tfrac{1}{\eta_{t-1}} \Big) = \sum_{t=1}^{T} M(\tilde{c}_{1:t}) \circ \tilde{c}_t \le M(\tilde{c}_{1:T}) \circ \tilde{c}_{1:T}. \tag{11}$$

In order to estimate $\mathbf{E}\hat{c}_{1:t}^{\widetilde{\text{IFPL}}}$, we take expectations on both sides. Then observe $\mathbf{E}M(\tilde{c}_{1:T}) \circ \tilde{c}_{1:T} \le \mathbf{E}M(\hat{c}_{1:T}) \circ \tilde{c}_{1:T} = \min_j\{\hat{c}^j_{1:T}\} - \frac{\mathbf{E}M(\hat{c}_{1:T}) \circ q_*}{\eta_T} \le \hat{c}^i_{1:T} - \frac{1}{\eta_T}$ by definition of M. The negative term on the l.h.s. of (11) may be bounded by $\sum_{t=1}^{T} M(\tilde{c}_{1:t}) \circ q_* \Big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \Big) \le \sum_{t=1}^{T} M(-q_*) \circ q_* \Big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \Big) = \frac{\max_i\{q^i_*\}}{\eta_T} \le \frac{1+\log n}{\eta_T}$ (see [KV03] or [HP05] for the last estimate). Plugging these estimates back into (11) while observing $\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \ge 0$ and $\eta_T = T^{-\frac{2}{3}} \big(\frac{\log n}{n}\big)^{\frac{2}{3}}$ (which holds because of $T \ge (n \log n)^2$), finally shows (10) and concludes the proof of the theorem. □

Proposition 3 *(High probability bound) For each $T \geq 1$ and $0 \leq \delta \leq 1$, the actual costs of bFPL are bounded with probability at least $1-\delta$ by*

$$c_{1:T}^{\text{bFPL}} \leq \mathbf{E}c_{1:T}^{\text{bFPL}} + \sqrt{2T \log \tfrac{2}{\delta}}.$$

Proof. Again we use the explicit notation from the proof of the previous theorem. It is easy to see that the sequence of random variables $X_T = \sum_{t=1}^{T} \big(c_t^{\text{bFPL}} - \mathbf{E}_{r_t,u_t,q_t}[c_t^{\text{bFPL}} \,|h_{<t}]\big)$ is a martingale w.r.t. the filter of sigma-algebras generated by the randomization history $h_{1:t}$. Moreover, its differences are bounded by $|X_t - X_{t-1}| \leq 1$. Consequently, by Azuma's inequality, the probability that X_t exceeds some $\lambda > 0$ is bounded by $\delta = 2\exp\big(-\frac{\lambda^2}{2T}\big)$. Solve this for λ to obtain the assertion. □

4 Using all observations

The algorithm bFPL considered so far does only uses a γ-fraction of all the input. It is thus a *label efficient* decision maker [CBLS04a, CBLS04b]. One possible way to specify a label efficient problem setup is to require that the master usually does not observe anything, and it incurs maximal cost if it decides to observe something [CBLS04b]. Since just before (5), we upper bounded the costs in case of exploration by 1, it is immediate that the same analysis and hence also Theorem 2 transfer to the label efficient case. [CBLS04b, Sec. 5] prove that there is a label efficient prediction problem such that *any* forecaster incurs a regret proportional to $t^{\frac{2}{3}}$. Hence the bound in Theorem 2 is essentially sharp for bFPL.

Of course, the usual bandit setup does not require the master to make use of only a tiny fraction of all information available. For weighted forecasters, it is very easy to produce an unbiased cost estimate if each round's inputs are used. It turns out that then regret bound proportional to $\sqrt{t}$ can be obtained [ACBFS03]. Unfortunately this is different for FPL, as here the sampling probabilities are not explicitely available. In the following, we will first discuss the computationally infeasible case assuming that we know the sampling probabilities. After that, we show how to approximate them by a Monte Carlo simulation to sufficient accuracy.

Surprisingly, it is possible to work with the plain FPL algorithm from (1), without exploration. We just have to use the correct estimated cost vectors,

$$\hat{c}_t^i = \begin{cases} c_t^i / \mathbf{P}(I_t^{\text{FPL}} = i) & \text{if } i = I_t^{\text{FPL}} \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

where I_t^{FPL} was FPL's choice at time t. We assume that the values $\mathbf{P}(I_t^{\text{FPL}} = i)$ are provided by some oracle.

It is not hard to adapt the proof of Theorem 2 to analyze FPL under these conditions. As in the steps up to (8),

$$\mathbf{E}_{q_t}[c_t^{\text{FPL}}|h_{<t}] = \mathbf{E}_{q_*,q_t,r_t,u_t}[\widetilde{\hat{c}_t^{\text{FPL}}}|h_{<t}] = \mathbf{E}_{q_*,q_t,r_t,u_t}[\widetilde{\hat{c}_t^{\text{FPL}(q_*)}}(q_t)|h_{<t}].$$

The overly explicit notation $\hat{c}_t^{\widetilde{\mathrm{FPL}}(q_*)}(q_t)$ serves to remind that the cost vector estimated is obtained using q_t, while $\widetilde{\mathrm{FPL}}$'s choice incurring cost stems from q_*. It is essential that q_t and q_* are independent. Observe that in general, $\mathbf{E}_{q_*,q_t,r_t,u_t}[\hat{c}_t^{\widetilde{\mathrm{FPL}}(q_*)}(q_t)|h_{<t}] \lneqq \mathbf{E}_{q_t,r_t,u_t}[\hat{c}_t^{\widetilde{\mathrm{FPL}}(q_t)}(q_t)|h_{<t}]$: the latter quantity, which is the actual estimated cost of $\widetilde{\mathrm{FPL}}$'s choice, is biased and too large.

Abbreviate $p^i = \mathbf{P}(I_t^{\widetilde{\mathrm{FPL}}} = i)$ and $\pi^i = \mathbf{P}(I_t^{\widetilde{\mathrm{IFPL}}} = i)$. Denote the exponential distribution by μ and integration with respect to $q^1 \dots q^n$ without the ith coordinate by $\int \dots d\mu(q^{\neq i})$. Moreover, for $x \in \mathbb{R}$, let $x^+ = \max\{x, 0\}$. Then, similarly to the proof of [HP05, Theorem 4],

$$\begin{aligned} p_i &= \int \int_{\max_{j\neq i}\{\eta_t(\hat{c}^i_{<t}-\hat{c}^j_{<t})+q^j\}}^{\infty} d\mu(q^i)d\mu(q^{\neq i}) = \int e^{-(\max_{j\neq i}\{\eta_t(\hat{c}^i_{<t}-\hat{c}^j_{<t})+q^j\})^+} d\mu(q^{\neq i}) \qquad (13)\\ &\leq \int e^{\frac{\eta_t}{p^i}} e^{-(\max_{j\neq i}\{\eta_t(\hat{c}^i_{<t}-\hat{c}^j_{<t})+q^j\}+\frac{\eta_t}{p^i})^+} d\mu(q^{\neq i})\\ &\leq e^{\frac{\eta_t}{p^i}} \int e^{-(\max_{j\neq i}\{\eta_t(\hat{c}^i_{1:t}-\hat{c}^j_{1:t})+q^j\})^+} d\mu(q^{\neq i}) = e^{\frac{\eta_t}{p^i}}\pi^i. \end{aligned}$$

Hence, $\pi^i \geq p^i e^{-\frac{\eta_t}{p^i}} \geq p^i\left(1-\frac{\eta_t}{p^i}\right) = p^i - \eta_t$, which implies

$$\begin{aligned} \mathbf{E}_{q_*,q_t,r_t,u_t}[\hat{c}_t^{\widetilde{\mathrm{FPL}}}|h_{<t}] &= \sum_{i=1}^n p^i \sum_{j=1}^n p^j \mathbb{1}_{i=j}\frac{c_t^i}{p^i} = \sum_{i=1}^n p^i c_t^i\\ &\leq \sum_{i=1}^n \pi^i c_t^i + n\eta_t = \mathbf{E}_{q_*,q_t,r_t,u_t}[\hat{c}_t^{\widetilde{\mathrm{IFPL}}}|h_{<t}] + n\eta_t. \end{aligned}$$

This shows the step from feasible to infeasible FPL. The last step from infeasible FPL to the best decision in hindsight proceeds as shown already above and in [KV03, HP05]. Like before, it causes the upper bound of the cumulative regret to increase by $\frac{\log n}{\eta_T}$. This is true for any $(q_{1:T}, r_{1:T}, u_{1:T})$, hence also in expectation. The total regret is thus upper bounded by $\frac{\log n}{\eta_T} + n\sum_{t=1}^T \eta_t$, and we have just proved:

Theorem 4 *The algorithm FPL (1), obtaining cost estimates according to (12) and with learning rate* $\eta_t = \sqrt{\frac{\log n}{2nt}}$ *achieves a regret of at most*

$$\mathbf{E}c_{1:T}^{\mathrm{FPL}} - c_{1:T}^i \leq 2\sqrt{2Tn\log n} \quad \textit{for any } i \in \{1\dots n\}. \qquad (14)$$

We would like to point to a quite remarkable symmetry break here. It is straightforward to formulate FPL and the analysis from Section 3 for *reward maximization* instead of cost minimization. Then the (perturbed) leader is the expert with the highest (perturbed) reward, and perturbations are added to the scores. In the full information game, this reward maximization is perfectly symmetric to cost minimization by just setting $\mathit{reward}_t^i = 1 - c_t^i$: all probabilities, distributions, and outcomes

will be exactly the same. This is different in the partial observation case: There, in case of reward, the expert by FPL is the only one which can gain score. This is an advantage, in contrast to the disadvantage in case of loss minimization: Here, the selected expert is the only one to worsen its score. Put it differently, there is an automatic exploration or self-stabilization in the cost minimization case. With this intuition, it is less surprising that we did not need explicit exploration in Theorem 4. The corresponding result for reward maximization would not hold, as simple counterexamples show. Formally, it is the step from FPL to infeasible FPL which fails: A computation similar to (13) only shows $\pi^i \leq p^i e^{\frac{\eta_t}{p^i}}$, which does not imply a sufficiently strong assertion in general. However, reintroducing the exploration rate γ_t, we may set $\eta_t = \frac{\gamma_t}{n}$. This implies $\frac{\eta_t}{p^i} \leq 1$ for all i, hence $e^{\frac{\eta_t}{p^i}} \leq 1 + 2\frac{\eta_t}{p^i}$. Letting $\gamma_t = \sqrt{\frac{n \log n}{t}}$, we can conclude a bound like (14).

4.1 A computationally feasible algorithm

We conclude this section by discussing a computationally feasible variant of FPL using all observations. This algorithm is constructed in a straightforward way: Select the current action $i = I_t^{\mathrm{FPL}}$ according to FPL and substitute the estimate $\hat{c}_t^i$ from (12) by $\hat{c}_t^i = \frac{c_t^i}{\hat{p}_t^i}$. It remains to estimate $\hat{p}_t^i$ by a Monte Carlo simulation.

There are two possibilities of error: either $\hat{p}_t^i$ overestimates p_t^i, or it underestimates p_t^i. The respective consequences are different: If $\hat{p}_t^i > p_t^i$, then the instantaneous cost of the selected expert is just underestimated. We can account for this by adding a small correction to the instantaneous regret. At the end of the game, we perform well with respect to the underestimated costs, which are upper bounded by the true costs. This does not cause any further problems. The case $\hat{p}_t^i < p_t^i$ is more critical, since then at the end of the game we perform well only w.r.t. overestimated costs. We therefore have to treat this case more carefully.

Problems arise if the true probability p_t^i is very close to 0, as then the Monte Carlo sample might contain very few or no hits and the variance of the estimated cost is high. Since FPL does not prevent this case, we reintroduce γ_t as an "exploration threshold". Let $\gamma_t = \frac{1}{2\sqrt{t}} \leq \frac{1}{2}$. We first assume that $p_t^i \geq \gamma_t$. If this assumption is false but we use $\hat{p}_t^i \geq \gamma_t$, then $\hat{p}_t^i$ is an overestimate and we have to consider an additional instantaneous regret. This case has probability at most γ_t. Consequently, as (true) instantaneous costs are always bounded by 1, the additional instantaneous regret is at most γ_t.

We sample the perturbed leader $k \in \mathbb{N}$ times and denote by $a^i(k)$ the number of times the leader happens to be expert i. Recall that expert i is the one already selected by FPL. By Hoeffding's inequality, the distribution of $\frac{a^i(k)}{k}$ is sharply peaked around its mean p^i:

$$\mathbf{P}\left[\frac{a^i(k)}{k} - p^i \geq \frac{\gamma_t^2}{\sqrt{2}}\right] \leq e^{-\gamma_t^4 k} \text{ and } \mathbf{P}\left[\frac{a^i(k)}{k} - p^i \leq -\frac{\gamma_t^2}{\sqrt{2}}\right] \leq e^{-\gamma_t^4 k}.$$

We choose k such that the probability bounds on the r.h.s. are at most γ_t, i.e. $e^{-\gamma_t^4 k} \leq \gamma_t$. Consequently we should sample $k = \lceil \gamma_t^{-4} \log(\gamma_t^{-1}) \rceil = \lceil 2t^2 \log(2\sqrt{t}) \rceil$ times. Hence the sampling complexity of the algorithm is $O(t^2 \log t)$. Let

$$\hat{p}_t^i := \max\left\{\gamma_t, \tfrac{a^i(k)}{k} - \tfrac{\gamma_t^2}{\sqrt{2}}\right\},$$

then $\hat{p}_t^i \leq p_t^i$ with probability at least $1 - \gamma_t$ (recall the assumption $p_t^i \geq \gamma_t$). Hence the possibility of overestimate $\hat{p}_t^i > p_t^i$ causes an additional regret of γ_t.

Finally we need to deal with possible underestimates. For some integer $m \geq 1$, the probability that $\hat{p}_t^i$ falls below $p_t^i - \frac{(\sqrt{m}+1)\gamma_t^2}{\sqrt{2}}$ is at most

$$\mathbf{P}\left[\tfrac{a^i(k)}{k} - p^i \leq -\tfrac{\sqrt{m}\gamma_t^2}{\sqrt{2}}\right] \leq e^{-m\gamma_t^4 k} \leq \gamma_t^m \tag{15}$$

by Hoeffding's inequality. We partition the interval $[\gamma_t, p_i^t)$ of all possible underestimates into subintervals $A_1 = \left[p_t^i - \frac{2\gamma_t^2}{\sqrt{2}}, p_t^i\right)$ and

$$A_m = \left[p_t^i - \tfrac{(\sqrt{m}+1)\gamma_t^2}{\sqrt{2}}, p_t^i - \tfrac{(\sqrt{m-1}+1)\gamma_t^2}{\sqrt{2}}\right), \quad m \geq 2.$$

We do not need to consider m with the property $A_m \cap [\gamma_t, p_i^t) = \emptyset$. That is, we can restrict to m small enough that $p_i^t - \sqrt{\frac{1}{2}}(\sqrt{m}+1)\gamma_t^2 \geq \gamma_t - \sqrt{\frac{1}{2}}\gamma_t^2$. Let M be the largest m for which this condition is satisfied, then one can easily see $\sqrt{m}+1 \leq \sqrt{M}+1 \leq \sqrt{2}(p - \gamma_t + \sqrt{\frac{1}{2}}\gamma_t^2)/\gamma_t^2$.

Claim 5 *If $m \leq M$, then $\frac{c_t^i}{p_t^i - (\sqrt{m}+1)\gamma_t^2/\sqrt{2}} \leq \frac{c_t^i}{p_t^i} + \gamma_t(\sqrt{m}+1)$.*

This follows by a simple algebraic manipulation. Consequently, for $\hat{p}_i^t \in A_m$, we have $\mathbf{E}\hat{c}_t^i \leq c_t^i + (\sqrt{m}+1)\gamma_t$. Moreover, $\hat{p}_t^i \in A_m$ occurs with probability at most γ_t^{m-1} according to (15). By bounding the expectation over all A_m, we thus obtain an additional regret of at most

$$\sum_{m=1}^{M}(\sqrt{m}+1)\gamma_t^m \leq \gamma_t \sum_{m=0}^{\infty}(m+2)\gamma_t^m \leq \frac{2\gamma_t}{1-\gamma_t} + \frac{\gamma_t^2}{(1-\gamma_t)^2} \leq 5\gamma_t,$$

since $\gamma_t \leq \frac{1}{2}$. Altogether, this proves the following theorem.

Theorem 6 *Let $\gamma_t = \frac{1}{2\sqrt{t}}$ be the exploration threshold. In each time step, after selecting one expert i, let FPL obtain an estimate $\hat{p}_t^i = \max\left\{\gamma_t, \frac{a^i(k)}{k} - \frac{\gamma_t^2}{\sqrt{2}}\right\}$ for $\mathbf{P}(I_t^{\mathrm{FPL}} = i)$, by sampling the perturbed leader $k = \lceil 2t^2 \log(2\sqrt{t}) \rceil$ times and counting the number of hits $a^i(k)$. Let the estimated cost of the selected expert be $\hat{c}_t^i = c_t^i/\hat{p}_t^i$, and the estimated cost of all other experts be zero. Then the algorithm FPL (1) with learning rate $\eta_t = \sqrt{\frac{\log n}{2nt}}$ achieves a regret of at most*

$$\mathbf{E}c_{1:T}^{\mathrm{FPL}} - c_{1:T}^i \leq 2\sqrt{2Tn\log n} + 7\sqrt{T} \quad \textit{for any } i \in \{1 \ldots n\}. \tag{16}$$

For $t = 1, 2, 3, \ldots$
set $\hat{c}_t^i = 0$ for $i \in \{i : t \geq \tau^i\}$ and $\hat{c}_t^i = \big(\gamma_t \min\{w^i : t \geq \tau^i\}\big)^{-1}$ for $i \notin \{i : t \geq \tau^i\}$
sample $r_t \in \{0, 1\}$ independently s.t. $P[r_t = 1] = \gamma_t$
If $r_t = 0$, set $I_t^b = \arg\min_{i:t\geq\tau^i} \big\{\hat{c}^i_{<t} + \frac{\log w^i - q_t^i}{\eta_t}\big\}$ (FPL on the active experts)
If $r_t = 1$, sample $I_t^b \in \{i : t \geq \tau^i\}$ according to the weights $\frac{w^i}{\sum_{i:t\geq\tau^i} w^i}$
play decision I_t^b and observe cost $c_t^{I_t^b}$
If $r_t = 1$, set $\hat{c}_t^{I_t^b} = \big[c_t^{I_t^b} \sum_{i:t\geq\tau^i} w^i\big] / \big[\gamma_t w^{I_t^b}\big]$

Figure 2: The algorithm bFPL for infinite expert class. The entering times τ^i, the exploration rate γ_t, and the learning rate η_t, will be specified in Theorem 7.

5 Infinite expert classes

Here, we sketch a variant of bFPL, taken from [PH05], with guaranteed worst-case performance against a bandit with countably infinitely many arms. So we consider the following setup: The adversary subsequently generates cost vectors $c_t \in [0, 1]^\infty$, and at each time t we have to select one index or expert i and incur its cost c_t^i. We learn only the cost of the selected expert.

As a prerequisite, we need that each of the infinitely many experts is associated with a *prior weight* w^i such that $\sum_i \leq 1$. Since in order to obtain a cost estimate $\hat{c}$, the observed cost is divided by the weight of the sampled expert, we have to be careful not to admit too small weights. We need to keep control of the maximum possible expected cost, since otherwise the step from FPL to IFPL would be problematic. One possibility to do so is defining an *entering time* $\tau^i \geq 1$ for each expert. Prior to τ^i, the expert is not active and cannot be chosen. We choose $\tau^i = \left\lceil \left(\frac{1}{w^i}\right)^{\frac{1}{\alpha}} \right\rceil$, with $0 < \alpha < 1$ to be defined later. Then it is not hard to see that the minimum weight of any active expert at time t is lower bounded: $\min\{w^i : t \geq \tau^i\} \geq t^{-\alpha}$. Letting the exploration rate be $\gamma_t = t^{-\beta}$ with $0 < \beta < 1$ to be defined later, the maximum unbiasedly estimated cost is at most $t^{\alpha+\beta}$. For the step from FPL to IFPL to go through, we thus may choose $\eta_t = t^{-2\alpha-2\beta}$. Then both steps from bFPL to FPL and from FPL to IFPL each cause a regret of at most $\sum_{t=1}^T t^{-\beta} \leq \frac{1}{1-\beta} T^{1-\beta}$. On the other hand, $\frac{1}{\eta_T}$ causes a regret of at most $T^{2\alpha+2\beta}$. In order to minimize these bounds simultaneously, we choose $\beta = \frac{1-2\alpha}{3}$.

In order to guarantee that the step from IFPL to some fixed expert holds, we have to correctly assign estimated costs to inactive experts. For example, if an expert enters the game and previously has been assigned no estimated cost at all, then a bound w.r.t. this expert may be difficult to obtain. We therefore assign *maximum possible* estimated costs to all inactive experts. Then one can show [PH05] that, evaluating the expected costs, the step IFPL to some fixed reference expert holds almost without modification. Clearly, the reference expert's estimated costs now exceed its true costs by at most $\sum_{t=1}^{\tau^i-1} t^{\alpha+\beta}$, which is easily shown to be upper bounded by $\left(\frac{1}{w^i}\right)^{1+\frac{1}{\alpha}+\frac{\beta}{\alpha}}$.

This gives another *additive* bound to the regret in terms of the weight of the reference expert – there is not multiplicative factor of $\frac{1}{w^i}$ any more. This is an artifact of the design of the algorithm and proof technique and does not mean that the new variant performs better than the old one. Actually, since $\alpha > 0$, the bound is now $O\big(t^{\frac{2}{3}+\frac{2\alpha}{3}}\big)$ as opposed to $O\big(t^{\frac{2}{3}}\big)$ before. Choosing a large α results in a small $\big(\frac{1}{w^i}\big)$ term, but the order in t gets large, while a small α has the opposite effect.

The complete algorithm is specified in Figure 2. The following statement, which improves on the bounds given in [PH05] (they are based on the workaround from [MB04]) is an example where we select $\alpha = \frac{1}{8}$.

Theorem 7 *Consider a bandit problem with countably many arms/experts, each expert i having a prior weight w^i such that the weights sum up to at most 1. Then the above described bFPL variant with entering times $\tau^i = \left\lceil \left(\frac{1}{w^i}\right)^8 \right\rceil$, exploration rate $\gamma_t = t^{-\frac{1}{4}}$, and learning rate $\eta_t = t^{-\frac{3}{4}}$, satisfies the regret bound*

$$\mathbf{E}c^{\text{bFPL}}_{1:T} - c^i_{1:T} \leq O\left(\left(\tfrac{1}{w^i}\right)^{11} + T^{\frac{3}{4}} \log w^i\right)$$

for all $T \geq 1$, any adaptive assignment of the cost vectors and any reference expert i.

The formal proof is omitted. It follows the outline of Theorem 2, using the arguments discussed above. Many of the arguments, including the step from IFPL to the reference expert, are formally carried out in [PH05].

6 Discussion

The main statement of this paper is the following:

> *If we have a regret minimization algorithm with a bound guaranteed against an oblivious adversary, and if the algorithm chooses the current action/expert by some independent random sampling based on past cumulative scores (e.g. FPL or weighted majority), then the same bound also holds against an adaptive adversary. This is true both for full and partial observations.*

We have used this argument for showing bounds for FPL in the adversarial bandit problem. The strategy to use only feedback from exploration rounds which is common for FPL achieves a regret bound of $O(t^{\frac{2}{3}})$. As the algorithm is label efficient, this bound is sharp. Using all observations allows to push the regret down to $O(\sqrt{t})$. Then however the sampling probabilities have to be approximated.

In the same way, it is possible to use our argument for the general geometric online optimization problem [MB04, AK04], also resulting in a $O(t^{\frac{2}{3}})$ regret bound against adaptive adversary. An interesting open problem is the following: Under

which conditions and how is it possible to use all observations in the geometric online optimization problem, hopefully arriving at a $O(\sqrt{t})$ bound?

We conclude with a note on *regret against an adaptive adversary*. We considered the *external* regret w.r.t. the best action/strategy/expert from a pool. There are two directions from here. One is to go to different regret definitions, such as internal regret. The other one is to change the reference and compare to the *hypothetical* performance of the best strategy, in this way accepting a stronger type of dependency of the future costs from the currently selected action (see e.g. [PH05] and the references therein). It is one of the major open problems to propose refined algorithms and prove better bounds in this model.

References

[ACBFS95] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proc. 36th Annual Symposium on Foundations of Computer Science (FOCS 1995)*, pages 322–331, Los Alamitos, CA, 1995. IEEE Computer Society Press.

[ACBFS03] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The non-stochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, Feb. 2003.

[AK04] B. Awerbuch and R. D. Kleinberg. Adaptive routing with end-to-end feedback: distributed learning and geometric approaches. In *STOC '04: Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 45–53, 2004.

[CBLS04a] N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Minimizing regret with label efficient prediction. In *17th Annual Conference on Learning Theory (COLT)*, volume 3120 of *Lecture Notes in Computer Science*, pages 77–92. Springer, 2004.

[CBLS04b] N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Regret minimization under partial monitoring. Technical report, 2004.

[FS97] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[Han57] J. Hannan. Approximation to Bayes risk in repeated plays. In M. Dresher, A. W. Tucker, and P. Wolfe, editors, *Contributions to the Theory of Games 3*, pages 97–139. Princeton University Press, 1957.

[HP05] M. Hutter and J. Poland. Adaptive online prediction by following the perturbed leader. *Journal of Machine Learning Research*, 6:639–660, 2005.

[KV03] A. Kalai and S. Vempala. Efficient algorithms for online decision. In *Proc. 16th Annual Conference on Learning Theory (COLT-2003)*, Lecture Notes in Artificial Intelligence, pages 506–521, Berlin, 2003. Springer.

[MB04] H. B. McMahan and A. Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *17th Annual Conference on Learning Theory (COLT)*, pages 109–123. Springer, 2004.

[PH05] J. Poland and M. Hutter. Defensive universal learning with experts. 2005. International Conference on Algorithmic Learning Theory (ALT), to appear.