Available online at www.sciencedirect.com

ScienceDirect

Theoretical Computer Science 397 (2008) 77–93

Theoretical Computer Science

www.elsevier.com/locate/tcs

Nonstochastic bandits: Countable decision set, unbounded costs and reactive environments[☆]

Jan Poland

ABB Switzerland Ltd. Corporate Research, Segelhof, CH - 5405 Baden, Switzerland

Abstract

The nonstochastic multi-armed bandit problem, first studied by Auer, Cesa-Bianchi, Freund, and Schapire in 1995, is a game of repeatedly choosing one decision from a set of decisions ("experts"), under partial observation: In each round t, only the cost of the decision played is observable. A regret minimization algorithm plays this game while achieving sublinear regret relative to each decision. It is known that an adversary controlling the costs of the decisions can force the player a regret growing as $t^{\frac{1}{2}}$ in the time t. In this work, we propose the first algorithm for a countably infinite set of decisions, that achieves a regret upper bounded by $O(t^{\frac{1}{2}+\varepsilon})$, i.e. arbitrarily close to optimal order. To this aim, we build on the "follow the perturbed leader" principle, which dates back to work by Hannan in 1957. Our results hold against an adaptive adversary, for both the expected and high probability regret of the learner w.r.t. each decision. In the second part of the paper, we consider reactive problem settings, that is, situations where the learner's decisions impact on the future behaviour of the adversary, and a strong strategy can draw benefit from well chosen past actions. We present a variant of our regret minimization algorithm which has still regret of order at most $t^{\frac{1}{2}+\varepsilon}$ relative to such strong strategies, and even sublinear regret not exceeding $O(t^{\frac{4}{5}})$ w.r.t. the *hypothetical* (without external interference) performance of a strong strategy. We show how to combine the regret minimizer with a universal class of experts, given by the countable set of programs on some fixed universal Turing machine. This defines a *universal learner* with sublinear regret relative to any computable strategy.

Keywords: Multi-armed bandit; Nonstochastic bandit; Countable decision set; Partial observations; Reactive environments

1. Introduction

For the last 15 years, online regret minimization algorithms have played an important role in learning theory. The setup for online regret minimization is most easily and accurately formalized as a *game* between a learner and an adversary. The game proceeds in discrete time $t = 1, 2, \dots,$ In each round t, the learner selects a decision from a set of possible decisions, while the adversary assigns costs to all possible decisions, both players without knowing the other's move. Then the learner's decision is revealed, and the learner incurs the corresponding cost.

[☆] This research was done while the author was with Hokkaido University, Graduate School of Information Science and Technology.
E-mail address: jan.poland@ch.abb.com.

doi:10.1016/j.tcs.2008.02.024

A crucial issue which impacts on both the learner's algorithm and its achievable performance is *what or how much does the learner observe*. In the full information game, also widely known as "prediction with expert advice", decisions are also called experts, and the learner observes the costs of all experts. This enables him to play a randomized strategy that guarantees, with high probability, a performance almost as well as the best decision in hindsight. More precisely, there are algorithms that achieve a regret of order $\sqrt{T \log n}$ at any horizon time T. Here, n is the (finite) number of available decisions, and regret is the difference between the learner's cumulative performance and the best expert in hindsight (after time T) [19,26,10]. (The reader who is not familiar with expert algorithms at all, should see the references, but at least Sections 2.1 and 2.2 for a formal definition of regret.)

Games where the learner does not observe the full cost vectors are termed "partial observation". The most important and widely studied partial observation game is the so-called "bandit" setup, where in each round t, the learner observes just the cost of the decision played, not of the alternatives. In analogy to gambling slot machines, decisions are also called "arms of the bandit" in this setup. Bandits have an important motivation in and were originally considered for medical treatment, where different available treatments for a certain disease need to be explored, but on the other hand, it is highly desirable to apply successful treatments as often and others as rarely as possible. In this way, bandits are one of the most simple mathematical formalizations of an *exploration vs. exploitation* tradeoff. Much work has been done on stochastic bandits, i.e. situations where each arm obeys a stationary probability distribution [5].

It may be surprising at a first glance that performance guarantees can be proved for *adversarial* bandits. The first respective result was shown by [2], who proved that there is an algorithm such that after any time T, the regret of the learner w.r.t. each arm (however *not* the best arm in hindsight) is at most $O(\sqrt{Tn \log n})$, with high probability, where again n is the finite number of arms. This bound is almost sharp, since the same authors show that an oblivious adversary (this is even weaker than an adaptive one, see Section 1.4) can force *any* learner a regret of $\Omega(\sqrt{Tn})$.

1.1. Results of this work

Most of the literature restricts to a finite number of decisions. In contrast, in this work we consider nonstochastic bandits with countably many arms. In this case, we cannot treat all arms equally, but we must introduce a *prior*: Each decision i, where they are enumerated $i = 1, 2, \ldots$, is assigned with a prior weight w^i, such that $\sum_{i=1}^{\infty} w^i \leq 1$. (The reader may have noticed that we use the terms "decision", "expert", and "arm" interchangeably.) The following theorem is the main result of Section 2.[1]

Theorem 1. *Consider a nonstochastic bandit problem with countably many arms $i = 1, 2, \ldots$, endowed with prior weights w^i such that $\sum_{i=1}^{\infty} w^i \leq 1$. The costs of the arms are controlled by an adaptive adversary. Then, for any $\varepsilon > 0$, there is an algorithm which achieves, for any time T and w.r.t. any arm i and with arbitrarily high probability, a regret of at most*

$$O\left(T^{(\frac{1}{2}+\varepsilon)} \log w^i + (w^i)^{-\frac{1}{2\varepsilon}}\right).$$

Since a lower bound of $\sqrt{T}$ is known, this says that the regret growth rate is close to optimal.

This result is new and improves on the best known bound before ($T^{\frac{2}{3}+\varepsilon}$) from [23]. In the following, we give a sketch of where are the difficulties to accomplish this and what are the contributions of this work. A key part of regret minimization algorithms is the learning rate, a parameter that balances exploration and exploitation and has to be tuned dynamically if the algorithm is supposed to perform well uniformly over time (that is, after any time T). Dynamic learning rate however hampers the analysis, in particular for the well-studied weighted averaging schemes. It was [15] who made popular a different type of algorithm, "follow the perturbed leader" (*FPL*), which dates back to [11] and substantially facilitates working with dynamic learning rate. The present work builds on the *FPL* strategy. However, another central ingredient for bandit regret minimizers is the concept of unbiased performance estimates, which is achieved by dividing the actually observed cost by the probability of selecting the actual arm. Since this probability is not directly available for the *FPL* strategy (as opposed to weighted averaging schemes), we propose a

[1] The results in this section are stated in order to provide an overview. The exact definitions, in particular of "long-term" and "hypothetical" regret used in Theorem 3, will be given later on in the technical parts.

way to estimate it with a Monte-Carlo sampling. Hence we need to bound the regret caused by this estimation error. In contrast, all FPL variants before were based on the "label efficient" (or rather "wasteful") way of using only the observations from designated exploration rounds. As one can show [7], this necessarily implies a worse regret bound of order at least $t^{\frac{2}{3}}$. For the analysis to go through, some steps require a finite set of decisions. We therefore need to restrict the number of arms we work with at any time and successively introduce the arms into the game.

A second contribution of this work concerns *reactive environments*. In this case, which is is subject of Section 3, we assume that the future behaviour of the adversary (we will also call it the environment) depends on the learner's decisions or actions, in a way, such that well chosen actions cause low cost in future. Then, there may be strong experts that propose such favourable actions. However, such a strong expert usually needs more than just one time step to act in a favourable way and then reap the benefits, we refer to this performance as the *long-term performance* of the expert. A special case is an expert *optimal after* t_0 *steps*, which displays optimal performance after it is followed for a fixed contiguous number t_0 of steps, from any start state of the system. We call this type of performance after t_0 time steps (that is, ignoring the initial t_0 adaptation steps) the *hypothetical performance* of the expert. It may be much stronger than the actual performance, which is repeatedly disturbed by other experts, after which the optimal expert needs to "recover".

An example is playing repeated prisoner's dilemma against the "tit-for-tat" adversary (see the beginning of Section 3 if you are not familiar with this game): Although in prisoner's dilemma, defecting is the dominant action, repeatedly cooperating is optimal after two steps. Thus, the long-term performance of cooperating is strong. Its hypothetical performance is even optimal (against tit-for-tat), since this is just the performance of exclusively cooperating, without disturbance from the defecting expert.

Our main tool for obtaining assertions for reactive environments is considering *unbounded costs*, precisely we allow the costs to grow in time. We will show (please compare again the previous footnote).

Theorem 2. *Suppose that the actual costs are bounded by a function increasing in time sufficiently slowly, hence, over all time steps, the costs are unbounded. Then a regret of at most*

$$O\left(T^{(\frac{1}{2}+\varepsilon)}\log w^i + (w^i)^{-(\frac{1}{\varepsilon}+\frac{1}{2})}\right)$$

is achievable.

Basing on Theorem 2, we will prove the following statement on the long-term and hypothetical regret in Section 3.

Theorem 3. *Suppose a reactive environment. For any $\varepsilon > 0$, there is an algorithm which achieves, for any time T, any t_0, and relative to the* long-term performance *of each expert, a regret of at most*

$$O\left(T^{(\frac{1}{2}+\varepsilon)}\log w^i + (w^i)^{-(\frac{1}{\varepsilon}+\frac{1}{2})}\right),$$

where w^i is the prior weight of the reference expert.

Moreover, there is an algorithm that performs well relative to the hypothetical *performance of each expert, where the regret does not exceed*

$$O\left(T^{\frac{4}{5}}(\log w^i + t_0) + (w^i)^{-12.5}\right).$$

Hence, if one strategy i is optimal after t_0 steps, this algorithm is performed almost optimally, with the above regret.

Reactive bandit problems have been first considered by [9]. We propose a technically much simpler way to achieve similar results, building on standard regret minimization algorithms. Moreover, while they do not state the growth rate of the regret for their algorithm explicitly, we show close to optimal growth rate. The idea is the following: Given the bounds for a regret minimization algorithm under the assumption that the instantaneous costs may slowly grow in time, we use this algorithm with uniformly bounded costs, but instead we *yield the control* to the respective selected expert for a gradually increasing number of time steps. In other words, we perform a change of time scale.

1.2. Motivation

Considering countably many arms or decisions or experts, straightforwardly generalizes the finite case. Moreover, the large body of literature dealing with finitely many arms, in reality only covers the special case of *uniform* prior weights: no prior preference is given to any decision. If we want to do otherwise and deal with nonuniform weights (for instance, we might trust some experts more than others), then even in the finite case we need techniques as proposed in this work.

There is a strong motivation for countable expert classes in the theory of computation. It is straightforward to interpret *any* program on some fixed universal Turing machine as an expert, by running it on the complete history of past observations (or on the part of it we decided to memorize) and interpreting its output appropriately as an action or decision. Since the set of all programs is countable, this construction naturally yields a countable expert class, and thus, combined with an appropriate regret minimization algorithm, a *universal* regret minimization algorithm or "universal agent".

There is a dual way to use the set of all programs on a universal Turing machine for learning, namely Bayesian learning. In fact, each program can be related to a (semi-)probability distribution, this construction is at the core of algorithmic information theory, see e.g. [18] (however it is more complicated than the direct correspondence to an expert). Then one can use the Bayes mixture over all these distributions for learning, actually it turns out that this Bayes mixture is equivalent to the a-priori probability distribution of the Turing machine. It has excellent learning properties, as first shown by Solomonoff [25]. For reactive environments, a construction of a universal agent (AIξ) has been suggested by Hutter [12]; however, this differs technically from the universal agent based on regret minimization constructed here.

1.3. Related work

We have already mentioned some relations of our work to others: Our regret minimization algorithm will basically solve similar problems (but with countably many arms) as does the "Exp3" algorithm for nonstochastic bandits due to Auer, Cesa-Bianchi, Freund, and Schapire [2,3]. However, it is not based on weighted averaging schemes [19,26,10], but on the follow the perturbed leader (*FPL*) principle [11,15]. The first full observation *FPL* algorithm for countable experts classes was suggested in [13,14], the first bandit *FPL* algorithm (with finite expert class) in [20]. The results in this work build and improve on [23,22]; however, these papers used explicit exploration as suggested by [20], which we abandon in this work. Actually, [20] construct a label efficient learner [6,7], which has a worse lower bound ($t^{\frac{2}{3}}$) on the regret than our upper bound ($t^{\frac{1}{2}+\varepsilon}$).

The present work is possibly the first to consider countable decisions sets in the bandit setup. However, continuous decision spaces have been studied in [16,17]. Moreover, [4] proposed an alternative method of using a nonuniform prior in the case of finitely many experts.

We indicated that *FPL*, as opposed to weighted averaging algorithms, may greatly facilitate the analysis, in particular for dynamic learning rate, which should be used if the time horizon T is not known in advance. We will now further discuss another feature of *FPL*: It permits the *efficient* treatment of *geometrical online optimization problems*, where the learner's decisions are linearly composed of base decisions, and it is assumed that the set of base decisions is of reasonable size, while the set of decisions is exponentially large. An example is the online shortest path problem in a graph with adversarially changing path costs. Most literature dealing with *FPL* actually considers this type of problem ([15,20] and many others).

The important issue of adaptive vs. oblivious adversary is discussed in the next subsection.

1.4. Adaptive and oblivious adversaries

In the context of game playing, it is natural to assume that the adversary is adaptive, i.e. it observes our decisions and adapts future behaviour accordingly. The other model frequently considered is that of oblivious adversary, which is equivalent to requiring that the adversary decides *all* cost vectors before the game starts. In the full information game, there is actually no difference in the worst case: Regret minimization strategies based on randomized sampling according to past costs have the *same* performance guarantees against oblivious and adaptive adversary, as shown in [14, Lemma 12]. In particular, all regret bounds are relative w.r.t. the best expert *in hindsight*.

In partial observation games, things are more complicated. In order to treat the adaptive adversary correctly, costs need to be modelled as a stochastic process, and there is no tool like [14, Lemma 12] which allows us to conclude from oblivious to adaptive adversary. As a consequence, many authors (e.g. [3]) focused their exposition on oblivious adversaries. For weighted averaging learners, it is, however, not hard to see that the same regret bounds *relative to each fixed expert* (not w.r.t. the best expert in hindsight) hold for adaptive adversary. For *FPL* in contrast, this is not obvious, since the analysis crucially builds on "lazy" randomization which does not perform well for adaptive adversary. The key assertion of [22] is that still, as soon as we switch from "lazy" to usual randomization, the desired regret bounds do hold for *FPL*.

However, the regret against the best arm in hindsight is still a strictly stronger notion, as pointed out by [20,8]: If R^i_T is the random variable denoting the regret relative to expert i at time T, then strictly $\mathbf{E}\max_i R^i_T > \max_i \mathbf{E}R^i_T$ holds. [20] sacrifice some regret growth rate in order to prove bounds w.r.t. the best expert in hindsight. The recent and sophisticated construction in [8] achieves this goal with optimal $t^{\frac{1}{2}}$ growth rate for a finite expert class.

In this work, we will consider adaptive adversary, and prove regret bound relative to each (fixed) expert.

1.5. Discussion

Before starting the technical presentation in the next section, we discuss our results and some open question on an informal level. What is the use of our regret minimization algorithm, and what are the new contributions of this work?

Although a regret growth rate $t^{\frac{1}{2}+\varepsilon}$ arbitrarily close to optimal can be achieved, the bounds seem hardly relevant for practical applications. The quantity that worries most is the $1/poly(w^i)$, which is exponential in the complexity k^i of the reference arm's complexity and typically huge. This quantity also appears in the lower bounds [3]. Hence it is important and, to our knowledge, open so far to find special cases which do not have this problem.

Another interesting nontrivial question is the following: Are the learning properties of regret minimization algorithms really desirable? What does "good" or even "optimal" learning mean for reactive settings? For illustration, we briefly present an example from [24], the reader can find more discussion there. Consider the repeated game of "chicken".[2] In this game, it is desirable for the learner to become the "dominant defector", i.e. to defect in the majority or even all of the cases while the opponent cooperates. Let's call an opponent "primitive" if he/she agrees to cooperate after a fixed number of consecutive defecting moves of the learner, and let's call "stubborn" if this number is high. Then *reFPeL*, which is the algorithm we propose for reactive environments, learns to be the dominant defector against any primitive opponent, however stubborn. On the other hand, if the opponent is some learning strategy which also tries to maximize profit and learns faster (we conducted the experiment with AIξ [12]), then *reFPeL* settles for cooperating, and the opponent will be the dominant defector. Interestingly, however, AIξ would not learn to defect against a stubborn primitive opponent. Hence, already in this simple reactive setting, it is not clear which learning behavior should be really considered "good" or even "optimal".

Although this work is possibly the first to consider countable decision spaces, an assertion like our first main result Theorem 1 can be proved for the Exp3 algorithm from [3], if the doubling trick is used and in each phase only finitely many, namely $O(t^{2\varepsilon})$, decisions are considered. The result then follows[3] without much effort from [3, Theorem 3.1]. The result of the present paper is still interesting, as it avoids the doubling trick but uses a smooth adaptation of the learning rate, and it is based on the *FPL* principle, as opposed to weighted averaging. Moreover, we will present the whole analysis against an adaptive adversary.

Compared with the Exp3 variant sketched in the last paragraph, parts of the present proofs may seem complicated. As the basic *FPL* analysis is quite elegant [15], this is primarily due to the fact that the explicit sampling probabilities for *FPL* algorithms are not known. On the other hand, they are necessary in order to get close to optimal bounds in

[2] This game, also known as "Hawk and Dove", can be interpreted as follows. Two coauthors write a paper, but each tries to spend as little effort as possible. If we succeed in letting the other do the whole work, there will be no cost. On the other hand, if no one does anything, there will be no paper, this translates to high costs for both. Finally, if both decide to cooperate, both incur some costs. We chose the cost matrix as $\binom{1\ 0.8}{0\ 0.5}$, the learner is the column player, and choosing the first column means to defect, the second to cooperate. The opponent is the row player, whose cost matrix is the transpose, i.e. $\binom{1\ \ 0}{0.8\ 0.5}$. For example, if the player defects and the opponent cooperates, then the player has no cost, while the opponent has cost of 0.8. Hence, in the repeated game, it is "socially optimal" to take turns cooperating and defecting.

[3] I am grateful to the referee who pointed this out.

the Bandit setup. Therefore, the error introduced by using estimated probabilities instead of true ones needs to be bounded, which the present work is probably the first to do.

Although technically not difficult, the unbounded losses, which are also possibly a new contribution of this work, allow some nice constructions. In particular, the treatment of reactive environments can be considerably simplified relative to [9]. Moreover, we define a "universal" learner based on expert advice, which is probably the first one that has provable guarantees w.r.t. all algorithms on some universal Turing machine.

2. Bandits with countably many arms

2.1. Notation

The learner's task is to choose repeatedly one arm of a bandit with countably many arms. We will use the terms "arm" and "expert" and "decision" interchangeably. Each arm $i \geq 1$ is assigned a weight $w^i \in (0, 1)$, such that $\sum_i w^i \leq 1$. We set $k^i = -\log w^i$ (and mean the natural logarithm), this is the *complexity* of the arm. By Kraft's inequality, the k^i correspond to description lengths in "nats" for the arms, relative to some prefix code.

Our regret minimization game involves sequence $c_1, c_2, \ldots$ of cost vectors in time t. Each cost vector c_t has countably infinite dimension, $c_t = (c_t^i)_{i=1}^{\infty}$, thus it contains the cost of each arm i at time t. Superscript indices always refer to arms, subscript indices to time. If we have an algorithm A, then the arm the algorithm chooses at time t is denoted by I_t^A, and its cost incurred by $c_t^A = c_t^{I_t^A}$. Note that, since in general the adaptive adversary depends on the randomized decisions of the learner, all costs are random variables.

Cumulative costs up to time T are denoted by $c_{1:T} = \sum_{t=1}^{T} c_t$ if time T is included and $c_{<T} = c_{1:T-1}$ otherwise. Observe that $c_{1:T}, c_{<T}$ are infinite dimensional vectors. Similarly, we define quantities $c_{1:T}^i, c_{<T}^i$, and moreover $c_{1:T}^A$ for an algorithm A etc. The *regret* of the cumulative cost of an algorithm A w.r.t. a fixed decision i is denoted by

$$\Delta_{1:T}[c^A, c^i] = c_{1:T}^A - c_{1:T}^i.$$

Note that we prefer this to the slightly heavier notation $\Delta_{1:T}[c_{1:T}^A, c_{1:T}^i]$, which is of course equivalent. If B is another algorithm, we define the regret of algorithm A relative to algorithm B analogously as $\Delta_{1:T}[c^A, c^B]$. Instantaneous regret (i.e. just at time t) is given by $\Delta_t[c^A, c^B]$. The notion of regret is sufficiently important to spend an extra symbol Δ. Actually, the regret is a *stochastic process* developing in time, and we will be always able to give bounds on the expected regret $\mathbf{E}\Delta_{1:T}[c^A, c^B]$, although there are no nontrivial bounds on any of the parts of the difference, $\mathbf{E}c_{1:T}^A$ or $\mathbf{E}c_{1:T}^B$.

2.2. The game

The following is the protocol of the nonstochastic bandit game we study:

```
For $t = 1, 2, 3, \ldots, T$
    Adversary selects cost vector $c_t \in [0, B_t]^\infty$
    Learner selects decision $i = I_t^{\text{Learner}}$
    Learner incurs and observes cost $c_t^i$
After time $T$, Learner's regret $\Delta_{1:T}[c^{\text{Learner}}, c^{i_0}]$ is evaluated
```

Here, the regret is measured relative to *any arm* i_0. The horizon T, the time at which the game ends and regret is evaluated, is *not known in advance* to the learner.

B_t is the *cost growth parameter*: The costs are not, as usual, uniformly bounded in $[0, 1]$, but in $[0, B_t]$, where B_t grows in t. As indicated in the introduction, B_t will be used in order to get to the long-term regret for reactive problems in Section 3. We assume that B_t is controlled by the learner. It could be also externally given, but in no case controlled by the adversary.

For simplicity, we assume the adversary to be deterministic throughout this paper. The results of this section immediately generalize to randomized adversary, as usually bounds for regret minimization algorithms do. In principle, this holds also for the results of Section 3, but the assertions regarding the time t_0 there are more difficult to state for a randomized adversary.

2.3. The algorithm *FPeL*

We can now define our regret minimization algorithm, *follow the perturbed estimated leader* (*FPeL*). Note that we do not need to care for explicit exploration, which is usually necessary (another algorithm that does not need explicit exploration was given by Allenberg [1]). The intuitive reason is the following: As becomes clear when looking at the algorithm, experts which are selected can only worsen their probability of being selected again; thus our *FPeL* explores implicitly. This is the opposite in other algorithms like Exp3 [3]. The algorithm contains some new notation, which is explained in the subsequent paragraphs.

```
Algorithm FPeL
For $t = 1, 2, 3, \ldots$
    Set $\hat{c}_t^i := 0$ for $i \in \{t \geq \tau\}$ and $\hat{c}_t^i := \frac{B_t}{\psi_t}$ for $i \notin \{t \geq \tau\}$
    Select and play $I_t^{\mathit{FPeL}} := \mathit{FPLsample}$
    Invoke FPLsample for $K := \lceil 16t^2 \log(2\sqrt{t}) \rceil$ times and
            set $a(K) :=$ the number of times $I_t^{\mathit{FPeL}}$ occurred
    Set $\hat{p}_t^{I_t^{\mathit{FPL}}} := \max\left\{\psi_t, \frac{a(K)}{K} - \frac{\psi_t^2}{\sqrt{2}}\right\}$
    Set $\hat{c}_t^{I_t^{\mathit{FPL}}} := c_t^{I_t^{\mathit{FPL}}} / \hat{p}_t^{I_t^{\mathit{FPL}}}$

Subroutine FPLsample
    Sample $q^i \overset{d.}{\sim} \mathit{Exp}$ independently for all $i \in \{t \geq \tau\}$
    Return $i^{\min} := \arg\min_{i:t \geq \tau^i} \{\eta_t \hat{c}_{<t}^i + k^i - q^i\}$
```

The subroutine *FPLsample* samples perturbation vectors q according to the *exponential distribution*. Recall that w^i are the weights and $k^i = -\log w^i$ are the complexities of the experts. For each expert i, we define an *introduction time* τ^i, that is a time from which the expert is used. By

$$\{t \geq \tau\} := \{i \geq 1 : t \geq \tau^i\}$$

we denote the set of experts that are active at time t. We choose the introduction times τ^i such that in each time step, always a finite number of experts is active.

Our algorithm makes use of a sequence of *learning rates* η_t that control the balance of exploration and exploitation, this is common in experts' algorithms. Moreover, we need *denominator thresholds*

$$\psi_t := \frac{1}{2\sqrt{t}},$$

bounding the denominator probability $\hat{p}_t^{I_t^{\mathit{FPL}}} := \max\left\{\psi_t, \frac{a(K)}{K} - \frac{\psi_t^2}{\sqrt{2}}\right\}$, which is used for obtaining "almost unbiased" cost estimates $\hat{c}_t^{\mathit{FPeL}}$. (The concept of unbiased cost estimates is widely used in weighted averaging based bandit algorithms, where the sampling probabilities are explicitly known. For *FPeL* in contrast, they can be only explicitly expressed for two experts.)

Each time randomness is used, it is assumed to be *independent* of the past randomness. Note in particular that the random vectors q are reinstantiated each time *FPLsample* is invoked, that is, the symbol q is repeatedly "reused", this simplifies notation. Before we start with the analysis, the reader should observe the following easy but important fact.

Proposition 4. *The algorithm FPeL is computationally feasible. In each time step, it employs only a finite number of experts. Moreover, the sampling complexity in order to determine* $\hat{p}_t^{\mathit{FPeL}}$ *is* $O(t^2 \log t)$.

2.4. Analysis

In this subsection, we will show the main technical results of this paper, ultimately arriving at the proof of Theorem 1. Similar to other work before [15,13,23], the proof is accomplished by considering two "intermediate"

t, T	denote the time and the horizon time, respectively
i	refers to an arm/expert/decision
c_t^i	$\in [0, B_t]$: cost of arm i at time t
B_t	≥ 1: upper bound on the costs at time t, grows in t
$c_{<T}$, $c_{1:T}$	cumulative cost vectors up to/including time T
η_t	> 0: learning rate, decreases in t
τ^i	≥ 1: introduction time for expert i
$\{t \geq \tau\}$	$= \{i : t \geq \tau^i\}$: set of arms active at time t
ψ_t	$= \frac{1}{2\sqrt{t}}$: denominator threshold
FPL	our main regret minimization algorithm
F*	virtual *FPL* variant using the “charging” perturbation q_*
IF*	virtual F* variant knowing the current estimated costs $\hat{c}_t$
p_t^i	$= \mathbf{P}[I_t^{\mathit{FPL}} = i]$: probability that *FPL*selects arm i in time t
q	vector of independently exponentially distributed perturbations, instantiated repeatedly
q_*	vector of independently exponentially distributed “charging” perturbations, instantiated once before the game starts
I_t^{FPL}	$= I_t^{F*} = I_t^{IF*}$: arm selected by *FPL* at time t
J_t^{F*}	arm that F* is charged at time t (selection w.r.t. q_*)
J_t^{IF*}	arm that IF* is charged at time t (selection w.r.t. q_*)
K	$= \lceil 16t^2 \log(2\sqrt{t}) \rceil$ at time t: #samples for estimating $\hat{p}_t^i$
$a(K)$	#occurrences of I_t^{FPL} in sampling for the estimate $\hat{p}_t^i$
$\hat{p}_t^i$	$= \max\left\{\psi_t, \frac{a(K)}{K} - \frac{\psi_t^2}{\sqrt{2}}\right\}$: estimate for p_t^i
$\hat{c}_t^i$	$= c_t^i / \hat{p}_t^i$: “almost unbiased” estimate for c_t^i
$\Delta_t[c^A, c^B]$	instantaneous regret of algorithm A relative to algorithm B

Fig. 1. List of notation.

algorithms F* and IF*. We show a sequence of regret bounds from *FPL* to F*, then from F* to IF*, and finally from IF* to a fixed arm i. *None of the intermediate algorithms is practically feasible* within the protocol defined in Section 2.2, they are “virtual” algorithms. However, all intermediate algorithms would be feasible with additional information, and in order to facilitate the analysis, the reader should imagine that all algorithms are in fact executed the way they are defined. Before we start, we introduce one more piece of notation, the true probability $p_t^i = \mathbf{P}[I_t^{\mathit{FPL}} = i]$ that *FPL* selects expert i at time t (recall that this quantity is not known exactly to the learner). Our notation is recapitulated in Fig. 1 (some notation displayed there is still to be introduced).

Our main analysis until Theorem 11 proceeds *in expectation*, we will state a high probability bound only at the very end. However, since the high probability bound relies on martingales, we really need the analysis for the *conditional expectation*, this will be stated as Corollary 12. The reader familiar with conditional expectation will notice that all bounds to come are in fact proven in conditional expectation: Lemmas 5, 6 and 9 hold instantaneously at time t in conditional expectation given the past randomness, while Lemma 10 holds in conditional expectation over the charging perturbation vector q_* (see (1)) given all other randomness. It will help the reader become familiar with these concepts keeping the conditional expectations in mind.

We now introduce the two intermediate algorithms F* and IF*.

F*: This algorithm proceeds exactly as *FPL* and also uses identical randomization q (recall that q is reinstantiated frequently). But additionally, in the beginning of the game, it samples a *single* infinitely dimensional “charging perturbation vector”

$$q_* = (q_*^i)_{i=1}^{\infty}, \text{ where } q_*^i \text{ are distributed indep. exponentially, for all } i. \tag{1}$$

By

$$J_t^{F*} = \arg\min_{i:t\geq\tau}\{\eta_t \hat{c}^i_{<t} + k^i - q^i_*\}, \tag{2}$$

we denote $F*$'s selection relative to the cost perturbation vector q_*. Then $F*$ still uses the original randomization q and the decisions $I_t^{F*} = I_t^{FPL}$ in order to build the estimates $\hat{c}_t$, but it is charged cost $c_t^{J_t^{F*}}$. Of course, we assume that the adversary still plays against *FPL*, the virtual algorithm participates by no means in the game. The following simple observation, namely that the expected costs of *FPL* and $F*$ coincide in each time step, is obviously true and starts the analysis.

Lemma 5. *For each* $t \geq 1$, *we have* $\mathbf{E}\Delta_t[c^{FPL}, c^{F*}] = 0$.

IF* is the other intermediate algorithm. It proceeds exactly as *FPL* and $F*$, using the identical randomness (meaning identical realizations, not just identically distributed samples), including the charging perturbation vector q_* from (1). The only difference to $F*$ is its selection J_t^{IF*} relative to the charging perturbation vector, which *assumes to know* $\hat{c}_t$ *already*

$$J_t^{IF*} = \arg\min_{i:t\geq\tau}\{\eta_t \hat{c}^i_{1:t} + k^i - q^i_*\}. \tag{3}$$

That is, $IF*$ is charged $c_t^{J_t^{IF*}}$, or if evaluated in terms of estimated costs, $\hat{c}_t^{J_t^{IF*}}$, while the cost estimates are still obtained using $I_t^{IF*} = I_t^{F*} = I_t^{FPL}$, hence, all three algorithms *FPL*, $F*$ and $IF*$ are using *identical* cost estimates. Recall that the intermediate algorithms are "virtual" and not actually feasible. Recall also that q_* is fixed at the beginning at the game, while the vector q is "reused" each time *FPLsample* is invoked.

Given Lemma 5, we may analyse $F*$ instead of *FPL*. The next Lemma relates the true costs c_t to the estimates $\hat{c}_t$.

Lemma 6. *When the denominator threshold is* $\psi_t = \frac{1}{2\sqrt{t}}$, *we have the following relations between c and* $\hat{c}$ *for each* $t \geq 1$.

(i) $\mathbf{E}\Delta_t[c^{F*}, \hat{c}^{F*}] \leq 2\psi_t B_t$,

(ii) $\hat{c}^i_t = \frac{c^i_t}{\hat{p}^i_t} \leq \frac{c^i_t}{p^i_t} + \psi_t B_t(\sqrt{m}+1) \leq \frac{c^i_t}{p^i_t} + \psi_t B_t(m+1)$ *with probability at least* $1 - \psi_t^m$ *for the expert i which was selected by* $F*$, *for all* $m \geq 1$,

(iii) $\mathbf{E}\Delta_t[\hat{c}^i, c^i] \leq 6\psi_t B_t$ *for all experts* $i \in \{t \geq \tau\}$.

Remark 7. Recall that $F*$ samples according to the perturbation q_t, but is charged according to the perturbation q_*. This means that $F*$'s expected estimated costs correctly evaluate to

$$\mathbf{E}\hat{c}_t^{F*} = \sum_{i\in\{t\geq\tau\}} p^i_t \sum_{j\in\{t\geq\tau\}} p^j_t \mathbb{1}_{i=j} \frac{c^i_t}{\hat{p}^i_t}.$$

That is, we need to sum over both probability distributions, and clearly, the cost is only different from zero if I_t^{FPL} and J_t^{F*} coincide.

Proof (*Of the Lemma*). Let $i = I_t^{FPL} = I_t^{F*} = I_t^{IF*}$ be the expert selected by *FPL*. Regarding the estimate $\hat{c}^i_t = \frac{c^i_t}{\hat{p}^i_t}$ for $\frac{c^i_t}{p^i_t}$, there are two possibilities of error: either $\hat{p}^i_t$ overestimates p^i_t, or it underestimates p^i_t. The respective consequences are different: If $\hat{p}^i_t > p^i_t$, then the instantaneous cost of the selected expert is just underestimated. When evaluating the algorithm in terms of estimated costs instead of true costs, we can account for this by adding a small correction to the instantaneous regret: This is done in (i). At the end of the game, we perform well with respect to the underestimated costs, which are upper bounded by the true costs.

The case $\hat{p}^i_t < p^i_t$ is covered by (ii)–(iv). It is more critical, since then at the end of the game we perform well only w.r.t. overestimated costs.

First we show (i). Problems would arise if the denominator $\hat{p}_t^i$ were very close to 0. This explains the name and the function of the *denominator threshold* $\psi_t = \frac{1}{2\sqrt{t}} \leq \frac{1}{2}$. We assume that $p_t^i \geq \psi_t$. If this assumption is false but we use $\hat{p}_t^i \geq \psi_t$, then $\hat{p}_t^i$ is an overestimate and we have to consider an additional instantaneous regret. In this case, we would have selected expert i with probability at most ψ_t, therefore the case has probability at most ψ_t. Consequently, as true instantaneous costs are always bounded by B_t, this causes an additional instantaneous regret of at most $\psi_t B_t$.

According to the definition of the algorithm *FPL*, the perturbed leader is sampled $K = \lceil 16t^2 \log(2\sqrt{t}) \rceil = \lceil \psi_t^{-4} \log(\psi_t^{-1}) \rceil$ times, and $a(K)$ is the number of times the leader happens to be expert i. By Hoeffding's inequality, the distribution of $\frac{a(K)}{K}$ is sharply peaked around its mean p_t^i:

$$\mathbf{P}\left[\frac{a(K)}{K} - p_t^i \geq \frac{\psi_t^2}{\sqrt{2}}\right] \leq \mathrm{e}^{-\psi_t^4 K} \quad \text{and} \quad \mathbf{P}\left[\frac{a(K)}{K} - p_t^i \leq -\frac{\psi_t^2}{\sqrt{2}}\right] \leq \mathrm{e}^{-\psi_t^4 K}.$$

Choosing $\hat{p}_t^i = \max\left\{\psi_t, \frac{a(K)}{K} - \frac{\psi_t^2}{\sqrt{2}}\right\}$ therefore implies that $\hat{p}_t^i \leq p_t^i$ with probability at least $1 - \psi_t$ (recall the assumption $p_t^i \geq \psi_t$). Hence the possibility of overestimate $\hat{p}_t^i > p_t^i$ causes another additional regret of $\psi_t B_t$, which proves (i).

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

$$\mathbf{P}\left[\frac{a(K)}{K} - p_t^i \leq -\frac{\sqrt{m}\psi_t^2}{\sqrt{2}}\right] \leq \mathrm{e}^{-m\psi_t^4 K} \leq \psi_t^m \tag{4}$$

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

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

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

Claim 8. *If $m \leq M$, then*

$$\frac{c_t^i}{p_t^i - (\sqrt{m}+1)\psi_t^2/\sqrt{2}} \leq \frac{c_t^i}{p_t^i} + \psi_t B_t(\sqrt{m}+1) \leq \frac{c_t^i}{p_t^i} + \psi_t B_t(m+1).$$

This follows by a simple algebraic manipulation. The claim implies (ii), because according to (4), $\hat{p}_t^i$ happens to be left of A_m with probability at most ψ_t^m.

Now, (iii) follows if we estimate the expectation over all A_m. We have just shown that for $\hat{p}_t^i \in A_m$, we have $\mathbf{E}\hat{c}_t^i \leq c_t^i + (m+1)\psi_t B_t$, and moreover this occurs with probability at most ψ_t^{m-1}. So, when passing back from the estimated to the true costs, this implies an upper bound on the additional regret of

$$\sum_{m=1}^{\infty} (m+1)\psi_t^m B_t \leq \frac{2\psi_t B_t}{1-\psi_t} + \frac{\psi_t^2 B_t}{(1-\psi_t)^2} \leq 6\psi_t B_t,$$

since $\psi_t \leq \frac{1}{2}$. For all other experts i not selected by *FPL*, we clearly have $\Delta_t[\hat{c}_t^i, c_t^i] = 0$, hence (iii) is proven. □

We now prove the step from *F** to *IF**, in terms of estimated costs.

Lemma 9. *Provided that $\eta_t B_t \leq 1$, we have*

$$\mathbf{E}\Delta_t[\hat{c}^{F*}, \hat{c}^{IF*}] \leq \eta_t |\{t \geq \tau\}| B_t^2 + 20\psi_t B_t \quad \textit{for all } t \geq 1.$$

Proof. Abbreviate $\pi_t^i = \mathbf{P}(I_t^{IF*} = i) = \mathbf{P}(J_t^{IF*} = i)$. Denote the exponential distribution by μ and integration with respect to $q^1 \dots q^n$ ($n = |\{t \geq \tau\}|$) without the ith coordinate by $\int \dots \mathrm{d}\mu(q^{\neq i})$. Moreover, for $x \in \mathbb{R}$, let $x^+ = \max\{x, 0\}$. Then, similarly to the proof of [14, Theorem 4],

$$\begin{aligned} p_t^i &= \int \int_{\max\limits_{j \neq i} \{\eta_t(\hat{c}^i_{<t} - \hat{c}^j_{<t}) + q^j + k^i - k^j\}}^{\infty} \mathrm{d}\mu(q^i)\, \mathrm{d}\mu(q^{\neq i}) = \int \mathrm{e}^{-(\max\limits_{j \neq i}\{\eta_t(\hat{c}^i_{<t} - \hat{c}^j_{<t}) + q^j + k^i - k^j\})^+} \mathrm{d}\mu(q^{\neq i}) \\ &\leq \int \mathrm{e}^{\frac{\eta_t B_t}{\hat{p}_t^i}}\, \mathrm{e}^{-(\max\limits_{j \neq i}\{\eta_t(\hat{c}^i_{<t} - \hat{c}^j_{<t}) + q^j + k^i - k^j\} + \frac{\eta_t B_t}{\hat{p}_t^i})^+} \mathrm{d}\mu(q^{\neq i}) \\ &\leq \mathrm{e}^{\frac{\eta_t B_t}{\hat{p}_t^i}} \int \mathrm{e}^{-(\max\limits_{j \neq i}\{\eta_t(\hat{c}^i_{1:t} - \hat{c}^j_{1:t}) + q^j + k^i - k^j\})^+} \mathrm{d}\mu(q^{\neq i}) = \mathrm{e}^{\frac{\eta_t B_t}{\hat{p}_t^i}} \pi_t^i. \end{aligned} \tag{5}$$

Hence, $\pi_t^i \geq p_t^i \mathrm{e}^{-\frac{\eta_t B_t}{\hat{p}_t^i}} \geq p_t^i \left(1 - \frac{\eta_t B_t}{\hat{p}_t^i}\right)$. From Lemma 6(ii), for each $m \geq 1$, we know that

$$\frac{c_t^i}{\hat{p}_t^i} \leq \frac{c_t^i}{p_t^i} + \psi_t B_t(\sqrt{m} + 1) \tag{6}$$

holds with high probability of at least $1 - \psi_t^m$. Also, simultaneously

$$\frac{\eta_t B_t}{\hat{p}_t^i} \leq \frac{\eta_t B_t}{p_t^i} + \psi_t(\sqrt{m} + 1) \tag{7}$$

holds because of $\eta_t B_t \leq 1$. Denote by $\mathbb{1}_{i=j}$ the indicator function that $i = j$ and recall Remark 7. Then, in the case that (6) and (7) hold, we have

$$\begin{aligned} \mathbf{E}\hat{c}_t^{F*} &= \sum_{i \in \{t \geq \tau\}} p_t^i \sum_{j \in \{t \geq \tau\}} p_t^j \mathbb{1}_{i=j} \frac{c_t^i}{\hat{p}_t^i} \leq \sum_{i \in \{t \geq \tau\}} p_t^i \sum_{j \in \{t \geq \tau\}} \left(\pi_t^j + p_t^j \frac{\eta_t B_t}{\hat{p}_t^i}\right) \mathbb{1}_{i=j} \frac{c_t^i}{\hat{p}_t^i} \\ &\leq \sum_{i \in \{t \geq \tau\}} p_t^i \sum_{j \in \{t \geq \tau\}} \pi_t^j \mathbb{1}_{i=j} \frac{c_t^i}{\hat{p}_t^i} + \sum_{i \in \{t \geq \tau\}} (p_t^i)^2 \left(\frac{\eta_t B_t}{p_t^i} + \psi_t(\sqrt{m} + 1)\right) \left(\frac{B_t}{p_t^i} + \psi_t B_t(\sqrt{m} + 1)\right) \\ &\leq \mathbf{E}\hat{c}_t^{IF*} + |\{t \geq \tau\}| \eta_t B_t^2 + 4\psi_t B_t(m + 1), \end{aligned}$$

where for the last estimate, we used $\eta_t B_t \leq 1$, $\psi_t \leq 1$, and $(\sqrt{m} + 1)^2 \leq 2(m + 1)$. As in the proof of Lemma 6(iii), we can bound the expectation over the sum from $m = 1 \dots \infty$, which shows the assertion. □

The following step from *IF** to any expert provides the last piece we need to complete the analysis in expectation.

Lemma 10. *Suppose that* $\sum_i \mathrm{e}^{-k^i} \leq 1$ *and* τ^i *depends monotonically on* k^i, *i.e.* $\tau^i \geq \tau^j$ *if and only if* $k^i \geq k^j$. *Assume decreasing learning rate* η_t. *For any* $t_0 \geq 1$, *all* $T \geq 1$ *and all experts* i,

$$\mathbf{E}\Delta_{t_0:T}[\hat{c}^{IF*}, \hat{c}^i] \leq \tfrac{k^i + 1}{\eta_T}.$$

Proof. This is a modification of the proof of [14, Theorem 2]. Without loss of generality, assume $t_0 = 1$. We will show that for fixed randomization q (and consequently fixed estimated costs $\hat{c}_t$),

$$\mathbf{E}\hat{c}_{1:T}^{IF*} \leq \min_{i \geq 1} \left\{ \hat{c}^i_{1:T} + \frac{k^i + 1}{\eta_T} \right\} \tag{8}$$

holds in expectation w.r.t. q_*. This implies the assertion. Recall that *IF** is charged cost according to the perturbation q_* and the selection J_t^{IF*}: throughout this proof, superscripts *IF** refer to J_t^{IF*}, e.g. $c_t^{IF*} = c_t^{J_t^{IF*}}$. Let $\eta_0 = \infty$ and

$$\lambda_t = \hat{c}_t + (k - q_*) \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right), \text{ which implies } \lambda_{1:t} = \hat{c}_{1:t} + \frac{k - q_*}{\eta_t}.$$

We use the abbreviated notation "$\min_{T\geq\tau}$" instead of "$\min_{i:T\geq\tau^i}$". Then, for all $T \geq 1$,

$$\sum_{t=1}^{T}\lambda_t^{\mathit{IF}*} \leq \min_{T\geq\tau}\lambda_{1:T}^i + \max_{T\geq\tau}\left\{\frac{q_*^i - k^i}{\eta_T}\right\} \tag{9}$$

can be shown by induction. It clearly holds for $T = 0$. For the induction step, we have to show

$$\begin{aligned}\min_{T\geq\tau}\lambda_{1:T}^i + \max_{T\geq\tau}\left\{\frac{q_*^i - k^i}{\eta_T}\right\} + \lambda_{T+1}^{\mathit{IF}*} &\leq \lambda_{1:T}^{J_{T+1}^{\mathit{IF}*}} + \max_{T+1\geq\tau}\left\{\frac{q_*^i - k^i}{\eta_{T+1}}\right\} + \lambda_{T+1}^{J_{T+1}^{\mathit{IF}*}} \\ &= \min_{T+1\geq\tau}\lambda_{1:T+1}^i + \max_{T+1\geq\tau}\left\{\frac{q_*^i - k^i}{\eta_{T+1}}\right\}.\end{aligned} \tag{10}$$

The inequality is obvious if $J_{T+1}^{\mathit{IF}*} \in \{T \geq \tau\}$. Otherwise, let

$$M = \arg\max\left\{q_*^i - k^i : i \in \{T \geq \tau\}\right\}.$$

Then

$$\begin{aligned}\min_{T\geq\tau}\lambda_{1:T}^i + \max_{T\geq\tau}\left\{\frac{q_*^i - k^i}{\eta_T}\right\} \ &\leq\ \lambda_{1:T}^M + \frac{q_*^M - k^M}{\eta_T} \ = \ \sum_{t=1}^{T}\hat{c}_t^M \leq \sum_{t=1}^{T}\frac{B_t}{\psi_t} \\ &= \sum_{t=1}^{T}\hat{c}_t^{J_{T+1}^{\mathit{IF}*}} \ \leq\ \lambda_{1:T}^{J_{T+1}^{\mathit{IF}*}} + \max_{T+1\geq\tau}\left\{\frac{q_*^i - k^i}{\eta_{T+1}}\right\}\end{aligned}$$

shows (10). Rearranging terms in (9), we see

$$\sum_{t=1}^{T}\hat{c}_t^{\mathit{IF}*} \leq \min_{T\geq\tau}\lambda_{1:T}^i + \max_{T\geq\tau^i}\left\{\frac{q_*^i - k^i}{\eta_T}\right\} + \sum_{t=1}^{T}(q_* - k)^{J_t^{\mathit{IF}*}}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right).$$

The assertion (8) then follows by taking expectations and using

$$\mathbf{E}\min_{T\geq\tau}\lambda_{1:T}^i \ \leq\ \min_{T\geq\tau}\left\{\hat{c}_{1:T}^i + \tfrac{k^i}{\eta_T} - \mathbf{E}\frac{q^i}{\eta_T}\right\} \leq \min_{i\geq 1}\left\{\hat{c}_{1:T}^i + \frac{k^i - 1}{\eta_T}\right\} \text{ and} \tag{11}$$

$$\mathbf{E}\sum_{t=1}^{T}(q - k)^{J_t^{\mathit{IF}*}}\left(\frac{1}{\eta_t} - \tfrac{1}{\eta_{t-1}}\right) \leq \mathbf{E}\max_{T\geq\tau}\left\{\frac{q^i - k^i}{\eta_T}\right\} \ \leq\ \frac{1}{\eta_T}. \tag{12}$$

The second inequality of (11) holds because τ^i depends monotonically on k^i, and $\mathbf{E}q^i = 1$, and maximality of $\hat{c}_{1:T}^i$ for $T < \tau_i$. The second inequality of (12) can be proven by a simple application of the union bound, see [14, Lemma 1]. □

We now combine the above results and derive an upper bound on the expected regret of *FPL* against an adaptive adversary.

Theorem 11. *Suppose* $\sum_i \mathrm{e}^{-k^i} \leq 1$. *Choose introduction times* $\tau^i = \lceil (w^i)^{-\frac{1}{\alpha}} \rceil$, *learning rate* $\eta_t = t^{-\frac{1}{2}-\varepsilon}$, *denominator threshold* $\psi_t = \frac{1}{2\sqrt{t}}$, *and instantaneous bounds on the costs* $B_t = t^\beta$. *Select* $\alpha, \beta \geq 0$ *such that* $\frac{\alpha}{2} + \beta = \varepsilon$. *Let* c_t *be some possibly adaptive assignment of cost vectors satisfying* $\|c_t\|_\infty \leq B_t$. *Then for each expert* i, *we have*

$$\mathbf{E}\Delta_{1:T}[c^{\mathit{FPL}}, c^i] \leq 2(w^i)^{-\frac{1}{\alpha}(1+\beta)} + T^{\frac{1}{2}+\varepsilon}(k^i + 31).$$

Proof. Set $t_0 = \tau^i$. For $t < t_0$, the cost of *FPL*, hence also the regret, is bounded by $\sum_{t=1}^{\tau^i - 1} B_t \leq 2(w^i)^{-\frac{1}{\alpha}(1+\beta)}$. The rest for $t \geq t_0$ follows by summing up all regret bounds in (this is the correct order) Lemmas 6(i), 9, 10, and 6(iii), observing that $|\{t \geq \tau\}| \leq t^\alpha$. □

The leading constant 31 of the (lower order) term $T^{\frac{1}{2}+\varepsilon}$ is not sharp; it stems from the necessity of taking into account the cost estimation error several times during the analysis.

In order to complete the proof of Theorem 1, we need a high probability bound. We will obtain this from Azuma's inequality. However, in order to apply Azuma's inequality, we first need to slightly generalize Theorem 11 to conditional expectations. Let $\mathcal{A}_t$ be the sigma algebra generated by all randomness up to time t, and consider

$$\mathbf{E}(X|\mathcal{A}_{t-1}),$$

which is the conditional expectation of some random variable X w.r.t. the sigma algebra $\mathcal{A}_{t-1}$. The sequence $(\mathcal{A}_t)_{t\geq 1}$ obviously is a filtration of sigma algebras.

Corollary 12. *Under the conditions of Theorem 11, we have*

$$\sum_{t=1}^{T}\mathbf{E}\big(\varDelta_t[c^{\mathit{FPL}},c^i]\big|\mathcal{A}_{t-1}\big)\leq 2(w^i)^{-\frac{1}{\alpha}(1+\beta)}+T^{\frac{1}{2}+\varepsilon}(k^i+31).$$

Proof. We need to check that the Lemmata shown so far also hold in conditional expectation. This is easy to see in all cases. Note in particular that Lemma 10 has been proven for *any fixed* randomness, where the expectation was taken only over the charging perturbation q_*. □

Now, the required high probability bound can be obtained from Azuma's inequality.

Lemma 13. *For each $T\geq 1$ and $\delta\in(0,1)$, with probability at least $1-\delta$ we have*

$$\varDelta_{1:T}[c^{\mathit{FPL}},c^i]\leq\sum_{t=1}^{T}\mathbf{E}\big(\varDelta_t[c^{\mathit{FPL}},c^i]\big|\mathcal{A}_{t-1}\big)+\sqrt{\left(2\ln\frac{4}{\delta}\right)\sum_{t=1}^{T}B_t^2}.$$

Proof. It is straightforward that the sequence of random variables

$$X_T=\varDelta_{1:T}[c^{\mathit{FPL}},c^i]-\sum_{t=1}^{T}\mathbf{E}\big(\varDelta_t[c^{\mathit{FPL}},c^i]\big|\mathcal{A}_{t-1}\big)$$

is a martingale w.r.t. the filtration $\mathcal{A}_t$ generated by the past randomness, since

$$\mathbf{E}(X_t|\mathcal{A}_{t-1})=X_{t-1}+\mathbf{E}\big(\varDelta_t[c^{\mathit{FPL}},c^i]|\mathcal{A}_{t-1}\big)-\mathbf{E}\big(\varDelta_t[c^{\mathit{FPL}},c^i]\big|\mathcal{A}_{t-1}\big)=X_{t-1}$$

holds. Its differences are bounded: $|X_t-X_{t-1}|\leq B_t$. Hence, it follows from Azuma's inequality (see e.g. [21]) that the probability that X_T exceeds some $\lambda>0$ is bounded by $2\exp\big(-\frac{\lambda^2}{2\sum_t B_t^2}\big)$. Requesting $\delta=2\exp\big(-\frac{\lambda^2}{2\sum_t B_t^2}\big)$ and solving for λ gives the assertion. □

We can now prove the following concretized version of Theorem 1:

Theorem 14. *Let $B_t\equiv 1$, i.e. costs are uniformly bounded. For given $\varepsilon>0$, choose introduction times $\tau^i=\lceil(w^i)^{-\frac{1}{2\varepsilon}}\rceil$ and learning rate $\eta_t=t^{-\frac{1}{2}-\varepsilon}$. Then both the expected and high probability regret of FPL relative to any arm i and at any time T can be bounded:*

$$\begin{aligned}\mathbf{E}\varDelta_{1:T}[c^{\mathit{FPL}},c^i]&=O\left(T^{(\frac{1}{2}+\varepsilon)}\log w^i+(w^i)^{-\frac{1}{2\varepsilon}}\right)\ \textit{and}\\ \varDelta_{1:T}[c^{\mathit{FPL}},c^i]&=O\left(T^{(\frac{1}{2}+\varepsilon)}\log w^i+(w^i)^{-\frac{1}{2\varepsilon}}\right)\ \textit{with probability}\\ &\qquad\qquad\textit{at least } 1-T^{-2}.\end{aligned}$$

Moreover, FPL is asymptotically optimal w.r.t. each expert, i.e. for all i,

$$\limsup_{T\to\infty}\frac{c_{1:T}^{\mathit{FPL}}-c_{1:T}^i}{T}\leq 0\quad\textit{almost surely.}$$

The asymptotic optimality is sometimes termed *Hannan-consistency*, in particular if the limit equals zero. We only show the upper bound.

Proof. Expected and high probability bounds follow from Theorem 11 and Corollary 12 combined with Lemma 13. Hence, only the asymptotic optimality remains to be shown. Since $\mathbf{P}\big[\frac{c^{FPL}_{1:T}-c^i_{1:T}}{T} > CT^{-\frac{1}{2}+\varepsilon}\big] \leq \frac{1}{T^2}$ holds for appropriate $C > 0$, this follows from the Borel–Cantelli Lemma. □

3. Reactive environments

The *FPL* algorithm considered so far performs well relative to any expert. In this section, we show an easy way to possibly improve the performance of some experts (and thus also of *FPL*) for a broad class of tasks, namely for *reactive environments*.

The reader is asked to think of broader learning tasks than just pulling the arm of a bandit repeatedly in this section, such as an agent which learns and acts in some environment. Also, the experts could be possibly complex procedures, performing some complicated computation and thereafter suggesting some decision or action. On the technical level, however, nothing changes: The learner repeatedly selects an expert and follows its advice.

As a motivating example, consider the repeated "prisoner's dilemma" against the tit-for-tat[4] strategy (this motivating example was also considered in [9]). If we use two strategies as experts, namely "always cooperate" and "always defect", then it is clear that always cooperating will have the best long-term reward. However, a standard expert advice or bandit learning algorithm will not discover this, since it compares only the costs in one step, which are always lower for the defecting expert (defecting is the "dominant action"). To put it differently, minimizing short-term regret is not at all a good idea here. E.g. always defecting has no regret, while for always cooperating the regret grows *linearly*. But this is only the case for short-term regret, i.e. if we restrict ourselves to time intervals of length one. Prisoner's dilemma is an example for a broader class of *reactive online decision problems* where evaluating the *long-term performance* renders some experts significantly stronger. In particular, there is an expert that is *optimal after t_0 time steps*, namely the cooperating expert.

Definition 15. (i) From the perspective of the learner, an *online decision problem* is just a game as defined in Section 2.2. However, we do not assume here that the goal of the adversary is necessarily maximizing the learner's regret.

(ii) We say an online decision problem to be *reactive* if the learner's actions impact on the adversary's behaviour in future. We are particularly interested in situations where the learner can benefit from well-chosen decisions in the past.

(iii) The *T-long-term performance* of an expert i is its cost $c^i_{T_0+1:T_0+T}$ for T time steps, starting from time T_0, if this expert i *plays all the time* from T_0 on, no other experts are allowed to interfere. We say that the *long-term performance* of the experts is evaluated, if for each T, from some time in the game on, each expert invoked is followed for T time steps or more.

(iv) The *hypothetical* cost or performance of an expert i *after t_0 time steps* is its long-term cost $c^i_{T_1+1:T_1+T}$, where it plays all the time from T_0 on, and the performance is evaluated from time $T_1 = T_0 + t_0$ on (so the initial t_0 steps are ignored). An expert is *strong or even optimal after t_0 steps* if its hypothetical cost after t_0 steps is low or even optimal, *starting from any state of the environment*. Here, the notions "low" and "optimal" depend on what is achievable for the respective environment.

For instance, in the prisoner's dilemma, where the opponent plays tit-for-tat, cooperating is optimal after two steps. If the learner chooses to defect in one round, the cooperative expert needs two contiguous rounds to recover, i.e. to get the opponent to cooperate again.

Recall from Section 2.2 that the adversarial (environment) is assumed to be *deterministic* throughout this work, which makes it possible to restrict to a fixed number t_0 of time steps. Also, we left the notions of strength and

[4] In the prisoner's dilemma, two players both decide independently if thy are *cooperating (C)* or *defecting (D)*. If both play C, they get both a small cost, if both play D, they get a large cost. However, if one plays C and one D, the cooperating player gets a very large loss and the defecting player no cost at all. Thus defecting is a *dominant* strategy. Tit-for-tat plays C in the first move and afterwards the opponent's respective preceding move.

optimality unspecified. Using more technical tools than we are willing to afford in this presentation, it is possible to treat stochastic environments, and to unambiguously define the optimal performance for a very broad class of environments (the "value"). Then, quantitative bounds for the regret minimization algorithms can be derived accordingly, which hold for a broad class of Markov decision processes (and even apply to POMDPs, provided that at least one of the experts suggests optimal actions). We leave the details to the interested reader.

The algorithm *reFPeL*. It is surprisingly easy to state an algorithm which performs well relative to the long-term performance and the even *hypothetical* performance of any expert after t_0 steps. We just have to make sure that the long-term performance of all experts is evaluated. We therefore define the algorithm *reFPeL*, *reactive follow the perturbed estimated leader*, as follows: Take the algorithm *FPeL* from Section 2.3 and rename the learner's scale from t to $\tilde{t}$. Then, in each time step $\tilde{t}$, give the control to a selected expert for *periods of increasing length* $B_{\tilde{t}}$. The costs at the learner's new time scale $\tilde{t}$ are defined as $\tilde{c}^i_{\tilde{t}} = \sum_{t=t(\tilde{t})}^{t(\tilde{t})+B_t-1} c^i_t$. (The points $t(\tilde{t})$ in basic time are defined recursively.) If the basic costs are uniformly bounded, $c^i_t \in [0, 1]$, the new learner's costs $\tilde{c}^i_t$ are then in $[0, B_t]$. The algorithm *reFPeL* has the following two performance guarantees, which imply Theorem 2 and Theorem 3 from the introduction.

Theorem 16 (*Performance Guarantee Relative to the* Actual *Costs*). *For given* $\varepsilon > 0$, *choose* $B_{\tilde{t}} = \lfloor \tilde{t}^{\frac{\varepsilon}{2}} \rfloor$, *introduction times* $\tau^i = \lceil (w^i)^{-\frac{1}{\varepsilon}} \rceil$, *and learning rate* $\eta_{\tilde{t}} = \tilde{t}^{-\frac{1}{2}-\varepsilon}$. *Then, both the expected regret* $\mathbf{E}\Delta_{1:T}[c^{\mathit{reFPeL}}, c^i]$ *and the high probability (probability at least* $1 - T^{-2}$*) regret* $\Delta_{1:T}[c^{\mathit{reFPeL}}, c^i]$ *of reFPeL relative to the* actual long-term performance *of any expert* i *and after any horizon* T, *are bounded by*

$$O\big(T^{(\frac{1}{2}+\varepsilon)} \log w^i + (w^i)^{-(\frac{1}{\varepsilon}+\frac{1}{2})}\big). \tag{13}$$

Also, reFPeL is asymptotically optimal: $\limsup_{T\to\infty} \frac{1}{T}(c^{\mathit{reFPeL}}_{1:T} - c^i_{1:T}) \le 0$ *almost surely for all experts* i.

Proof. A bound such as (13) in terms of $\tilde{T}$ instead of T follows from Theorem 11. Since $\tilde{T} \le T$, this immediately implies (13). □

Theorem 17 (*Performance Guarantee Relative to the* Hypothetical *Costs*). *Let* $B_{\tilde{t}} = \lfloor \tilde{t}^{\frac{1}{4}} \rfloor$, *introduction times* $\tau^i = \lceil (w^i)^{-\frac{1}{10}} \rceil$, *and learning rate* $\eta_{\tilde{t}} = \tilde{t}^{-\frac{2}{5}}$. *Then, for any* t_0, *the regret (expected and high probability) relative to the* hypothetical performance after t_0 steps *of any expert* i *is at most*

$$O\big(T^{\frac{4}{5}}(\log w^i + t_0) + (w^i)^{-12.5}\big), \tag{14}$$

at the horizon time T. *Hence, if an expert* i *is strong or even optimal after* t_0 *time steps, the same holds for reFPeL with the above regret* (14). *Also, reFPeL is asymptotically optimal relative to the hypothetical costs.*

Proof. On the learner's time scale $\tilde{t}$, we have the regret of expert i's actual performance w.r.t. its hypothetical performance is at most $\tilde{t} \cdot t_0$. Since t is of order $\tilde{t}^{\frac{5}{4}}$, on the original time scale t this regret is of order $t_0 \cdot t^{\frac{4}{5}}$. Selecting $\varepsilon = \frac{3}{10}$, $\alpha = \frac{1}{10}$, $\beta = \frac{1}{4}$ in Theorem 11 and observing $\tilde{T} \le T$ hence implies the assertion. □

Since we can handle countably infinite expert classes, we may specify a *universal* experts algorithm. To this aim, let expert i be derived from the ith (valid) program prog^i on some fixed universal Turing machine. The ith program can be well-defined, e.g. by representing programs as binary strings and lexicographically ordering them [12]. Before the expert is consulted, the relevant input, consisting e.g. of the complete history of observations, is written to the input tape of the corresponding program. If the program halts, an appropriate part of the output is interpreted as the expert's recommendation. E.g. if the decision is binary, then the first bit suffices. (If the program does not halt, we may go for well-definedness; just fill its output tape with zeros, see the next paragraph for the computability issue.) Each expert is assigned a prior weight by $w^i = 2^{-\mathrm{length}(\mathrm{prog}^i)}$, where $\mathrm{length}(\mathrm{prog}^i)$ is the length of the corresponding program and we assume the program tape to be binary. This construction parallels the definition of Solomonoff's *universal prior* [25].

Corollary 18. *If reFPeL is used together with a universal expert class as specified in the preceding paragraph, then it displays long-term performance asymptotically at least as well as* any computable strategy i. *The upper bound on the regret growth is exponential in the complexity* k^i *and proportional to* $t^{\frac{1}{2}+\varepsilon}$ *(against the actual performance) or* $t^{\frac{4}{5}}$ *(against the hypothetical performance).*

Note that our universal learner is not computable, since we cannot check if the computation of an expert halts. (This is just like other universal learners, e.g. AIξ [12].) On the other hand, if used with computable experts, the algorithm is computationally feasible (at each time t we need to consider only finitely many experts). Moreover, it is easy to impose an additional constraint on the computation time of each expert and abort the expert's computation after C_t operations on the Turing machine. We may choose some (possibly rapidly) growing function C_t, e.g. $C_t = 2^t$. The resulting learner is fully computable and has small regret with respect to all resource bounded strategies.

Acknowledgments

Many people contributed to this paper. Marcus Hutter had many of the key ideas that led to [13,14] and laid the foundation for the present work. Gabor Lugosi pointed me at the problem of the [15] analysis with adaptive adversary, and Tom Hayes and Varsha Dani pointed me at the general problem with the "regret in hindsight" notion with adaptive bandits. Thanks to Thomas Zeugmann and Jussi Kujala for helpful discussions. Thanks also to the anonymous referees for their careful reading and their valuable suggestions. I am especially grateful to one referee for pointing out several errors in a previous version of the paper, and for his help to improve the bounds. Financial support from JSPS 21st century COE programme C01 is gratefully acknowledged.

References

[1] C. Allenberg, Individual sequence prediction — upper bounds and application for complexity, in: COLT '99: Proceedings of the twelfth annual conference on Computational learning theory, ACM Press, 1999, pp. 233–242.

[2] P. Auer, N. Cesa-Bianchi, Y. Freund, 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), IEEE Computer Society Press, Los Alamitos, CA, 1995, pp. 322–331.

[3] P. Auer, N. Cesa-Bianchi, Y. Freund, R.E. Schapire, The nonstochastic multiarmed bandit problem, SIAM Journal on Computing 32 (1) (2002) 48–77.

[4] B. Awerbuch, R.D. Kleinberg, Competitive collaborative learning, in: Learning Theory, 18th Annual Conference on Learning Theory (COLT), 2005, pp. 233–248.

[5] D.A. Berry, B. Fristedt, Bandit Problems: Sequential Allocation of Experiments, Chapman and Hall, London, 1985.

[6] N. Cesa-Bianchi, G. Lugosi, G. Stoltz, Minimizing regret with label efficient prediction, in: 17th Annual Conference on Learning Theory (COLT), in: Lecture Notes in Computer Science, vol. 3120, Springer, 2004, pp. 77–92.

[7] N. Cesa-Bianchi, G. Lugosi, G. Stoltz, Regret minimization under partial monitoring, Technical Report, 2004.

[8] V. Dani, T. Hayes, How to beat the adaptive multi-armed bandit, Technical Report cs.DS/0602053,arXiv.org, February 2006.

[9] D. Pucci de Farias, N. Megiddo, How to combine expert (and novice) advice when actions impact the environment?, in: S Thrun, L. Saul, B. Schölkopf (Eds.), Advances in Neural Information Processing Systems 16, MIT Press, Cambridge, MA, 2004.

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

[11] J. Hannan, Approximation to Bayes risk in repeated plays, in: M. Dresher, A.W. Tucker, P. Wolfe (Eds.), Contributions to the Theory of Games 3, Princeton University Press, 1957, pp. 97–139.

[12] M. Hutter, Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability, Springer, Berlin, 2004.

[13] M. Hutter, J. Poland, Prediction with expert advice by following the perturbed leader for general weights, in: International Conference on Algorithmic Learning Theory (ALT), 2004, pp. 279–293.

[14] M. Hutter, J. Poland, Adaptive online prediction by following the perturbed leader, Journal of Machine Learning Research 6 (2005) 639–660.

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

[16] R.D. Kleinberg, Nearly tight bounds for the continuum-armed bandit problem, in: Advances in Neural Information Processing Systems 17, MIT Press, Cambridge, MA, 2005, pp. 697–704.

[17] R. Kleinberg, Anytime algorithms for multi-armed bandit problems, in: SODA '06: Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm, ACM Press, 2006, pp. 928–936.

[18] M. Li, P.M.B. Vitányi, An introduction to Kolmogorov complexity and its applications, 2nd edition, Springer, 1997.

[19] N. Littlestone, M.K. Warmuth, The weighted majority algorithm, in: 30th Annual Symposium on Foundations of Computer Science, IEEE, Research Triangle Park, North Carolina, 1989, pp. 256–261.

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

[21] R. Motwani, P. Raghavan, Randomized Algorithms, Cambridge University Press, Cambridge, England, 1995.

[22] J. Poland, FPL analysis for adaptive bandits, in: 3rd Symposium on Stochastic Algorithms, Foundations and Applications (SAGA), 2005, pp. 58–69.

[23] J. Poland, M. Hutter, Defensive universal learning with experts, in: 17th International Conference on Algorithmic Learning Theory (ALT), 2005, pp. 356–370.

[24] J. Poland, M. Hutter, Universal agents in repeated matrix games, 2006. Presented at 8th Workshop on Game Theoretic and Decision Theoretic Agents (in conjunction with AAMAS).

[25] R.J. Solomonoff, Complexity-based induction systems: Comparisons and convergence theorems, IEEE Trans. Inform. Theory 24 (1978) 422–432.

[26] V.G. Vovk, Aggregating strategies, in: Proc. Third Annual Workshop on Computational Learning Theory, ACM Press, Rochester, New York, 1990, pp. 371–383.