Defensive Universal Learning with Experts

Jan Poland[1] and Marcus Hutter[2]

[1] Grad. School of Inf. Sci. and Tech., Hokkaido University, Japan
jan@ist.hokudai.ac.jp, http://www-alg.ist.hokudai.ac.jp/~jan
[2] IDSIA, Galleria 2, CH-6928 Manno (TI), Switzerland
marcus@idsia.ch, http://www.idsia.ch/~marcus

TCS-TR-A-05-4 — July 2005 — IDSIA-15-05

Abstract. This paper shows how universal learning can be achieved with expert advice. To this aim, we specify an experts algorithm with the following characteristics: (a) it uses only feedback from the actions actually chosen (bandit setup), (b) it can be applied with countably infinite expert classes, and (c) it copes with losses that may grow in time appropriately slowly. We prove loss bounds against an adaptive adversary. From this, we obtain a master algorithm for "reactive" experts problems, which means that the master's actions may influence the behavior of the adversary. Our algorithm can significantly outperform standard experts algorithms on such problems. Finally, we combine it with a universal expert class. The resulting universal learner performs – in a certain sense – almost as well as any computable strategy, for any online decision problem. We also specify the (worst-case) convergence speed, which is very slow.
Keywords. Prediction with expert advice, responsive environments, partial observation game, bandits, universal learning, asymptotic optimality.

1 Introduction

Expert advice has become a well-established paradigm of machine learning in the last decade, in particular for prediction. It is very appealing from a theoretical point of view, as performance guarantees usually hold in the worst case, without any (statistical) assumption on the data. Such assumptions are generally required for other statistical learning methods, often however not resulting in stronger guarantees.

Using expert advice in the standard way seems a rather bad idea in some cases where the decisions of the learner or master algorithm influence the behavior of the environment or adversary. One example is the repeated prisoner's dilemma when the opponent plays "tit for tat" (see Section 4). This was noted and resolved by [1], who introduced a "strategic expert algorithm" for so-called reactive environments. Their algorithm works with a finite class of experts and attains asymptotically optimal behavior. No convergence speed is asserted, and the analysis is quite different from that of standard experts algorithms.

In this paper, we show how the more general task with a countably infinite expert class can be accomplished, building on standard experts algorithms, and simultaneously also bounding the convergence rate ($t^{-\frac{1}{10}}$, which can be actually improved to $t^{-\frac{1}{3}+\varepsilon}$). To this aim, we will combine techniques from [2–5] and obtain a master algorithm which performs well on loss functions that may *increase in time*. Then this is applied to (possibly) reactive problems by yielding the control to the selected expert for an increasing period of time steps. Using a universal expert class defined by the countable set of all programs on some fixed universal Turing machine, we obtain an algorithm which is in a sense asymptotically optimal with respect to *any* computable strategy. An easy additional construction guarantees that our algorithm is computable, in contrast to other universal approaches which are non-computable [6]. To our knowledge, we also propose the first algorithm for non-stochastic bandit problems with countably many arms.

The paper is structured as follows. Section 2 introduces the problem setup, the notation, and the algorithm. In Sections 3, we give the (worst-case) analysis of the master algorithm. The implications to active experts problems and a universal master algorithms are given in Section 4. We discuss our results in Section 5.

2 The master algorithm

Setup. We are acting in an online decision problem. "We" is here an abbreviation for the master algorithm which is to be designed. An "online decision problem" is to be understood in a very general sense, it is just a sequence of decisions each of which results in some loss. This could be e.g. a prediction task, a repeated game, etc. In each round, that is at each time step t, we have access to the recommendations of countably infinitely many "experts" or strategies. (For simplicity, we restrict our notation to a countably infinite expert class, all results also hold for finite classes.) We do not specify what exactly a "recommendation" is – we just follow the advice of one expert. *Before* we reveal our move, the adversary has to assign losses $\ell_t^i \geq 0$ to *all* experts i. There is an upper bound $B_t \geq \|\ell_t\|_\infty$ on the maximum loss the adversary may use. This quantity may depend on t and is not controlled by the adversary. After the move, only the loss of the selected expert i is revealed. Our goal is to perform nearly as well as the best available strategy (expert) in terms of cumulative loss, after any number T of time steps which is not known in advance. The difference between our loss and the loss of some expert is also termed *regret*. We consider the general case of an *adaptive* adversary, which may assign losses depending on our past decisions.

If there are only finitely many experts or strategies, then it is common to give no prior preferences to any of them. Formally, this is realized by defining uniform *prior weights* $w^i = \frac{1}{n}$ for each expert i. This is not possible for countably infinite

For $t = 1, 2, 3, \ldots$
  set $\hat{\ell}_t^i = \hat{B}_t$ for $i \notin \{t \geq \tau\}$ (see (2))
  sample $r_t \in \{0, 1\}$ independently s.t. $P[r_t = 1] = \gamma_t$
  If $r_t = 0$ Then
    invoke *FPL*(t) and play its decision
    set $\hat{\ell}_t^i = 0$ for $i \in \{t \geq \tau\}$
  Else
    sample $I_t^{FoE} := u_t$ "uniformly", see (1), and play $I := I_t^{FoE}$
    set $\hat{\ell}_t^I = \ell_t^I / (u_t^I \gamma_t)$ and $\hat{\ell}_t^i = 0$ for $i \in \{t \geq \tau\} \setminus \{I\}$

Fig. 1. The algorithm *FoE*. The exploration rate γ_t will be specified in Corollary 8.

Sample $q_t^i \overset{d.}{\sim} Exp$ (i.e. $\mathbf{P}(q_t^i \geq x) = e^{-x}$ for $x \geq 0$) indep. $\forall i \in \{t \geq \tau\}$
select and play $I_t^{FPL} = \arg\min_{i: t \geq \tau} \{\eta_t \hat{\ell}_{<t}^i + k^i - q_t^i\}$

Fig. 2. The algorithm *FPL*. The learning rate η_t will be specified in Corollary 8.

expert classes, as there is no uniform distribution on the natural numbers. In this case, we need some non-uniform prior $(w^i)_{i \in \mathbb{N}}$ and require $w^i > 0$ for all experts i and $\sum_i w^i \leq 1$. We also define the complexity of expert i as $k^i = -\ln w^i$. This quantity is important since in the full observation game (i.e. after our decision we get to know the losses of *all* experts), the regret can usually be bounded by some function of the best expert's complexity.

Our algorithm "Follow or Explore" (*FoE*, specified in Fig. 1) builds on McMahan and Blum's online geometric optimization algorithm [4]. It is a bandit version of a "Follow the Perturbed Leader" experts algorithm. This approach to online prediction and playing repeated games has been pioneered by Hannan [2]. For the full observation game and uniform prior, [3] gave a very elegant analysis which is clearly different from the standard analysis of exponential weighting schemes. It has one advantage over other aggregating algorithms such as exponential weighting schemes: the analysis is not complicated if the learning rate is dynamic rather than fixed in advance. A dynamic learning rate is necessary if there is no target time T known in advance. For non-uniform prior, an analysis was given in [5]. The following issues are important for *FoE*'s design.

Exploration. Since we are playing the bandit game (as opposed to the full information game), we need to explore sufficiently [7, 4]. At each time step t, we decide randomly according to some exploration rate $\gamma_t \in (0, 1)$ whether to explore or not. If so, we would like to choose an expert according to the prior distribution. There is a caveat: In order to make the analysis go through, we have to assure that we are working with *unbiased* estimates of the losses. This is achieved by dividing the observed loss by the probability of choosing the expert. But this quantity could become arbitrarily large if we admit arbitrarily small weights. We address this problem by *finitizing* the expert pool at each time t.

For each expert i, we define an *entering time* τ^i, that is, expert i is active only for $t \geq \tau^i$. We denote the set of active experts at time t by $\{t \geq \tau\} = \{i : t \geq \tau^i\}$. For exploration, the prior is then replaced by the finitized prior distribution u_t,

$$\mathbf{P}(u_t = i) = \frac{w^i \mathbb{I}_{t \geq \tau^i}}{\sum_j w^j \mathbb{I}_{t \geq \tau^j}}. \tag{1}$$

Consequently, the maximum unbiasedly estimated instantaneous loss is (note that the exploration probability also scales with the exploration rate γ_t)

$$\hat{B}_t = \frac{B_t}{\gamma_t \min\{w^i : t \geq \tau^i\}}. \tag{2}$$

It is convenient for the analysis to assign estimated loss of $\hat{B}_t$ to all currently inactive experts. Observe finally that in this way, our master algorithm *FoE* always deals with a finite expert class and is thus computable.

Follow the perturbed leader (*FPL*, specified in Fig. 2) is invoked if *FoE* does not explore. Just following the "leader" (the best expert so far) may not be a good strategy [3]. Instead we subtract an exponentially distributed perturbation q_t from the current score (the complexity penalized past loss) of the experts. An important detail of the *FPL* subroutine is the *learning rate* $\eta_t > 0$, which should be adaptive if the total number of steps T is not known in advance. Please see e.g. [3, 5] for more details. Also the variant of *FPL* we use (specified in Fig. 2) works on the finitized expert pool.

Note that each time randomness is used, it is assumed to be *independent* of the past randomness. Performance is evaluated in terms of true or estimated cumulative loss, this is specified in the notation. E.g. for the true loss of *FPL* up to and including time T we write $\ell^{FPL}_{1:T}$, while the estimated loss of *FoE* and not including time T is $\hat{\ell}^{FoE}_{<T}$.

3 Analysis on the master's time scale

The following analysis uses McMahan and Blum's trick [4] in order to prove bounds against adaptive adversary. With a different argument, it is possible to circumvent Lemma 6, thus achieving better bounds [8]. This will be briefly discussed in the last section.

Let $B_t \geq 0$ be some sequence of upper bounds on the instantaneous losses, $\gamma_t \in (0,1)$ be a sequence of exploration rates, and $\eta_t > 0$ be a *decreasing* sequence of learning rates. The analysis proceeds according to the following diagram (where L is an informal abbreviation for the loss and always refers to cumulative loss, but sometimes additionally to instantaneous loss).

$$L^{FoE} \lesssim \mathbf{E}L^{FoE} \lesssim \mathbf{E}L^{FPL} \lesssim \mathbf{E}\hat{L}^{FPL} \lesssim \mathbf{E}\hat{L}^{IFPL} \lesssim \mathbf{E}\hat{L}^{best} \lesssim L^{best} \tag{3}$$

Each " $\lesssim$ " means that we bound the quantity on the left by the quantity on the right plus some additive term. The first and the last expressions are the losses of the *FoE* algorithm and the best expert, respectively. The intermediate quantities

Sample $q_t^i \stackrel{d.}{\sim} Exp$ independently for all $i \in \{t \geq \tau\}$ select and play $I_t^{FPL} = \arg\min_{i:t\geq\tau}\{\eta_t \hat{\ell}_{1:t}^i + k^i - q_t^i\}$

Fig. 3. The algorithm *IFPL*. The learning rate η_t will be specified in Corollary 8.

belong to different algorithms, namely *FoE*, *FPL*, and a third one called *IFPL* for "infeasible" FPL [3]. *IFPL*, as specified in Fig. 3, is the same as *FPL* except that it has access to an oracle providing the current estimated loss vector $\hat{\ell}_t$ (hence infeasible). Then it assigns scores of $\eta_t \hat{\ell}_{1:t}^i + k^i - q_t^i$ instead of $\eta_t \hat{\ell}_{<t}^i + k^i - q_t^i$.

The randomization of *FoE* and *FPL* gives rise to two filtrations of σ-algebras. By $\mathcal{A}_t$ we denote the σ-algebra generated by the *FoE*'s randomness up to time t, meaning *only* the random variables $\{u_{1:t}, r_{1:t}\}$. Then $(\mathcal{A}_t)_{t\geq 0}$ is a filtration ($\mathcal{A}_0$ is the trivial σ-algebra). We may also write $\mathcal{A} = \bigcup_{t\geq 0} \mathcal{A}_t$. Similarly, $\mathcal{B}_t$ is the σ-algebra generated by the *FoE*'s *and FPL*'s randomness up to time t (i.e. $\mathcal{B}_t \widehat{=} \{u_{1:t}, r_{1:t}, q_{1:t}\}$). Then clearly $\mathcal{A}_t \subset \mathcal{B}_t$ for each t.

The reader should think of the expectations in (3) as of both ordinary and *conditional* expectations. Conditional expectations are mostly with respect to *FoE*'s past randomness $\mathcal{A}_{t-1}$. These conditional expectations of some random variable X are abbreviated by

$$\mathbf{E}_t[X] := \mathbf{E}[X|\mathcal{A}_{t-1}].$$

Then $\mathbf{E}_t[X]$ is an $\mathcal{A}_{t-1}$-measurable random variable, meaning that its value is determined for fixed past randomness $\mathcal{A}_{t-1}$. Note in particular that the estimated loss vectors $\hat{\ell}_t^i$ are random vectors which depend on *FoE*'s randomness $\mathcal{A}_t$ up to time t. In this way, *FoE*'s (and *FPL*'s and *IFPL*'s) actions depend on *FoE*'s past randomness. Note, however, that they do not depend on *FPL*'s randomness $q_{1:t}$.

We now start with proving the diagram (3). In order to understand the analysis, it is important to consider each intermediate algorithm as a stand-alone procedure which is actually executed (with an oracle if necessary) on the specified inputs (e.g. on the estimated losses) and has the asserted performance guarantees (e.g. again on the estimated losses).

Lemma 1. $\left[L^{FoE} \lesssim \mathbf{E}L^{FoE}\right]$ *For each $T \geq 1$ and $\delta_T \in (0,1)$, with probability at least $1 - \frac{\delta_T}{2}$ we have*

$$\ell_{1:T}^{FoE} \leq \sum_{t=1}^{t} \mathbf{E}_t \ell_t^{FoE} + \sqrt{(2\ln\tfrac{4}{\delta_T})\textstyle\sum_{t=1}^T B_t^2}.$$

Proof. The sequence of random variables $X_T = \sum_{t=1}^T \left[\ell_t^{FoE} - \mathbf{E}_t \ell_t^{FoE}\right]$ is a martingale with respect to the filtration $\mathcal{B}_t$ (not $\mathcal{A}_t$!). In order to see this, observe $\mathbf{E}[\ell_T^{FoE}|\mathcal{B}_{T-1}] = \mathbf{E}\big(\mathbf{E}[\ell_T^{FoE}|\mathcal{A}_{T-1}]\big|\mathcal{B}_{T-1}\big)$ and $\mathbf{E}[\ell_t^{FoE}|\mathcal{B}_{T-1}] = \ell_t^{FoE}$ for $t < T$, which implies

$$\mathbf{E}[X_T|\mathcal{B}_{T-1}] = \sum\nolimits_{t=1}^{T} \big(\mathbf{E}[\ell_t^{FoE}|\mathcal{B}_{T-1}] - \mathbf{E}\big[\mathbf{E}[\ell_t^{FoE}|\mathcal{A}_{t-1}]\big|\mathcal{B}_{T-1}\big]\big)$$
$$= \sum\nolimits_{t=1}^{T-1} \big(\ell_t^{FoE} - \mathbf{E}[\ell_t^{FoE}|\mathcal{A}_{t-1}]\big) = X_{T-1}.$$

Its differences are bounded: $|X_t - X_{t-1}| \leq B_t$. Hence, it follows from Azuma's inequality (see e.g. [9]) that the probability that X_T exceeds some $\lambda > 0$ is bounded by $p = 2\exp\big(-\frac{\lambda^2}{2\sum_t B_t^2}\big)$. Requesting $\frac{\delta_T}{2} = p$ and solving for λ gives the assertion. □

Lemma 2. $\big[\mathbf{E}\ell^{FoE} \lesssim \mathbf{E}\ell^{FPL}\big]$ *$\mathbf{E}_t\ell_t^{FoE} \leq (1-\gamma_t)\mathbf{E}_t\ell_t^{FPL} + \gamma_t B_t$ holds $\forall t \geq 1$.*

This follows immediately from the specification of *FoE*. Clearly, a corresponding assertion for the ordinary expectations holds by just taking expectations on both sides. This is the case for all subsequent lemmas, except for Lemma 6.

The next lemma relating $\mathbf{E}L^{FPL}$ and $\mathbf{E}\hat{L}^{FPL}$ is technical but intuitively clear. It states that in expectation, the real loss suffered by *FPL* is the same as the estimated loss. This is simply because the loss estimate is unbiased. A combination of this and the previous lemma was shown in [4]. Note that $\hat{\ell}_t^{FPL}$ is the loss $\hat{\ell}_t^I$ estimated by *FoE*, but for the expert $I = I_t^{FPL}$ chosen by *FPL*.

Lemma 3. $\big[\mathbf{E}L^{FPL} \lesssim \mathbf{E}\hat{L}^{FPL}\big]$ *For each $t \geq 1$, we have $\mathbf{E}_t\ell_t^{FPL} = \mathbf{E}_t\hat{\ell}_t^{FPL}$.*

Proof. Let $f_t^i = \mathbf{P}[I_t^{FPL} = i|\mathcal{A}_{t-1}]$ be the probability distribution over actions i which *FPL* uses at time t, depending on the past randomness $\mathcal{A}_{t-1}$. Let u_t be the finitized prior distribution (1) at time t. Then

$$\mathbf{E}_t[\hat{\ell}_t^{FPL}](1-\gamma_t)\cdot 0 + \gamma_t \sum_{i=1}^{\infty} f_t^i[(1-u_t^i)\cdot 0 + u_t^i\hat{\ell}_t^i|_{r_t=1\wedge I_t^{FoE}=i}] \sum_{i=1}^{\infty} f_t^i\ell_t^i = \mathbf{E}_t[\ell_t^{FPL}],$$

where $\hat{\ell}_t^i|_{r_t=1\wedge I_t^{FoE}=i} = \ell_t^i/(u_t^i\gamma_t)$ is the estimated loss under the condition that *FoE* decided to explore ($r_t = 1$) and chose action $I_t^{FoE} = i$. □

The following lemma relates the losses of *FPL* and *IFPL*. It is proven in [3] and [5]. We give the full proof, since it is the only step in the analysis where we have to be careful with the upper loss bound B_t. Let $\hat{B}_t$ be the upper bound on the estimated loss (2). (We remark that also for *weighted averaging forecasters*, losses which grow sufficiently slowly do not cause any problem in the analysis. In this way, it is straightforward to modify the algorithm by Auer et al. [10] for reactive tasks with a finite expert class.)

Lemma 4. $\big[\mathbf{E}\hat{L}^{FPL} \lesssim \mathbf{E}\hat{L}^{IFPL}\big]$ *For all $t \geq 1$, $\mathbf{E}_t\hat{\ell}_t^{FPL} \leq \mathbf{E}_t\hat{\ell}_t^{IFPL} + \gamma_t\eta_t\hat{B}_t^2$ holds.*

Proof. If $r_t = 0$, $\hat{\ell}_t = 0$ and thus $\hat{\ell}_t^{FPL} = \hat{\ell}_t^{IFPL}$ holds. This happens with probability $1-\gamma_t$. Otherwise we have

$$\mathbf{E}_t\hat{\ell}_t^{FPL} = \sum_{i=1}^{\infty} \int \mathbb{I}_{I_t^{FPL}=i}\hat{\ell}_t^i d\mu(x), \tag{4}$$

where μ denotes the (exponential) distribution of the perturbations, i.e. $x_i := q_t^i$ and density $\mu(x) := \mathrm{e}^{-\|x\|_\infty}$. The idea is now that if action i was selected by *FPL*, it is – because of the exponentially distributed perturbation – with high probability also selected by *IFPL*. Formally, we write $u^+ = \max(u,0)$ for $u \in \mathbb{R}$, abbreviate $\lambda = \hat{\ell}_{<t} + k/\eta_t$, and denote by $\int \ldots d\mu(x_{\neq i})$ the integration leaving out the ith action. Then, using $\eta_t\lambda_i - x_i \leq \eta_t\lambda_j - x_j \ \forall j$ if $I_t^{FPL} = i$ in the first equation, and $\hat{B}_t \geq \hat{\ell}_t^i - \hat{\ell}_t^j$ in the last line, we get

$$\begin{aligned}
\int \mathbb{1}_{I_t^{FPL}=i}\hat{\ell}_t^i d\mu(x) &= \int \int\limits_{x_i \geq \max_{j\neq i}\{\eta_t(\lambda_i-\lambda_j)+x_j\}} \hat{\ell}_t^i d\mu(x_i) d\mu(x_{\neq i}) = \int \hat{\ell}_t^i \,\mathrm{e}^{-(\max_{j\neq i}\{\eta_t(\lambda_i-\lambda_j)+x_j\})^+} d\mu(x_{\neq i}) \\
&\leq \int \hat{\ell}_t^i \,\mathrm{e}^{\eta_t\hat{B}_t}\mathrm{e}^{-(\max_{j\neq i}\{\eta_t(\lambda_i-\lambda_j)+x_j\}+\eta_t\hat{B}_t)^+} d\mu(x_{\neq i}) \\
&\leq \mathrm{e}^{\eta_t\hat{B}_t} \int \hat{\ell}_t^i \,\mathrm{e}^{-(\max_{j\neq i}\{\eta_t(\lambda_i+\hat{\ell}_t^i-\lambda_j-\hat{\ell}_t^j)+x_j\})^+} d\mu(x_{\neq i}) \\
&= \mathrm{e}^{\eta_t\hat{B}_t} \int \mathbb{1}_{I_t^{IFPL}=i}\hat{\ell}_t^i d\mu(x).
\end{aligned}$$

Summing over i and using the analog of (4) for *IFPL*, we see that if $r_t = 1$, then $\mathbf{E}_t\hat{\ell}_t^{FPL} \leq \mathrm{e}^{\eta_t\hat{B}_t}\mathbf{E}_t\hat{\ell}_t^{IFPL}$ holds. Thus $\mathbf{E}_t\hat{\ell}_t^{IFPL} \geq \mathrm{e}^{-\eta_t\hat{B}_t}\mathbf{E}_t\hat{\ell}_t^{FPL} \geq (1-\eta_t\hat{B}_t)\mathbf{E}_t\hat{\ell}_t^{FPL} \geq \mathbf{E}_t\hat{\ell}_t^{FPL} - \eta_t\hat{B}_t^2$. The assertion now follows by taking expectations w.r.t. r_t. □

The next lemma relates the losses of *IFPL* and the best action in hindsight. For an oblivious adversary (which means that the adversary's decisions do not depend on our past actions), the proof is quite simple [3]. An additional step is necessary for an adaptive adversary [11].

Lemma 5. $\left[\mathbf{E}\hat{L}^{IFPL} \lesssim \mathbf{E}\hat{L}^{best}\right]$ *Assume 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 all* $T \geq 1$ *and all* $i \geq 1$*,*

$$\sum_{t=1}^{T} \mathbf{E}_t\hat{\ell}_t^{IFPL} \leq \hat{\ell}_{1:T}^i + \tfrac{k^i+1}{\eta_T}.$$

Proof. This is a modification of the corresponding proofs in [3] and [5]. We may fix the randomization $\mathcal{A}$ and suppress it in the notation. Then we only need to show

$$\mathbf{E}\hat{\ell}_{1:T}^{IFPL} \leq \min_{i\geq 1}\{\hat{\ell}_{1:T}^i + \tfrac{k^i+1}{\eta_T}\}, \tag{5}$$

where the expectation is with respect to *IFPL*'s randomness $q_{1:T}$.

Assume first that the adversary is oblivious. We define an algorithm A as a variant of *IFPL* which samples only one perturbation vector q in the beginning and uses this in each time step, i.e. $q_t \equiv q$. Since the adversary is oblivious, A is equivalent to *IFPL* in terms of expected performance. This is all we need to show (5). Let $\eta_0 = \infty$ and $\lambda_t = \hat{\ell}_t + (k-q)\big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\big)$, then $\lambda_{1:t} = \hat{\ell}_{1:t} + \frac{k-q}{\eta_t}$.

Recall $\{t \geq \tau\} = \{i : t \geq \tau^i\}$. We argue by induction that for all $T \geq 1$,

$$\sum_{t=1}^{T} \lambda_t^A \leq \min_{T \geq \tau} \lambda_{1:T}^i + \max_{T \geq \tau} \left\{ \tfrac{q^i - k^i}{\eta_T} \right\}. \tag{6}$$

This clearly holds for $T = 0$. For the induction step, we have to show

$$\begin{aligned} \min_{T \geq \tau} \lambda_{1:T}^i + \max_{T \geq \tau} \left\{ \tfrac{q^i - k^i}{\eta_T} \right\} + \lambda_{T+1}^A &\leq \lambda_{1:T}^{I_{T+1}^A} + \max_{T+1 \geq \tau} \left\{ \tfrac{q^i - k^i}{\eta_{T+1}} \right\} + \lambda_{T+1}^{I_{T+1}^A} \\ &= \min_{T+1 \geq \tau} \lambda_{1:T+1}^i + \max_{T+1 \geq \tau} \left\{ \tfrac{q^i - k^i}{\eta_{T+1}} \right\}. \end{aligned} \tag{7}$$

The inequality is obvious if $I_{T+1}^A \in \{T \geq \tau\}$. Otherwise, let $J = \arg\max \big\{ q^i - k^i : i \in \{T \geq \tau\} \big\}$. Then

$$\begin{aligned} \min_{T \geq \tau} \lambda_{1:T}^i + \max_{T \geq \tau} \left\{ \tfrac{q^i - k^i}{\eta_T} \right\} \leq \lambda_{1:T}^J + \tfrac{q^J - k^J}{\eta_T} &= \sum_{t=1}^{T} \hat{\ell}_t^J \leq \sum_{t=1}^{T} \hat{B}_t \\ &= \textstyle\sum_{t=1}^{T} \hat{\ell}_t^{I_{T+1}^A} \leq \lambda_{1:T}^{I_{T+1}^A} + \max_{T+1 \geq \tau} \left\{ \tfrac{q^i - k^i}{\eta_{T+1}} \right\} \end{aligned}$$

shows (7). Rearranging terms in (6), we see

$$\sum_{t=1}^{T} \hat{\ell}_t^A \leq \min_{T \geq \tau} \lambda_{1:T}^i + \max_{T \geq \tau^i} \left\{ \tfrac{q^i - k^i}{\eta_T} \right\} + \sum_{t=1}^{T} (q-k)^{I_t^A} \left(\tfrac{1}{\eta_t} - \tfrac{1}{\eta_{t-1}} \right)$$

The assertion (5) – still for oblivious adversary and $q_t \equiv q$ – then follows by taking expectations and using

$$\mathbf{E} \min_{T \geq \tau} \lambda_{1:T}^i \;\leq\; \min_{T \geq \tau} \{ \hat{\ell}_{1:T}^i + \tfrac{k^i}{\eta_T} - \mathbf{E} \tfrac{q^i}{\eta_T} \} \leq \min_{i \geq 1} \{ \hat{\ell}_{1:T}^i + \tfrac{k^i - 1}{\eta_T} \} \text{ and} \tag{8}$$

$$\mathbf{E} \sum_{t=1}^{T} (q-k)^{I_t^A} \left(\tfrac{1}{\eta_t} - \tfrac{1}{\eta_{t-1}} \right) \leq \mathbf{E} \max_{T \geq \tau} \left\{ \tfrac{q^i - k^i}{\eta_T} \right\} \;\leq\; \tfrac{1}{\eta_T}. \tag{9}$$

The second inequality of (8) holds because τ^i depends monotonically on k^i, and $\mathbf{E} q^i = 1$, and maximality of $\hat{\ell}_{1:T}^i$ for $T < \tau_i$. The second inequality of (9) can be proven by a simple application of the union bound, see e.g. [5, Lem.1].

Sampling the perturbations q_t independently is equivalent under expectation to sampling q only once. So assume that q_t are sampled independently, i.e. that *IFPL* is played against an oblivious adversary: (5) remains valid. In the last step, we argue that then (5) also holds for an *adaptive* adversary. This is true because the future actions of *IFPL* do not depend on its past actions, and therefore the adversary cannot gain from deciding after having seen *IFPL*'s decisions. This argument can be made formal, as shown in [11, Lemma 12]. (Note the subtlety that the future actions of *FoE* would depend on its past actions.) □

Finally, we give a relation between the estimated and true losses (adapted from [4]).

Lemma 6. $\left[\mathbf{E}\hat{L}^{best} \lesssim L^{best}\right]$ *For each* $T \geq 1$, $\delta_T \in (0,1)$, *and* $i \geq 1$, *we have*

(*i*) $\hat{\ell}^i_{1:T} \leq \ell^i_{1:T} + \sqrt{(2\ln\frac{4}{\delta_T})\sum_{t=1}^T \hat{B}_t^2} + \sum_{t=1}^{\tau^i-1} \hat{B}_t$ *w.p.* $1-\frac{\delta_T}{2}$ *and hence*

(*ii*) $\mathbf{E}\hat{\ell}^i_{1:T} \leq \ell^i_{1:T} + \sqrt{(2\ln\frac{4}{\delta_T})\sum_{t=1}^T \hat{B}_t^2} + \frac{\delta_T}{2}\sum_{t=1}^T \hat{B}_t + \sum_{t=1}^{\tau^i-1} \hat{B}_t$.

Proof. For $t \geq \tau^i$, $X_t = \hat{\ell}^i_{1:t} - \ell^i_{1:t}$ is a martingale, since

$$\mathbf{E}[X_t|\mathcal{A}_{t-1}] = \mathbf{E}[\hat{\ell}^i_{1:t}|\mathcal{A}_{t-1}] - \ell^i_{1:t} = X_{t-1} + \mathbf{E}[\hat{\ell}^i_t|\mathcal{A}_{t-1}] - \ell^i_t = X_{t-1}.$$

It is clear that $X_{\tau^i-1} \leq \sum_{t=1}^{\tau^i-1} \hat{B}_t$. Moreover, $|X_t - X_{t-1}| \leq \hat{B}_t$ for $t \geq \tau^i$, i.e. we have bounded differences. By Azuma's inequality, the actual value $X_T - X_{\tau^i-1}$ does not exceed $\sqrt{(2\ln\frac{4}{\delta_T})\sum_{t=1}^T \hat{B}_t^2}$ with probability $1-\frac{\delta_T}{2}$. This proves (*i*). To arrive at (*ii*), take expectations and observe that (*i*) fails with probability at most $\frac{\delta_T}{2}$, in which case $\hat{\ell}^i_{1:T} \leq \sum_{t=1}^T \hat{B}_t$ holds. □

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

Theorem 7. [*FoE* against an adaptive adversary] *Let* $\sum_i \mathrm{e}^{-k^i} \leq 1$, τ^i *depend monotonically on* k^i, *and the learning rate* η_t *be decreasing. Let* ℓ_t *be some possibly adaptive assignment of (true) loss vectors satisfying* $\|\ell_t\|_\infty \leq B_t$. *Then for all experts* i, *we have with probability at least* $1-\delta_T$

$$\ell^{FoE}_{1:T} \leq \ell^i_{1:T} + \frac{k^i+1}{\eta_T} + \sum_{t=1}^{\tau^i-1}\hat{B}_t + \sum_{t=1}^{T}\gamma_t\eta_t\hat{B}_t^2 + \sum_{t=1}^{T}\gamma_t B_t + \sqrt{(2\ln\tfrac{4}{\delta_T})}\left(\sqrt{\sum_{t=1}^{T}\hat{B}_t} + \sqrt{\sum_{t=1}^{T}B_t^2}\right).$$

Consequently, in expectation, we have

$$\mathbf{E}\ell^{FoE}_{1:T} \leq \ell^i_{1:T} + \frac{k^i+1}{\eta_T} + \sum_{t=1}^{\tau^i-1}\hat{B}_t + \sum_{t=1}^{T}\gamma_t\eta_t\hat{B}_t^2 + \sum_{t=1}^{T}\gamma_t B_t + \sqrt{(2\ln\tfrac{4}{\delta_T})\sum_{t=1}^{T}\hat{B}_t} + \frac{\delta_T}{2}\sum_{t=1}^{T}\hat{B}_t.$$

Proof. This follows by summing up all excess terms in the above lemmas. Recall that we only need to take expectations on both sides of the assertions of Lemmas 2–5 in order to obtain the second bound on the expectation (and we don't need Lemma 1 there). □

Corollary 8. *Assume the conditions of Theorem 7 and choose* $\eta_t = t^{-\frac{3}{4}}$ *and* $\gamma_t = t^{-\frac{1}{4}}$. *Then*

(*i*) $B_t \equiv 1, \tau^i = \lceil (w^i)^{-8}\rceil \ \Rightarrow \mathbf{E}\ell^{FoE}_{1:T} \leq \ell^i_{1:T} + O\big((\frac{1}{w^i})^{11} + k^i T^{\frac{7}{8}}\sqrt{\ln T}\big)$,

(*ii*) $B_t \equiv 1, \tau^i = \lceil (w^i)^{-8}\rceil \ \Rightarrow \ell^{FoE}_{1:T} \leq \ell^i_{1:T} + O\big((\frac{1}{w^i})^{11} + k^i T^{\frac{7}{8}}\sqrt{\ln T}\big)$ *w.p.* $1-T^{-2}$,

(*iii*) $B_t = t^{\frac{1}{16}}, \tau^i = \lceil (w^i)^{-16}\rceil \Rightarrow \mathbf{E}\ell^{FoE}_{1:T} \leq \ell^i_{1:T} + O\big((\frac{1}{w^i})^{22} + k^i T^{\frac{7}{8}}\sqrt{\ln T}\big)$, *and*

(*iv*) $B_t = t^{\frac{1}{16}}, \tau^i = \lceil (w^i)^{-16}\rceil \Rightarrow \ell^{FoE}_{1:T} \leq \ell^i_{1:T} + O\big((\frac{1}{w^i})^{22} + k^i T^{\frac{7}{8}}\sqrt{\ln T}\big)$ *w.p.* $1-T^{-2}$, *for all i and $T \geq 1$ (recall $k^i = -\ln w^i$). Moreover, in both cases (bounded and growing B_t) FoE is asymptotically optimal w.r.t. each expert, i.e. for all i,*

$$\limsup_{T\to\infty} \frac{\ell_{1:T}^{FoE} - \ell_{1:T}^i}{T} \leq 0 \qquad \textit{almost surely.}$$

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

Proof. Assertions (i)-(iv) follow from the previous theorem: Set $\delta_T = T^{-2}$, abbreviate $w_T^{\min} = \min\{w^i : t \geq \tau^i\}$, and observe that for $\tau^i = \lceil (w^i)^{-\alpha} \rceil$ and $B_t = t^\beta$, we have

$$w_T^{\min} = \min\{w^i : T \geq \lceil (w^i)^{-\alpha} \rceil\} \geq \min\{w^i : T^{-\frac{1}{\alpha}} \leq w^i \rceil\} \geq T^{-\frac{1}{\alpha}} \qquad \text{and}$$

$$\sum_{t=1}^{\tau^i - 1} \hat{B}_t \leq (\tau^i - 1)\hat{B}_{\tau^i - 1} \leq \frac{(w^i)^{-\alpha} B_{\tau^i - 1}}{\gamma_{\tau^i - 1} w_{\tau^i - 1}^{\min}} \leq \frac{(w^i)^{-\alpha}(w^i)^{-\alpha\beta}}{(w^i)^{\frac{\alpha}{4}} w^i}$$

(note $w_{\tau^i-1}^{\min} \geq (\tau^i - 1)^{-\frac{1}{\alpha}} \geq (w^i)^{(-\alpha)(-\frac{1}{\alpha})}$). Then (i) and (ii) follow from $\alpha = 8$, $\beta = 0$, and (iii) and (iv) follow from $\alpha = 16$, $\beta = \frac{1}{16}$. The asymptotic optimality finally follows from the Borel-Cantelli Lemma, since according to (ii) and (iv),

$$\mathbf{P}\left[\frac{\ell_{1:T}^{FoE} - \min_i \ell_{1:T}^i}{T} > CT^{-\frac{1}{8}}\sqrt{\ln T}\right] \leq \frac{1}{T^2}$$

for an appropriate $C > 0$. □

As mentioned in the first paragraph of this section, it is possible to avoid Lemma 6, thus arriving at better bounds. E.g. in (i), choosing $\tau^i = \lceil (\frac{1}{w^i})^8 \rceil$, $\gamma_t = t^{-\frac{1}{4}}$, and $\eta_t = t^{-\frac{3}{4}}$, a regret bound of $O\big((\frac{1}{w^i})^{11} + k^i T^{\frac{3}{4}}\big)$ can be shown. Of course, also a corresponding high probability bound like (ii) holds. Likewise, for a similar statement as (iii), we may set $\tau^i = \lceil (\frac{1}{w^i})^{16} \rceil$, $B_t = t^{\frac{1}{8}}$, $\gamma_t = t^{-\frac{1}{4}}$, and $\eta_t = t^{-\frac{3}{4}}$, arriving at a regret bound of $O\big((\frac{1}{w^i})^{23} + k^i T^{\frac{3}{4}}\big)$ Generally, in this way any regret bound $O\big((\frac{1}{w^i})^c + k^i T^{\frac{2}{3}+\varepsilon}\big)$ is possible, at the cost of increasing c where $\varepsilon \to 0$.

4 Reactive environments and a universal master algorithm

Regret can become a quite subtle notion if we start considering reactive environments, i.e. care for future consequences of a decision. An extreme case is the "heaven-hell" example: We have two experts, one always playing 0 ("saying a prayer"), the other one always playing 1 ("cursing"). If we always follow the first expert, we stay in heaven and get no loss in each step. As soon as we "curse" only once, we get into hell and receive maximum loss in all subsequent time steps. Clearly, *any* algorithm without prior knowledge must "fail" in this situation.

set $\tilde{t} = 1$
For $t = 1, 2, 3, \ldots$
invoke $FoE(t)$ and play its decision for B_t basic time steps
set $\tilde{t} = \tilde{t} + B_t$

Fig. 4. The algorithm $\widetilde{FoE}$, where B_t is specified in Corollary 9.

One way to get around this problem is taking into account the *actual* (realization of the) game we are playing. For instance, after "cursing" once, also the praying expert goes to hell together with us and subsequently has maximum loss. Hence, were are interested in a regret defined as $\mathbf{E}\ell_{1:T} - \ell^i_{1:T}$ as in the previous section. So what is missing? This becomes clear in the following example.

Consider the repeated "prisoner's dilemma" against the tit-for-tat[3] strategy [1]. 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 standard expert advice or bandit master algorithm will not discover this, since it compares only the losses in one step, which are always lower for the defecting expert. 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 to time intervals of length one.

We therefore give the control to a selected expert for *periods of increasing length*. Precisely, we introduce a new time scale $\tilde{t}$ (the *basic* time scale) at which we have single games with losses $\tilde{\ell}_{\tilde{t}}$. The master's time scale t does not coincide with $\tilde{t}$. Instead, at each t, the master gives control to the selected expert i for $B_t \geq 1$ single games and receives loss $\ell^i_t = \sum_{\tilde{t}=\tilde{t}(t)}^{\tilde{t}(t)+B_t-1} \tilde{\ell}^i_{\tilde{t}}$. (The points $\tilde{t}(t)$ in basic time are defined recursively, see Fig. 4.) Assume that the game has bounded instantaneous losses $\tilde{\ell}^i_{\tilde{t}} \in [0,1]$. Then the master algorithm's instantaneous losses are bounded by B_t. We denote the algorithm, which is completely specified in Fig. 4, by $\widetilde{FoE}$. Then the following assertion is an easy consequence of the previous results.

Corollary 9. *Assume $\widetilde{FoE}$ plays a repeated game with bounded instantaneous losses $\tilde{\ell}^i_{\tilde{t}} \in [0,1]$. Choose $\gamma_t = t^{-\frac{1}{4}}$, $\eta_t = t^{-\frac{3}{4}}$, $B_t = \lfloor t^{\frac{1}{16}} \rfloor$ and $\tau^i = \lceil (w^i)^{-16} \rceil$. Then for all experts i and all $\tilde{T} \geq 1$,*

$$\tilde{\ell}^{\widetilde{FoE}}_{1:\tilde{T}} \leq \tilde{\ell}^i_{1:\tilde{T}} + O\big((\tfrac{1}{w^i})^{22} + k^i \tilde{T}^{\frac{9}{10}}\big) \text{ w.p. } 1 - \tilde{T}^{-\frac{32}{17}} \text{ and}$$
$$\mathbf{E}\tilde{\ell}^{\widetilde{FoE}}_{1:\tilde{T}} \leq \tilde{\ell}^i_{1:\tilde{T}} + O\big((\tfrac{1}{w^i})^{22} + k^i \tilde{T}^{\frac{9}{10}}\big).$$

[3] 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 loss, if both play D, they get a large loss. However, if one plays C and one D, the cooperating player gets a very large loss and the defecting player no loss 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.

Consequently, $\limsup_{T\to\infty}(\tilde\ell^{\widetilde{FoE}}_{1:\tilde T} - \tilde\ell^i_{1:\tilde T})/\tilde T \le 0$ *a.s. The rate of convergence is at least* $\tilde T^{-\frac{1}{10}}$*, and it can be improved to* $\tilde T^{-\frac{1}{3}+\varepsilon}$ *at the cost of a larger power of* $\frac{1}{w^i}$*.*

Proof. This follows from changing the time scale from t to $\tilde t$ in Corollary 8: $\tilde t$ is of order $t^{1+\frac{1}{16}}$. Consequently, the regret bound is $O\big((\frac{1}{w^i})^{22} + k^i \tilde T^{\frac{15}{17}}\sqrt{\ln \tilde T}\big) \le O\big((\frac{1}{w^i})^{22} + k^i \tilde T^{\frac{9}{10}}\big)$. □

Broadly spoken, this means that $FoE_{\tilde T}$ performs asymptotically as well as the best expert. Asymptotic performance guarantees for the Strategic Experts Algorithm have been derived in [1]. Our results improve upon this by providing a rate of convergence. One can give further corollaries, e.g. in terms of flexibility as defined in [1].

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 p^i of some fixed universal Turing machine. The ith program can be well-defined, e.g. by representing programs as binary strings and lexicographically ordering them [6]. Before the expert is consulted, the relevant input 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 for well-definedness just fill its output tape with zeros.) Each expert is assigned a prior weight by $w^i = 2^{-\mathrm{length}(p^i)}$, where length$(p^i)$ 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* [12].

Corollary 10. *If* $\widetilde{FoE}$ *is used together with a universal expert class as specified above and the parameters* $\eta_t, \gamma_t, B_t, \delta_T$ *are chosen as in Corollary 9, then it performs asymptotically at least as well as* any computable expert i. *The upper bound on the rate of convergence is exponential in the complexity* k^i *and proportional to* $\tilde t^{-\frac{1}{10}}$ *(improvable to* $\tilde t^{-\frac{1}{3}+\varepsilon}$*).*

The universal prior has been used to define a universal agent AIXI in a quite different way [13, 6]. Note that like the universal prior and the AIXI agent, our universal experts algorithm is not computable, since we cannot check if a the computation of an expert halts. 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 master algorithm is fully computable and has small regret with respect to all resource bounded strategies.

It is important to keep in mind that Corollaries 9 and 10 give assertions relative to the experts' performance merely on the *actual* action-observation sequence. In other words, if we wish to assess how well $\widetilde{FoE}$ does, we have to evaluate the actual *value* of the best expert [14]. Note that the whole point of our increasing time construction is to cause this actual value to coincide with the value under *ideal* conditions. For passive tasks, this coincidence always holds with any experts algorithm. With $\widetilde{FoE}$, the actual and the ideal value of an expert coincide in many further situations, such as "finitely controllable tasks". By this we mean cases where the best expert can drive the environment into some optimal state in a fixed finite number of time steps. An instance is the prisoner's dilemma with tit-for-tat being the opponent. The following is an example for a formalization of this statement.

Proposition 11. *Suppose $\widetilde{FoE}$ acts in a (fully or partially observable) Markov Decision Process. Let there be a computable strategy which is able to reach an ideal (that is optimal w.r.t. reward) state sequence in a fixed number of time steps. Then $\widetilde{FoE}$ performs asymptotically optimal.*

This statement may be generalized to cases where only a close to optimal state sequence is reached with high probability. However, we need assumptions on the closeness to optimality for a given target probability, which are compatible with the sampling behavior of $\widetilde{FoE}$.

Not all environments have this or similar nice properties. As mentioned above, any version of *FoE* would not perform well in the "heaven-hell" example. The following is a slightly more interesting variant of the heaven-hell task, where we might wish to learn optimal behavior, however *FoE* will not. Consider the heaven-hell example from the beginning of this section, but assume that if at time t I am in hell and I "pray" for t consecutive time steps, I will get back into heaven. Then it is not hard to see that *FoE*'s exploration is so dominant that almost surely, *FoE* will eventually stay in hell.

Simulations with some 2×2 matrix games show similar effects, depending on the opponent. We briefly discuss the repeated game of "chicken"[4]. In this game, it is desirable for the learner to become the "dominant defector", i.e. to defect in the majority of the cases while the opponent cooperates. Let's call an opponent "primitive" if he agrees to cooperate after a fixed number of consecutive defecting moves of *FoE*, and let's call him "stubborn" if this number is high. Then *FoE* 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 be the dominant defector and learns faster (we conducted the experiment with AIXI [6]), then *FoE* settles for cooperating, and the opponent will be the dominant defector. Interestingly however, AIXI would not learn to defect against a stubborn primitive opponent. Under this point of view, it seems questionable

[4] 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 one succeeds to let the other do the whole work, he has a high reward. On the other hand, if no one does anything, there will be no paper and thus no reward. Finally, if both decide to cooperate, both get some reward. We choose the loss matrix as $\binom{1 \;\; 0}{0.8 \;\; 0.5}$, the learner is the column player, the opponent's loss matrix is the transpose, choosing the fist column means to defect, the second to cooperate. Hence, in the repeated game, it is socially optimal to take turns cooperating and defecting.

that there is something like a universally optimal balance of exploration vs. exploitation in active learning at all.

5 Discussion

An alternative argument for adaptive adversary. As mentioned in the beginning of Section 3, the analysis we gave uses a trick from [4]. Such a trick seems necessary, as the basic FPL analysis only works for oblivious adversary. The simple argument from [11] which we used in the last paragraph of the proof of Lemma 5 works only for full observation games (note that considering the estimated losses, we were actually dealing with full observations there). In order to obtain a similar result in the partial observation case, we may argue as follows. We let the game proceed for T time steps with independent randomization against an adaptive adversary. Then we analyze *FoE*'s performance *in retrospective*. In particular, we note that for the losses assigned by the adversary, *FoE*'s expected regret coincides with the regret of another, virtual algorithm, which uses (in its FPL subroutine) identical perturbations $q_t \equiv q$. Performing the analysis for this virtual algorithm, we arrive at the desired assertion, however without needing Lemma 6. This results in tighter bounds as stated above. The argument is formally elaborated in [8].

Actual learning speed and lower bounds. In practice, the bounds we have proven seem irrelevant except for small expert classes, although asserting almost sure optimality and even a convergence rate. The exponential of the complexity $\frac{1}{w^i}$ may be huge. Imagine for instance a moderately complex task and some good strategy, which can be coded with mere 500 bits. Then its prior weight is 2^{-500}, a constant which is not distinguishable from zero in all practical situations. Thus, it seems that the bounds can be relevant at most for small expert classes with uniform prior. This is a general shortcoming of bandit style experts algorithms: For uniform prior a lower bound on the expected loss which scales with $\sqrt{n}$ (where n is the size of the expert class) has been proven [10].

In order to get a lower bound on *FoE*'s regret in the time T, observe that *FoE* is a *label-efficient* learner [15, 16]: According to the definition in [16], we may assume that in each exploration step, we incur maximal loss B_t. It is immediate that the same analysis then still holds. For label-efficient prediction, Cesa-Bianchi et al. [16] have shown a lower regret bound of $O(T^{\frac{2}{3}})$. Since according to the remark at the end of Section 3, we have an upper bound of $O\big((\frac{1}{w^i})^c + k^i T^{\frac{2}{3}+\varepsilon}\big)$, this is almost tight except for the additive $(\frac{1}{w^i})^c$ term. It is an open problem to state a lower bound simultaneously tight in both $\frac{1}{w^i}$ and T.

Even if the bounds, in particular $\frac{1}{w^i}$, seem not practical, maybe *FoE* would learn sufficiently quickly in practice anyway? We believe that this is not so in most cases: The design of *FoE* is too much tailored towards worst-case environments, *FoE* is too *defensive*. Assume that we have a "good" and a "bad" expert, and *FoE* learns this fact after some time. Then it still would spend a relatively huge fraction of $\gamma_t = t^{-\frac{1}{4}}$ to exploring the bad expert. Such defensive behavior seems only acceptable if we are already starting with a class of good experts.

Acknowledgment. This work was supported by SNF grant 2100-67712.02 and JSPS 21st century COE program C01.

References

1. de Farias, D.P., Megiddo, N.: How to combine expert (and novice) advice when actions impact the environment? In Thrun, S., Saul, L., Schölkopf, B., eds.: Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA (2004)
2. Hannan, J.: Approximation to Bayes risk in repeated plays. In Dresher, M., Tucker, A.W., Wolfe, P., eds.: Contributions to the Theory of Games 3. Princeton University Press (1957) 97–139
3. Kalai, A., Vempala, S.: Efficient algorithms for online decision. In: Proc. 16th Annual Conference on Learning Theory (COLT-2003). Lecture Notes in Artificial Intelligence, Berlin, Springer (2003) 506–521
4. McMahan, H.B., Blum, A.: Online geometric optimization in the bandit setting against an adaptive adversary. In: 17th Annual Conference on Learning Theory (COLT). Volume 3120 of Lecture Notes in Computer Science., Springer (2004) 109–123
5. Hutter, M., Poland, J.: Prediction with expert advice by following the perturbed leader for general weights. In: International Conference on Algorithmic Learning Theory (ALT). (2004) 279–293
6. Hutter, M.: Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability. Springer, Berlin (2004)
7. Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E.: Gambling in a rigged casino: The adversarial multi-armed bandit problem. In: Proc. 36th Annual Symposium on Foundations of Computer Science (FOCS 1995), Los Alamitos, CA, IEEE Computer Society Press (1995) 322–331
8. Poland, J.: FPL analysis for adaptive bandits. Technical report, Hokkaido University (2005)
9. Motwani, R., Raghavan, P.: Randomized Algorithms. Cambridge University Press, Cambridge, England (1995)
10. Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E.: The nonstochastic multiarmed bandit problem. SIAM Journal on Computing **32** (2002) 48–77
11. Hutter, M., Poland, J.: Adaptive online prediction by following the perturbed leader. Journal of Machine Learning Research **6** (2005) 639–660
12. Solomonoff, R.J.: Complexity-based induction systems: comparisons and convergence theorems. IEEE Trans. Inform. Theory **24** (1978) 422–432
13. Hutter, M.: Towards a universal theory of artificial intelligence based on algorithmic probability and sequential decisions. Proc. 12^{th} European Conference on Machine Learning (ECML-2001) (2001) 226–238
14. de Farias, D.P., Megiddo, N.: Exploration-exploitation tradeoffs for experts algorithms in reactive environments. In: Advances in Neural Information Processing Systems 17. (2005)
15. Cesa-Bianchi, N., Lugosi, G., Stoltz, G.: Minimizing regret with label efficient prediction. In: 17th Annual Conference on Learning Theory (COLT). Volume 3120 of Lecture Notes in Computer Science., Springer (2004) 77–92
16. Cesa-Bianchi, N., Lugosi, G., Stoltz, G.: Regret minimization under partial monitoring. Technical report (2004)