
Master Algorithms for Active Experts Problems based on Increasing Loss Values

Jan Poland
Marcus Hutter
 IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland

JAN@IDSIA.CH
 MARCUS@IDSIA.CH
 WWW.IDSIA.CH

Abstract

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 master algorithms for “active 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. This results in a (computationally infeasible) *universal master algorithm* which performs – in a certain sense – almost as well as any computable strategy, for any online problem.

Keywords. Prediction with expert advice, responsive environments, partial observation game, bandits, universal learning, asymptotic optimality.

1. Introduction

Expert algorithms have been popular since about fifteen years ago [LW89]. They are appropriate for online prediction or repeated decision making or repeated game playing (we call these setups *online problems* for brevity), based on a class of “experts”. In each round, each expert gives a recommendation. From this, we derive a master decision. After that, losses (or rewards) are assigned to each expert by the environment, also called adversary. Our goal is to perform almost as well as the best expert in hindsight in the long run. In other words, we try to minimize the regret.

The early papers deal with the full information game,

where we get to know the losses of each expert after each round. The analysis holds for the *worst case*, where the environment is fully adversarial and tries to maximize our regret in the long run. Later, [ACBFS95] gave a worst-case analysis for the *bandit* setup, where the master algorithm knows only the loss of its own decision after each round. This has been further generalized to *label-efficient prediction* [HP97] and *partial monitoring* [CBLS04].

Recently, [FM04] introduced a *strategic experts algorithm* which performs well for a broader class of environments. The algorithm has still asymptotically optimal properties against a worst-case adversary. Additionally, it may perform much better than a standard experts algorithm in more favorable situations, when the actions influence the behavior of the environment. We refer to these as *active experts problems*. One example is the repeated prisoner’s dilemma when the opponent is willing to cooperate under certain conditions (see Section 5 for some details). However, [FM04] give only asymptotic guarantees, but no convergence rate.

In this paper, we introduce a different algorithm for active experts problems with the same asymptotic guarantees, but in addition a convergence rate (of $t^{-\frac{1}{10}}$) is shown. Both algorithm and analysis are assembled from a standard “toolkit”, basing on [KV03, MB04]. The basic idea is the following: We use the bandit experts algorithm by [MB04], but allow the losses to increase with time t . This allows us to give control to one expert for an *increasing* period of time steps.

Secondly, we generalize our analysis to the case of *infinitely many* experts, basing on [HP04b]. The master algorithm stays computable (if the experts are), since only a finite (with time increasing) number of experts is involved. Allowing infinitely many experts also permits to define a *universal expert class* by means of all programs on some universal Turing machine. (This construction is quite common in Algorithmic Information Theory, see e.g. [Hut04].) Thus, we obtain a *uni-*

versal master algorithm, which we show to perform in a certain sense almost as well as *any computable* strategy on *any online problem*. Thus, we introduce a new approach to universal artificial intelligence, which is in a sense dual to the AIXI model based on Bayesian learning [Hut04]. Although the master algorithm is computable, the resulting universal agent is not (like the AIXI model), since the experts may be non-responsive.

The paper is structured as follows. Section 2 introduces the problem setup, the notation, and the algorithm. In Sections 3 and 4, we give the (worst-case) analysis for finite and infinite expert classes. The implications to active experts problems and a universal master algorithms are given in Section 5. Section 6 contains discussion and conclusions.

2. The Algorithm

Our task is an online decision problem. That is, we have to make a sequence of decisions, each of which results in a certain loss we incur. “We” is an abbreviation for the master algorithm which is to be designed. For concreteness, you may imagine the task of playing a game repeatedly. In each round, i.e. at each time step t , we have access to the recommendations of $n \in \mathbb{N} \cup \{\infty\}$ “experts” or strategies. 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 on the maximum loss the adversary may use, i.e. $\ell_t \in [0, B_t]^n$. This quantity may depend on t and is known to us. After the move, only the loss of the selected expert i is revealed. This is the *bandit setup*, as opposed to the full information game where we get to know the losses *all* experts. Our goal is to perform nearly as well as the best available strategy 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 is a finite number n of experts or strategies, then it is common to give no prior preferences to any of them. Formally, we define *prior weights* $w^i = \frac{1}{n}$. Moreover, we define the complexity of expert i as $k^i = -\ln w^i$. This arises in the full observation game, where the regret can be bounded by some function of the best expert’s complexity. On the other hand, if there are reasons not to trust all strategies equally in the beginning, we may use a non-uniform prior w . This is mandatory for infinitely many experts. We then require $w^i > 0$ for all experts i and $\sum_i w^i \leq 1$.

For $t=1,2,3,\dots$
 Sample $r_t \in \{0,1\}$ independently s.t. $P[r_t=1]=\gamma_t$
 If $r_t=0$ Then
 Play $FPL(t)$ ’s decision ($I_t^{FoE} := I_t^{FPL}$)
 Set $\hat{\ell}_t^i = 0$ for all $1 \leq i \leq n$
 Else
 Sample $I_t^{FoE} \in \{1\dots n\}$ uniformly & play $I := I_t^{FoE}$
 Let $\hat{\ell}_t^I = \ell_t^I n / \gamma_t$ and $\hat{\ell}_t^i = 0$ for all $i \neq I$

Figure 1. The algorithm FoE

Sample $q_t^i \stackrel{d}{\sim} Exp$ independently for $1 \leq i \leq n$
 select and play $I_t^{FPL} = \arg \min_{1 \leq i \leq n} \{\eta_t \hat{\ell}_{<t}^i + k^i - q_t^i\}$

Figure 2. The algorithm $FPL(t)$

Our algorithm “Follow or Explore” (FoE) builds on McMahan and Blum’s online geometric optimization algorithm. (For finite n and uniform prior, it even is their algorithm, save for the adaptive parameters.) 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 [Han57]. For the full observation game, [KV03] gave a very elegant analysis which is distinct from the standard analysis of exponential weighting schemes. It is particularly handy 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.

The algorithm is composed of two standard ingredients: *exploration* and *follow the (perturbed) leader*. Since we are playing the bandit game (as opposed to the full information game), we need to explore sufficiently. Otherwise, there could be a strategy which we think is poor (and thus never play), but in reality it is good. 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 choose an expert according to the uniform distribution (or the prior distribution, compare (5), in case of non-uniform priors). After observing the loss of the selected expert, we want to give an *unbiased estimate* of the true loss vector. We achieve that by dividing the observed loss by the probability of exploring this expert, and estimate the unobserved losses of all other experts by zero. We call the resulting loss vector $\hat{\ell}_t$.

When not exploring, we follow some strategy which performed well in the past. It may be not advisable to pick always the *best* strategy so far - the adver-

sary could fool us in this case. Instead we introduce a *perturbation* for each expert and follow the advice of the strategy with the best perturbed score. In order to assign a score to each expert, note that we have only access to the *estimated* losses $\hat{\ell}_t$. Let $\hat{\ell}_{<T}^i = \sum_{t=1}^{T-1} \hat{\ell}_t^i$ be the estimated cumulative past loss of expert i . Then his complexity-penalized score is defined as $\eta_T \hat{\ell}_{<T}^i + k^i$, i.e. high scores are bad. Here, $\eta_T > 0$ is the *learning rate*. The perturbed score is then given by $\eta_T \hat{\ell}_{<T}^i + k^i - q^i$, where the perturbations q^i are chosen independently exponentially distributed. This ensures a convenient analysis.

The algorithms “Follow or Explore” *FoE* and “Follow the perturbed Leader” *FPL* are fully specified in Figures 1 and 2. Note that each time randomness is used, it is assumed to be *independent* of the past randomness. Note also that all algorithms occurring in this paper work with the *estimated losses* $\hat{\ell}$. We may evaluate their performance in terms of true or estimated losses, this is specified in the notation. E.g. for the true loss of *FPL* up to and including time T we write $L^{FPL} = \ell_{1:T}^{FPL}$, while the estimated loss is $\hat{L}^{FPL} = \hat{\ell}_{1:T}^{FPL}$.

3. Analysis for Uniform Prior

In this section we assume a uniform prior $w \equiv \frac{1}{n}$ over finitely many experts. (The general case is treated in the next section.) We assume that $B_t \geq 0$ is some sequence of upper bounds on the true losses, $\gamma_t \in (0,1)$ is a sequence of exploration rates, and $\eta_t > 0$ is a *decreasing* sequence of learning rates.

The analysis is according to the following diagram:

$$L^{FoE} \lesssim \mathbf{E}L^{FoE} \lesssim \mathbf{E}L^{FPL} \lesssim \mathbf{E}\hat{L}^{FPL} \lesssim \mathbf{E}\hat{L}^{IFPL} \lesssim \hat{L}^{best} \lesssim L^{best} \quad (1)$$

The symbol L is used informally for the cumulative loss $\ell_{1:T}$. Each “ \lesssim ” means that we bound the quantity on the left by the quantity on the right plus some additive terms. The first and the last expressions are the losses of the *FoE* algorithm and the best expert, respectively. The intermediate quantities belong to different algorithms, namely *FoE*, *FPL*, and a third one called *IFPL* for “infeasible” FPL [KV03]. *IFPL* 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 $\eta_t \hat{\ell}_{<t}^i + k^i - q_t^i$. We assume that *IFPL* uses the same randomization as *FPL* (i.e. the respective q_t are the same).

The randomization of *FoE* and *FPL* gives rise to two filters of σ -algebras. By \mathcal{A}_t for $t \geq 0$ we denote the σ -algebra generated by the *FoE*’s randomness $\{u_{1:t}, r_{1:t}\}$ up to time t . 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. $\mathcal{B}_t \hat{=} \{u_{1:t}, r_{1:t}, q_{1:t}\}$). Then clearly $\mathcal{A}_t \subset \mathcal{B}_t$ for each t .

The arguments below rely on *conditional expectations* – the expectations in (1) should also be understood conditional. In particular we will often need the conditional expectations with respect to *FoE*’s past randomness \mathcal{A}_{t-1} , abbreviated as

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

where X is some random variable. 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 (only). 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 $q_{1:t}$. Finally, I_t^{FoE} and ℓ_t^{FoE} are \mathcal{A}_t^i measurable, i.e. depend on $u_{<t}, r_{<t}, q_t$, but are independent of $q_{<t}$.

We now start proving the diagram (1). It is helpful to consider each intermediate algorithm as a stand-alone procedure which is actually executed (with an oracle if necessary) and has the asserted performance guarantees (e.g. in terms of expected losses).

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

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

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

$$\begin{aligned} \mathbf{E}(X_T | \mathcal{B}_{T-1}) &= \\ &= \sum_{t=1}^T (\mathbf{E}[\ell_t^{FoE} | \mathcal{B}_{T-1}] - \mathbf{E}[\mathbf{E}[\ell_t^{FoE} | \mathcal{A}_{t-1}] | \mathcal{B}_{T-1}]) \\ &= \sum_{t=1}^{T-1} (\ell_t^{FoE} - \mathbf{E}[\ell_t^{FoE} | \mathcal{A}_{t-1}]) = X_{T-1}. \end{aligned}$$

Its differences are bounded: $|X_t - X_{t-1}| \leq B_t$. Hence, it follows from Azuma’s inequality that the probability that X_T exceeds some $\lambda > 0$ is bounded by $p = 2 \exp(-\frac{\lambda^2}{2 \sum_{t=1}^T B_t^2})$. Requesting $\frac{\delta_T}{2} = p$ and solving for λ gives the assertion. \square

The relation $\mathbf{E}L^{FoE} \lesssim \mathbf{E}L^{FPL}$ follows immediately from the specification of the algorithm *FoE*.

Lemma 2 [$\mathbf{E}L^{FoE} \lesssim \mathbf{E}L^{FPL}$] For each $t \geq 1$, we have $\mathbf{E}_t \ell_t^{FoE} \leq (1-\gamma_t) \mathbf{E}_t \ell_t^{FPL} + \gamma_t B_t$.

The next lemma relating $\mathbf{E}L^{FPL}$ and $\mathbf{E}\hat{L}^{FPL}$ is technical but intuitively clear. It states that in (conditional) 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 with the previous lemma was shown in [MB04].

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

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 .

Proof. Let $f_t^i = f_t^i(\mathcal{A}_{t-1}) = \mathbf{P}[I_t^{FPL} = i | \mathcal{A}_{t-1}]$ be the probability distribution over actions i which FPL uses at time t , depending on the past randomness \mathcal{A}_{t-1} . Let $u_t = [1 \dots 1]/n$ be the uniform distribution at time t (for non-uniform weights this will be replaced appropriately later). Then

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

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 $I_t^{FoE} = i$. \square

The following lemma from [KV03] relates the losses of FPL and $IFPL$. We repeat the proof, since it is the crucial and only step in the analysis where we have to be careful with the upper loss bound B_t . Let $\hat{B}_t = B_t(n/\gamma_t)$ denote the upper bound on the instantaneous estimated losses.

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

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

$$\mathbf{E}_t \hat{\ell}_t^{FPL} = \sum_{i=1}^n \int \mathbb{1}_{I_t^{FPL}=i} \hat{\ell}_t^i d\mu(x), \quad (2)$$

where μ denotes the (exponential) distribution of the perturbations, i.e. $x_i := q_t^i$ and density $\mu(x) := e^{-\|x\|_\infty}$. 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 $\int \dots d\mu(x_{\neq i})$ the integration leaving out the i th action. Then, using $\eta_t \lambda_i - x_i \leq \eta_t \lambda_j - x_j$ for all j if $I_t^{FPL} = i$ in the first line,

and $\hat{B}_t \geq \hat{\ell}_t^i - \hat{\ell}_t^j$ in the fourth line, we get

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

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

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 was given in [KV03]. An additional step is necessary for an adaptive adversary. We omit the proof here, the reader may reconstruct it from the proof of Lemma 9.

Lemma 5 [$\mathbf{E}\hat{L}^{IFPL} \lesssim \hat{L}^{best}$] Assume decreasing learning rate η_t and $\sum_i e^{-k^i} \leq 1$. For all $T \geq 1$ and $1 \leq i \leq n$, we have $\sum_{t=1}^T \mathbf{E}_t \hat{\ell}_t^{IFPL} \leq \hat{\ell}_{1:T}^i + \frac{k^i}{\eta_T}$ (recall that $\hat{\ell}_{1:T}^i$ is a random variable depending on \mathcal{A}_t).

Finally, we give a relation between the estimated and true losses, adapted from [MB04].

Lemma 6 [$\hat{L}^{best} \lesssim L^{best}$] For each $T \geq 1$, $\delta_T \in (0, 1)$, and $1 \leq i \leq n$, w.p. at least $1 - \frac{\delta_T}{2}$ we have

$$\hat{\ell}_{1:T}^i \leq \ell_{1:T}^i + \sqrt{(2 \ln \frac{4}{\delta_T}) \sum_{t=1}^T \hat{B}_t^2}. \quad (3)$$

Proof. $X_t = \hat{\ell}_{1:t}^i - \ell_{1:t}^i$ is a martingale, since

$$\begin{aligned} \mathbf{E}[X_t | \mathcal{A}_{t-1}] &= \mathbf{E}[\hat{\ell}_{1:t}^i | \mathcal{A}_{t-1}] - \ell_{1:t}^i \\ &= X_{t-1} + \mathbf{E}[\hat{\ell}_t^i | \mathcal{A}_{t-1}] - \ell_t^i = X_{t-1}. \end{aligned}$$

Its differences are bounded: $|X_t - X_{t-1}| \leq \hat{B}_t$. By Azuma's inequality, its actual value at time T does not exceed $\sqrt{(2 \ln \frac{4}{\delta_T}) \sum_{t=1}^T \hat{B}_t^2}$ w.p. $1 - \frac{\delta_T}{2}$. \square

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 n be finite and $k^i = \ln n$ for all $1 \leq i \leq n$. Let η_t be decreasing, and $\ell_t \in [0, B_t]^n$ some possibly adaptive assignment of loss vectors. Then for all experts i ,*

$$\begin{aligned} \ell_{1:T}^{FoE} &\leq \ell_{1:T}^i + \sqrt{(2 \ln \frac{4}{\delta_T})} \left(\sqrt{\sum_{t=1}^T \frac{B_t^2 n^2}{\gamma_t}} + \sqrt{\sum_{t=1}^T B_t^2} \right) \\ &+ \frac{\ln n}{\eta_T} + \sum_{t=1}^T \frac{\eta_t B_t^2 n^2}{\gamma_t} + \sum_{t=1}^T \gamma_t B_t \quad \text{w.p. } 1 - \delta_T \text{ and} \\ \mathbf{E} \ell_{1:T}^{FoE} &\leq \ell_{1:T}^i + \frac{\ln n}{\eta_T} + \sum_{t=1}^T \frac{\eta_t B_t^2 n^2}{\gamma_t} + \sum_{t=1}^T \gamma_t B_t \\ &+ \sqrt{(2 \ln \frac{4}{\delta_T})} \sum_{t=1}^T \frac{B_t^2 n^2}{\gamma_t} + \frac{\delta_T}{2} \sum_{t=1}^T \frac{B_t n}{\gamma_t}. \end{aligned}$$

Proof. The first high probability bound follows by summing up all excess terms in the above lemmas, observing that $\hat{B}_t = B_t(n/\gamma_t)$. For the second bound on the expectation, we take expectations in Lemmas 2-5, while Lemma 1 is not used. For Lemma 6, a statement in expectation is obtained as follows: (3) fails w.p. at most $\frac{\delta_T}{2}$, in which case $\hat{\ell}_{1:T}^i - \ell_{1:T}^i \leq \sum_{t=1}^T \hat{B}_t$. \square

Corollary 8 *Under the conditions of Theorem 7,*

- (i) $B_t \equiv 1 \Rightarrow \mathbf{E} \ell_{1:T}^{FoE} \leq \ell_{1:T}^i + O(n^2 T^{\frac{3}{4}} \sqrt{\ln T})$,
- (ii) $B_t \equiv 1 \Rightarrow \ell_{1:T}^{FoE} \leq \ell_{1:T}^i + O(n^2 T^{\frac{3}{4}} \sqrt{\ln T})$,
- (iii) $B_t = t^{\frac{1}{8}} \Rightarrow \mathbf{E} \ell_{1:T}^{FoE} \leq \ell_{1:T}^i + O(n^2 T^{\frac{7}{8}} \sqrt{\ln T})$,
- (iv) $B_t = t^{\frac{1}{8}} \Rightarrow \ell_{1:T}^{FoE} \leq \ell_{1:T}^i + O(n^2 T^{\frac{7}{8}} \sqrt{\ln T})$,

for all i and T . Here, (ii) and (iv) hold with probability $1 - T^{-2}$. Moreover, in both cases (bounded and growing B_t) *FoE* is asymptotically optimal, i.e.

$$\limsup_{T \rightarrow \infty} \frac{1}{T} (\ell_{1:T}^{FoE} - \min_i \ell_{1:T}^i) \leq 0 \quad \text{almost surely.}$$

$B_t = t^{\frac{1}{8}}$ in (iii) and (iv) is just one choice to achieve asymptotic optimality while the losses may grow unboundedly. Asymptotic optimality is sometimes termed *Hannan-consistency*, in particular if the limit equals zero. We only show the upper bound.

Proof. (i) and (ii) follow by applying the previous theorem to $\eta_t = t^{-\frac{1}{2}}$, $\gamma_t = t^{-\frac{1}{4}}$, $\delta_T = T^{-2}$, and observing $\sum_{t=1}^T t^\alpha \leq \int_0^{T+1} t^\alpha \leq 2(T+1)^{1+\alpha}$ for $\alpha \geq -\frac{1}{2}$. In order to obtain (iii) and (iv), set $\eta_t = t^{-\frac{3}{4}}$, $\gamma_t = t^{-\frac{1}{4}}$, and $\delta_T = T^{-2}$. The asymptotic optimality finally follows from the Borel-Cantelli Lemma, since

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

for an appropriate $C > 0$ according to (ii) and (iv). \square

For $t=1,2,3,\dots$
 Sample $r_t \in \{0,1\}$ independently s.t. $P[r_t=1] = \gamma_t$
 If $r_t=0$ Then
 Invoke $FPL^\tau(t)$ and play its decision
 Set $\hat{\ell}_t^i = 0$ for $i \in \{t \geq \tau\}$
 Else
 Sample I_t w.r.t. u_t in (5) and play $I := I_t^{FoE^\tau}$
 Set $\hat{\ell}_t^I = \ell_t^I / (u_t^I \gamma_t)$ and $\hat{\ell}_t^i = 0$ for $i \in \{t \geq \tau\} \setminus \{I\}$
 Set $\hat{\ell}_t^i = \hat{B}_t$ for $i \notin \{t \geq \tau\}$

Figure 3. The algorithm FoE^τ

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

Figure 4. The algorithm $FPL^\tau(t)$

4. Infinitely Many Experts and Arbitrary Priors

The following considerations are valid for both finitely and infinitely many experts with arbitrary prior weights w^i . For notational convenience, we write $n = \infty$ in the latter case. When admitting infinitely many experts, two difficulties arise: Since the prior weights of the experts sum up to one and thus become arbitrarily small, the estimated losses – obtained by dividing by these weights – would possibly get arbitrarily large. We therefore introduce, for each expert i , a time $\tau^i \geq 1$ at which the expert enters the game. All algorithms *FoE*, *FPL*, *IFPL* are substituted by counterparts FoE^τ , FPL^τ , $IFPL^\tau$ which use expert i only for $t \geq \tau^i$. Thus, the maximum estimated loss possibly assigned to these *active* experts is

$$\hat{B}_t = B_t / [\gamma_t \min\{w^i : t \geq \tau^i\}]. \quad (4)$$

We denote the set of active experts at time t by $\{t \geq \tau\} = \{i : t \geq \tau^i\}$. Experts which have not yet entered the game are given an estimated loss of \hat{B}_t . This also solves the computability problem: Since at every time t only a finite number of experts is involved, FoE^τ is computable (if each expert is). The algorithms FoE^τ and FPL^τ are specified in Figures 3 and 4.

Again, the analysis follows the outline (1). Lemmas 1–4 have equivalent counterparts, the proofs of which remain almost unchanged. In Lemma 3, the “uniform” distribution over experts u_t now becomes

$$u_t^i = w^i \mathbb{I}_{t \geq \tau^i} / [\sum_j w^j \mathbb{I}_{t \geq \tau^j}]. \quad (5)$$

The upper bound on the estimated loss \hat{B}_t in Lemma

4 is given by (4). We only need to prove assertions corresponding to Lemmas 5 and 6.

Lemma 9 [$\mathbf{E}\hat{L}^{IFPL^\tau} \lesssim \hat{L}^{best^\tau}$] *Assume that $\sum_i e^{-k^i} \leq 1$ 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 $1 \leq i \leq n$, we have*

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

Proof. This is a modification of the corresponding proofs in [KV03] and [HP04b]. We may fix the randomization \mathcal{A} and suppress it in the notation. Then we only need to show

$$\mathbf{E} \hat{\ell}_{1:T}^{IFPL^\tau} \leq \min_{1 \leq i \leq n} \left\{ \hat{\ell}_{1:T}^i + \frac{k^i+1}{\eta_T} \right\}, \quad (6)$$

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

Assume first that the adversary is oblivious. We define an algorithm A as a variant of $IFPL^\tau$ 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 $IFPL^\tau$ in terms of expected performance. This is all we need to show (6). Let $\eta_0 = \infty$ and $\lambda_t = \hat{\ell}_t + (k-q)\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)$, 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 $T \geq 1$,

$$\sum_{t=1}^T \lambda_t^A \leq \min_{T \geq \tau} \lambda_{1:T}^i + \max_{T \geq \tau} \left\{ \frac{q^i - k^i}{\eta_T} \right\}. \quad (7)$$

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

$$\begin{aligned} \min_{T \geq \tau} \lambda_{1:T}^i + \max_{T \geq \tau} \frac{q^i - k^i}{\eta_T} + \lambda_{T+1}^A &\leq \lambda_{1:T+1}^{I_{T+1}^A} \\ + \max_{T+1 \geq \tau} \frac{q^i - k^i}{\eta_{T+1}} + \lambda_{T+1}^{I_{T+1}^A} &= \min_{T+1 \geq \tau} \lambda_{1:T+1}^i + \max_{T+1 \geq \tau} \frac{q^i - k^i}{\eta_{T+1}}. \end{aligned} \quad (8)$$

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

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

shows (8). Rearranging terms in (7), we see

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

The assertion (6) – still for oblivious adversary and $q_t \equiv q$ – then follows by taking expectations and using

$$\begin{aligned} \mathbf{E} \min_{T \geq \tau} \lambda_{1:T}^i &\leq \min_{T \geq \tau} \left\{ \hat{\ell}_{1:T}^i + \frac{k^i}{\eta_T} - \mathbf{E} \frac{q^i}{\eta_T} \right\} \stackrel{(*)}{\leq} \min_{1 \leq i \leq n} \left\{ \hat{\ell}_{1:T}^i + \frac{k^i-1}{\eta_T} \right\} \\ \text{and } \mathbf{E} \sum_{t=1}^T (q-k)^{I_t^A} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) &\leq \mathbf{E} \max_{T \geq \tau} \left\{ \frac{q^i - k^i}{\eta_T} \right\} \leq \frac{1}{\eta_T}. \end{aligned}$$

Here, $(*)$ 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 last inequality can be proven by an application of the union bound [HP04b, 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 $IFPL^\tau$ is played against an oblivious adversary: (6) remains valid. In the last step, we argue that then (6) also holds for an *adaptive* adversary. This is true because the future actions of $IFPL^\tau$ do not depend on its past actions, and therefore the adversary cannot gain from deciding after having seen $IFPL^\tau$'s decisions. (For details see [HP04a]. Note the subtlety that the future actions of FoE^τ would depend on its past actions.) \square

Lemma 10 [$\hat{L}^{best^\tau} \lesssim L^{best}$] *For each $T \geq 1$, $\delta_T \in (0,1)$, and $1 \leq i \leq n$, we have $\hat{\ell}_{1:T}^i \leq \ell_{1:T}^i + \sqrt{(2 \ln \frac{4}{\delta_T}) \sum_{t=1}^T \hat{B}_t^2} + \sum_{t=1}^{\tau^i-1} \hat{B}_t$ w.p. $1 - \frac{\delta_T}{2}$.*

This corresponds to Lemma 6. The proof proceeds in a similar way: we have to note that $\hat{\ell}_{1:t}^i - \ell_{1:t}^i$ is a martingale only for $t \geq \tau^i$, and $\hat{\ell}_{< \tau^i}^i$ exceeds $\ell_{< \tau^i}^i$ by at most $\sum_{t=1}^{\tau^i-1} \hat{B}_t$. Then the following theorem corresponds to Theorem 7 and is proven likewise.

Theorem 11 [FoE^τ against an adaptive adversary] *Let n be finite or infinite, $\sum_i e^{-k^i} \leq 1$, τ^i depend monotonically on k^i , and the learning rate η_t be decreasing. Let ℓ_t some possibly adaptive assignment of (true) loss vectors satisfying $\|\ell_t\|_\infty \leq B_t$. Then for all experts i , we have*

$$\begin{aligned} \ell_{1:T}^{FoE^\tau} &\leq \ell_{1:T}^i + \sqrt{(2 \ln \frac{4}{\delta_T}) \left(\sqrt{\sum_{t=1}^T \frac{B_t^2}{\gamma_t^2 (w_t^*)^2}} + \sqrt{\sum_{t=1}^T B_t^2} \right)} \\ &\quad + \frac{k^i+1}{\eta_T} + \sum_{t=1}^{\tau^i-1} \frac{B_t}{\gamma_t w_t^*} + \sum_{t=1}^T \frac{\eta_t B_t^2}{\gamma_t (w_t^*)^2} + \sum_{t=1}^T \gamma_t B_t \end{aligned}$$

with probability $1 - \delta_T$, where $w_t^* = \min\{w^i : t \geq \tau^i\}$. A corresponding statement holds for the expectation (compare Theorem 7).

Corollary 12 *Assume the conditions of Theorem 11.*

Then for all i and T , the following holds w.p. $1 - \delta_T$.

- (i) $B_t \equiv 1, \tau^i = \lceil (w^i)^{-8} \rceil$
 $\Rightarrow \ell_{1:T}^{FoE} \leq \ell_{1:T}^i + O\left(\left(\frac{1}{w^i}\right)^{11} + T^{\frac{7}{8}}\sqrt{\ln T}\right)$, and
(ii) $B_t = t^{\frac{1}{16}}, \tau^i = \lceil (w^i)^{-16} \rceil$
 $\Rightarrow \ell_{1:T}^{FoE} \leq \ell_{1:T}^i + O\left(\left(\frac{1}{w^i}\right)^{22} + T^{\frac{7}{8}}\sqrt{\ln T}\right)$.

Corresponding assertions are true for the expectation (compare Corollary 8). In both cases (bounded and growing B_t) FoE is asymptotically optimal w.r.t. each expert: $\limsup_{T \rightarrow \infty} \frac{1}{T} (\ell_{1:T}^{FoE} - \ell_{1:T}^i) \leq 0$ a.s. for all i .

Proof. Let $\eta_t = t^{-\frac{3}{4}}, \gamma_t = t^{-\frac{1}{4}}$, and $\delta_T = T^{-2}$. For $\tau^i = \lceil (w^i)^{-\alpha} \rceil$ and $B_t = t^\beta$, we have $w_{\tau^i}^* = \min\{w^i : T \geq \lceil (w^i)^{-\alpha} \rceil\} \geq \min\{w^i : T^{-\frac{1}{\alpha}} \leq w^i\} \geq T^{-\frac{1}{\alpha}}$ and

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

(observe $w_{\tau^i-1}^* \geq (\tau^i - 1)^{-\frac{1}{\alpha}} \geq (w^i)^{-(\alpha)(-\frac{1}{\alpha})}$). Then set $\alpha=8, \beta=0$, for (i) and $\alpha=16, \beta=\frac{1}{16}$ for (ii). Asymptotic optimality is shown as in Corollary 8. \square

5. Active Expert Problems and a Universal Master Algorithm

If the adversary's goal is just to maximize our (expected) regret, then it is well known what he can achieve (at least for uniform prior, see e.g. the lower bound in [CB97, ACBFS02]). We are interested in different situations. An example is the repeated playing of the "Prisoner's dilemma" against the Tit-for-Tat¹ strategy [FM04]. If we use two strategies as experts, namely "always cooperate" and "always defect", then it is clear that always cooperating will have the better long-term reward. It is also clear that a 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.

We therefore propose to give the control to a selected expert for *periods of increasing length*. Precisely, we introduce a new time scale \tilde{t} at which we have single games with losses $\tilde{\ell}_t$. 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 \tilde{T}_t single games and

¹In the prisoner's dilemma, two players both decide independently if they 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. A Tit-for-Tat player play C in the first move and afterwards the opponent's respective preceding move.

receives loss $\ell_t^i = \sum_{\tilde{t}=\tilde{t}(t)}^{\tilde{t}(t)+\tilde{T}_t-1} \tilde{\ell}_t^i$. Assume that the game has bounded instantaneous losses $\tilde{\ell}_t^i \in [0,1]$. Then the master algorithm's instantaneous losses are bounded by \tilde{T}_t . We denote this algorithm by $FoE_{\tilde{T}}$ or $FoE_{\tilde{T}}^i$.

Corollary 13 Assume $FoE_{\tilde{T}}$ (or $FoE_{\tilde{T}}^i$, respectively) plays a repeated game with bounded instantaneous losses $\tilde{\ell}_t^i \in [0,1]$. Let the exploration and learning rates be $\gamma_t = t^{-\frac{1}{4}}$ and $\eta_t = t^{-\frac{3}{4}}$. In case of uniform prior, choose $\tilde{T}_t = \lfloor t^{\frac{1}{8}} \rfloor$ ($\tau^i \equiv 0$). In case of arbitrary prior let $\tilde{T}_t = \lfloor t^{\frac{1}{16}} \rfloor$ and $\tau^i = \lceil (w^i)^{-16} \rceil$. Then for all experts i and all \tilde{T} , suppressing the dependence on the prior of expert i , we have

$$\begin{aligned} \ell_{1:\tilde{T}}^{FoE_{\tilde{T}}} &\leq \ell_{1:\tilde{T}}^i + O(\tilde{T}^{\frac{9}{10}}) \text{ w.p. } 1 - \tilde{T}^{-2} \text{ and} \\ \mathbf{E} \ell_{1:\tilde{T}}^{FoE_{\tilde{T}}} &\leq \ell_{1:\tilde{T}}^i + O(\tilde{T}^{\frac{9}{10}}). \end{aligned}$$

Consequently, $\limsup_{T \rightarrow \infty} (\ell_{1:T}^{FoE_{\tilde{T}}} - \ell_{1:T}^i) / \tilde{T} \leq 0$ almost surely. The rate of convergence is at least $\tilde{T}^{-\frac{1}{10}}$. The same assertions hold for $FoE_{\tilde{T}}^i$.

Proof. This follows from changing the time scale from t to \tilde{t} in Corollaries 8 and 12: \tilde{t} is of order $t^{1+\frac{1}{8}}$ in the uniform case and $t^{1+\frac{1}{16}}$ in the general case. Then the bounds are $\tilde{T}^{\frac{9}{8}}\sqrt{\ln \tilde{T}}$ in the former and $\tilde{T}^{\frac{15}{17}}\sqrt{\ln \tilde{T}}$ in the latter case. Both are upper bounded by $\tilde{T}^{\frac{9}{10}}$. \square

Broadly spoken, this means that $FoE_{\tilde{T}}$ performs asymptotically as well as the best expert. Asymptotic guarantees for the Strategic Experts Algorithm have been derived by [FM04]. Our results approve upon this by providing a rate of convergence. One can give further corollaries, e.g. in terms of flexibility as defined by [FM04].

It is also possible to specify a *universal* experts algorithm. To this aim, let expert i be derived from the i th program p^i of some fixed universal Turing machine. The i th program can be well-defined, e.g. by representing programs as binary strings and lexicographically ordering them [Hut04]. Before the expert is consulted, the relevant input is written to the input tape of the corresponding program. If the program halts, the appropriate number of first bits 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 $w^i = 2^{-\text{length}(p^i)}$, where $\text{length}(p^i)$ 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* [Sol78]. This has been used to define a universal agent AIXI in a quite different way by [Hut04]. Note that like the universal prior and AIXI, our universal

agent is not computable, since we cannot check if a program halts. It is however straightforward to impose a bound on the computation time which for instance increases rapidly in t . If used with computable experts, the algorithm is computationally feasible. The universal master algorithm performs well with respect to *any* computable strategy.

Corollary 14 *Assume the universal set of experts specified in the last paragraph. If FoE_T^x is applied with $\gamma_t = t^{-\frac{1}{4}}$, $\eta_t = t^{-\frac{3}{4}}$, $\tilde{T}_t = \lfloor t^{\frac{1}{10}} \rfloor$, and $\tau^i = \lceil (w^i)^{-16} \rceil$, then it performs asymptotically at least as good as any computable expert i . The rate of convergence is exponential in the complexity k^i and proportional to $\tilde{T}^{-\frac{1}{10}}$.*

6. Discussion

For large or infinite expert classes, the bounds we have proven are irrelevant in practice, although asserting almost sure optimality and even a convergence rate: the exponential of the complexity is far too huge. Imagine for instance a moderately complex task and some good strategy, which can be coded with mere 500 bits. Then its 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 experts algorithms: For uniform prior a lower bound on the expected loss which is linear in \sqrt{n} has been proven [ACBFS02].

If the bounds are not practically relevant, maybe the algorithms are so? We leave this interesting question unanswered. Intuitively, it might seem that the algorithms proposed here are too much tailored towards worst-case bounds and fully adversarial setups. For example, the exploration rate of $t^{-\frac{1}{4}}$ is quite high. Master algorithms which are less “cautious” might perform better for many practical problems. Finally, it would be nice to investigate the differences between the proposed expert style approach and other definitions of universal agents, such as by [Hut04].

Acknowledgement: This work was supported by SNF grant 2100-67712.02.

References

- [ACBFS95] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proc. 36th Annual Symposium on Foundations of Computer Science (FOCS 1995)*, pages 322–331, Los Alamitos, CA, 1995. IEEE Computer Society Press.
- [ACBFS02] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.
- [CB97] N. Cesa-Bianchi et al. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.
- [CBLS04] N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Regret minimization under partial monitoring. Technical report, 2004.
- [FM04] D. Pucci de Farias and N. Megiddo. How to combine expert (and novice) advice when actions impact the environment? In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.
- [Han57] J. Hannan. Approximation to Bayes risk in repeated plays. In M. Dresher, A. W. Tucker, and P. Wolfe, editors, *Contributions to the Theory of Games 3*, pages 97–139. Princeton University Press, 1957.
- [HP97] D. Helmbold and S. Panizza. Some label efficient learning results. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 218–230. ACM Press, 1997.
- [HP04a] M. Hutter and J. Poland. Adaptive online prediction by following the perturbed leader. Technical Report IDSIA-30-04, 2004.
- [HP04b] M. Hutter and J. Poland. Prediction with expert advice by following the perturbed leader for general weights. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 279–293, 2004.
- [Hut04] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2004. 300 pages, <http://www.idsia.ch/~marcus/ai/uaibook.htm>.
- [KV03] A. Kalai and S. Vempala. Efficient algorithms for online decision. In *Proc. 16th Annual Conference on Learning Theory (COLT-2003)*, Lecture Notes in Artificial Intelligence, pages 506–521, Berlin, 2003. Springer.
- [LW89] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. In *30th Annual Symposium on Foundations of Computer Science*, pages 256–261, Research Triangle Park, North Carolina, 1989. IEEE.
- [MB04] H. B. McMahan and A. Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *17th Annual Conference on Learning Theory (COLT)*, volume 3120 of *Lecture Notes in Computer Science*, pages 109–123. Springer, 2004.
- [Sol78] R. J. Solomonoff. Complexity-based induction systems: comparisons and convergence theorems. *IEEE Trans. Information Theory*, IT-24:422–432, 1978.