Adaptive Bandits:
Towards the best history-dependent strategy

Anonymous Author 1
Unknown Institution 1

Anonymous Author 2
Unknown Institution 2

Anonymous Author 3
Unknown Institution 3

Abstract

We consider multi-armed bandit games with possibly adaptive opponents. We introduce models Θ of constraints based on equivalence classes on the common history (information shared by the player and the opponent) which define two learning scenarios: (1) The opponent is constrained, i.e. he provides rewards that are stochastic functions of equivalence classes defined by some model $\theta^* \in \Theta$. The regret is measured with respect to (w.r.t.) the best history-dependent strategy. (2) The opponent is arbitrary and we measure the regret w.r.t. the best strategy among all mappings from classes to actions (i.e. the best history-class-based strategy) for the best model in Θ. This allows to model opponents (case 1) or strategies (case 2) which handles finite memory, periodicity, standard stochastic bandits and other situations.

When $\Theta = \{\theta\}$, i.e. only one model is considered, we derive *tractable* algorithms achieving a *tight* regret (at time T) bounded by $\tilde{O}(\sqrt{TAC})$, where C is the number of classes of θ. Now, when many models are available, all known algorithms achieving a nice regret $O(\sqrt{T})$ are unfortunately *not tractable* and scale poorly with the number of models $|\Theta|$. Our contribution here is to provide *tractable* algorithms with regret bounded by $T^{2/3}C^{1/3}\log(|\Theta|)^{1/2}$.

1 Introduction

Designing medical treatments for patients infected by the Human Immunodeficiency Virus (HIV) is challenging due to the ability of the HIV to mutate into new viral strains that become, with time, resistant to a specific drug [8]. Thus we need to alternate between drugs. The standard formalism of *stochastic bandits* (see [17]) used for designing medical treatment strategies models each possible drug as an arm (action) and the immediate efficiency of the drug as a reward. In this setting, the rewards are assumed to be i.i.d., thus the optimal strategy is constant in time. However in the case of adapting viruses, like the HIV, no constant strategy (i.e., a strategy that constantly uses the same drug) is good on the long term. We thus need to design new algorithms (together with new performance criteria) to handle a larger class of strategies that may depend on the whole treatment history (i.e., past actions and rewards).

More formally, we consider a sequential decision making problem with rewards provided by a possibly adaptive opponent. The general game is defined as follows: At each time-step t, the decision-maker (or player, or agent) selects an action $a_t \in \mathcal{A}$ (where $\mathcal{A}$ is a set of $A = |\mathcal{A}|$ possible actions), and simultaneously the opponent (or adversary or environment) chooses a reward function $r_t : \mathcal{A} \mapsto [0,1]$. The agent receives the reward $r_t(a_t)$. In this paper we consider the so-called *bandit information* setting where the agent only sees the rewards of the chosen action, and not the other rewards provided by the opponent. The goal of the agent is to maximize the cumulative sum of the rewards received, i.e. choose a sequence of actions $(a_t)_{t\le T}$ that maximizes $\sum_{t=1}^{T} r_t(a_t)$.

Motivating example In order to better understand our goal, consider the following very simple problem for which no standard *bandit* algorithm achieves good cumulative rewards.

The set of actions is $\mathcal{A} = \{a, b\}$, and the opponent is defined by: $r(a) = 1$ and $r(b) = 0$ if the last action of the player is b, and $r(a) = 0$ and $r(b) = 1$ if the last action is a. Finally $r(a) = r(b) = 1$ for the first action.

In that game, playing a constant action a (or b) yields a cumulative reward of $T/2$ at time T. On the other hand, a player that would switch its actions at each round would obtain a total rewards of T, which is op-

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9. Copyright 2011 by the authors.

timal. Although this opponent is very simple, the loss of using any usual multi-armed bandit algorithm (such as UCB [3] and EXP3 [4]) instead of this simple switching strategy will be linear in T.

Adaptive opponents In this paper, we consider the setting when the opponent is *adaptive*, in the sense that the reward functions can be arbitrary measurable functions of the past common history, where by common history we mean all the observed rewards $(r_s(a_s))_{s<t}$ and actions $(a_s)_{s<t}$ played before current time t. We write $h_{<t}$ or simply h the common history up to time t, so we can write $r_t(a) = r(h_{<t}, a)$.

Due to the motivating example, we naturally want to compare to the best history-dependent strategy against the adaptive opponent, and introduce a more challenging notion of regret (see Section 2.1) than usual for that purpose. Since this may be not possible without assumptions on the opponent or the comparison strategies (see [18]), we consider some model of constraints, and thus we want to adapt to a class Θ of possible constraints. The question is: can we adapt to the (unknown) model of constraints of the opponent?

Adversarial Bandits in literature A first approach when considering adversarial bandits providing arbitrary rewards (when no constraint is put on the complexity of the adversary) is to assess the performance of the player in terms of the best strategy that is constant in time (best constant action), which defines the external regret [4, 11]. However, since this approach does not consider limitations on the strategy of the opponent with respect to the history, it can only give partial answer to the question of adaptivity to the best possible history-dependent strategy against a given opponent.

In [4], the authors extend the class of comparison strategies to piecewise constant strategies with at most S switches. The corresponding Exp3S (aka ShiftBand) algorithm achieves a regret of order $\sqrt{TSA\log(T^3A)}$, provided that T is large enough. However, against the opponent described in the previous section, the best strategy would need to switch $S = T/2$ times, thus this algorithm still suffers a linear regret compared to the optimal strategy.

The notion of internal regret (see [9]), which compares the loss of an online algorithm to the loss of a modified algorithm that consistently replaces one action by another, has been also considered in many works [12, 19, 7, 10]. In [5], the authors propose a way to convert any external regret minimization algorithm into an algorithm minimizing an extended notion of internal regret, using the so-called modifications rules that are functions $h, a \to b$, where h is the history, and a and b are actions. This enables to compare the actions selected by the algorithm to an alternative sequence and thus to assess the performance of the algorithm to other slightly perturbed algorithm. Assuming that the opponent's strategy can be described with the modification rules, then we might also see the corresponding modified regret minimization algorithm as adaptive to the opponent, in some sense. However, the proposed algorithm uses exponentially many internal variables and will not provide tight performance bounds in terms of regret w.r.t. the best history-based strategy, that we consider in Section 2.1.

On a more theoretical aspect, the work by [18] addresses the learnability problem in reactive environments (adaptive opponents). The authors introduce the notion of value-stable and recoverable environments, and show that environments satisfying such mild conditions are learnable. This also means that it is *not possible* to obtain sublinear regret w.r.t. the best strategy against any arbitrary opponent: we need to consider limitations of the opponent. Note also that the main proof of the paper by [18] is constructive, but unfortunately the would-be corresponding player is not implementable.

Tractability Since bandit algorithms are the base stone for Reinforcement Learning (RL) algorithms, it is thus important if not crucial to consider numerically efficient algorithms. The question of adaptability is not trivial because of tractability: Although the works of [5] and [18] already provide adaptive algorithms, none of them would be tractable in our setting (even with only one θ). Moreover, for a pool of possible behaviors Θ of the opponent (see Section 3), we define the Θ-regret w.r.t. the best possible strategy for the best model $\theta \in \Theta$. We then show (in Section 3) that our problem can be seen as a special instance of sleeping bandits. The best regret bounds known with tractable algorithms would be of order $\tilde{O}((TC_\Theta)^{4/5})$ (see [14]) whereas there exists a non-tractable algorithm achieving $\tilde{O}(\sqrt{TC_\Theta})$, where $C_\Theta = \sum_{\theta\in\Theta} C_\theta$ and C_θ is the complexity of model θ. If the regret of the second algorithm nicely scales with the time horizon T, both of them provide loose bounds for large $|\Theta|$. So the question is: can we design *tractable* algorithms that can *adapt* to a *large* pool of models of constraints?

Our contribution The main contribution of this paper is a new way of considering adversarial opponents. For some equivalence relation Φ on histories, we write $[h]_\Phi$ for the equivalence class of the history h w.r.t. Φ. We introduce Φ-constrained opponents that are such that the reward functions only depend on the equivalence classes of history, i.e. $r_t(a) = r([h_{<t}], a)$. Similarly, one can consider classes of strategies of the form $\mathcal{H}/\Phi \mapsto \mathcal{A}$, where $\mathcal{H}/\Phi$ is the set of equivalence classes of histories. Interestingly, such equivalence re-

lations were also introduced in [13], with the goal to build relevant equivalence relations for Reinforcement Learning. The author provides useful insights, but no performance analysis. Our model of constraints, although seemingly simple, has two main advantages: (1) the notion of Φ-regret (see Section 2) captures the regret w.r.t. such strategies and is expressive enough to handle many kinds of situations (like finite memory, periodicity, etc) and thus enables to define opponents that can be anything *from the worst possible* (fully adversarial), *to a simple stochastic* multi-armed bandit. (2) such a model leads to simple and efficient algorithms that are built directly from standard algorithms, and yet achieve significantly good performances.

The introductory Section 2 starts with a single model and provides algorithms with expected regret w.r.t. the optimal history-based strategy bounded by $O(\sqrt{TAC\log A})$, where C is a measure of the complexity (number of equivalence classes of $\mathcal{H}/\Phi$) of the opponent, and a lower bound $\Omega(\sqrt{TAC})$. This applies to the switching opponent described in the introduction. The complexity of those algorithms is C times the complexity of the standard algorithms they are built from (namely UCB and Exp3), as opposed to the complexity of order A^C for algorithms that would be derived directly from [5] in our setting. Note also that for the special case of a Φ-constrained opponent with a known model Φ, one can consider a RL point of view instead, and apply algorithms such as [16].

Our main contribution in this paper is to consider the more challenging situation where we have a pool of possible models Θ. In this case, we provide tractable algorithms with Θ-regret of order (see Section 3). $(TA)^{2/3}(C_{\theta^*}\log(T))^{1/3}\log(|\Theta|)^{1/2}$ when the opponent belongs to the pool (i.e. $\theta^*\in\Theta$, in which case we compare the performance to that of the optimal history-based strategy), and $T^{2/3}(A\overline{C}\log(A))^{1/3}\log(|\Theta|)^{1/2}$ where $\overline{C} = \max_\theta C_\theta$, when it does not (in which case we compare to the best $\mathcal{H}/\Phi_\theta$-history-class-based strategy for the best model $\theta\in\Theta$).

We finally report numerical experiments in Section 4 which compares standards algorithms for bandits (from stochastic to adversarial) (UCB, MOSS, EXP3, ShiftBand) to the algorithms proposed here.

2 Definitions and preliminary results

2.1 Model of constrained opponents

Let $\mathcal{H}$ be the set of all histories, i.e. sequences of action played and information received. Let $\Phi : \mathcal{H} \to Y$ be a given function mapping histories to an abstract space Y, and let $\mathcal{H}/\Phi$ denote the class of equivalence of histories w.r.t. the relation $h_1 \sim h_2$ if and only if $\Phi(h_1) = \Phi(h_2)$. We write also $[h]_\Phi$ (or $[h]$ when there is no ambiguity) for the equivalence class of h.

Based on an equivalence-class Φ, one can define Φ-constrained opponents, which are intuitively the opponents that are Φ-classwise stochastic:

Definition 1 *A Φ-constrained opponent is a function $f : \mathcal{H}/\Phi \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ is the set of distribution over the set $\mathcal{A}$, taking values in $[0,1]$ (i.e. we assume that all rewards belongs to the interval $[0,1]$).*

Examples: Definition 1 covers many situations:

- When $\Phi(h) = 1$ for all $h\in\mathcal{H}$, then $\mathcal{H}/\Phi$ consists of only one class, and Definition 1 reduces to a stochastic multi-armed bandit.
- When $\Phi_m : \mathcal{H} \to \mathcal{A}^m$ is $\Phi_m(h) = a_1...a_m$, where $a_1,...,a_m$ are the last sequence of m actions, this corresponds to opponents with finite short-term memory of length m. In this case, there are $|\mathcal{A}|^m$ equivalence classes. The example of the introduction corresponds to this case with $m = 1$.
- When $\Phi : \mathcal{H} \to \{0,...,m-1\}$ is defined by $\Phi(h) = |h| \bmod m$, where $|h|$ is the length of the history in term of number of time steps, it corresponds to reward functions that come from time-periodic distributions. Here, there are m different classes.

Regret against the best history-class-based strategy: If we consider a **Φ-constrained** opponent, then one can define for each class $c\in\mathcal{H}/\Phi$, and action $a\in\mathcal{A}$ the expected reward $\mu_c(a) = \mathbb{E}[r(c,a)]$. We define the expected history-class-based regret for the equivalence class defined by Φ, also called *stochastic Φ-regret*, as:

$$R_T^\Phi = \mathbb{E}\Big(\sum_{t=1}^T \max_{a\in\mathcal{A}} \mu_{[h_{<t}]}(a) - \mu_{[h_{<t}]}(a_t))\Big), \quad (1)$$

where $(a_t)_{t\le T}$ is the sequence of actions played, $h_{<t}$ is the history observed by the player up to time t, and $\max_a \mu_{[h_{<t}]}(a)$ is the best action, which respect to the expected rewards provided by the opponent, given the history-class $[h_{<t}]$.

Now, for an **arbitrary** adversary and an equivalence class Φ, one can define a (non-stochastic) regret w.r.t. the best $\mathcal{H}/\Phi$-history-class-based strategy, also called *adversarial Φ-regret*,

$$\tilde{R}_T^\Phi = \sup_{g:\mathcal{H}/\Phi\to\mathcal{A}} \mathbb{E}\Big(\sum_{t=1}^T \Big[r_t(g([h_{<t}])) - r_t(a_t)\Big]\Big), \quad (2)$$

where $g([h_{<t}])$ is the action that a strategy g would play given the history-class $[h_{<t}]$ activated at time t, where g belongs to the set of strategies that are constant per $\mathcal{H}/\Phi$-history-class, i.e. mappings $\mathcal{H}/\Phi\to\mathcal{A}$.

In both cases, the expectation is w.r.t. all sources of randomness: the possible internal randomization of the player, and the possible random rewards provided by the opponent.

Note that by definition the Φ-regret is always bigger than the external regret (i.e. w.r.t. the best constant action), and that in the case when Φ defines only one class, those two notions of regret reduce to their usual definitions in stochastic and adversarial bandits, respectively.

2.2 Upper bounds on the Φ-regret

In the case we play against a constrained opponent, we observe from the definition of the Φ-regret (1) that if we introduce $R_T(c) = \mathbb{E}\big[\sum_{t=1}^T \big(\max_a \mu_{[h_{<t}]}(a) - \mu_c(a_t)\big)\mathbb{I}_{[h_{<t}]=c}\big]$ for a class $c \in \mathcal{H}/\Phi$, then $R_T^\Phi = \sum_{c\in\mathcal{H}/\Phi} R_T(c)$. This enables us to use usual stochastic bandit algorithms, such as UCB [3], per history-class, and the resulting behavior will enable to minimize the stochastic Φ-regret.

Similarly, if we consider an arbitrary opponent, and an equivalence class Φ, by using usual adversarial bandit algorithms, such as Exp3 [4], per history-class, one can minimize the per-class regret $\mathbb{E}\big[\sum_{t=1}^T \big(r_t(g(c)) - r_t(a_t)\big)\mathbb{I}_{[h_{<t}]=c}\big]$ w.r.t. any constant-per-class strategy g, thus minimizing the adversarial Φ-regret $\tilde{R}_T^\Phi$.

The two corresponding algorithms, called respectively Φ-UCB and Φ-EXP3, are described in Figure 1 (α and η are parameters) and we report the regret upper-bounds in the next result.

Theorem 1 *In the case of a Φ-constrained opponent, using the Φ-UCB algorithm with parameter $\alpha > 1/2$, we have the distribution-dependent bound:*

$$R_T^\Phi \leq \sum_{c\in\mathcal{H}/\Phi;\mathbb{E}(I_c(T))>0}\ \sum_{a\in\mathcal{A};\Delta_c(a)>0} \frac{4\alpha\log(T)}{\Delta_c(a)} + \Delta_c(a)c_\alpha$$

where $I_c(T) = \sum_{t=1}^T \mathbb{I}_{[h_{<t}]=c}$, the per-class gaps $\Delta_c(a) \overset{\text{def}}{=} \sup_{b\in\mathcal{A}} \mu_c(b) - \mu_c(a)$, and the constant $c_\alpha = 1 + \frac{4}{\log(\alpha+1/2)}\big(\frac{\alpha+1/2}{\alpha-1/2}\big)^2$. We also have a distribution-free bound (i.e. which does not depend on the gaps):

$$R_T^\Phi \leq \sqrt{TA\overline{C}\Big(4\alpha\log(T) + c_\alpha\Big)}$$

where $\overline{C} = |\{c \in |\mathcal{H}/\Phi|; \mathbb{E}(I_c(T)) > 0\}|$ is the number of classes that may be activated.

Now, in the case of an arbitrary opponent, using Φ-Exp3 algorithm, we have:

$$\tilde{R}_T^\Phi \leq \frac{3}{\sqrt{2}}\sqrt{T\overline{C}A\log(A)}.$$

The proof of these statements is reported in the supplementary material and directly derives from the analysis detailed in [6] and the previous remarks. Note that one can use other bandit algorithms (such as UCB-V [2], MOSS [1]) and derive straightforwardly the corresponding result for the Φ-regret.

```
For each round $t = 1, 2, \dots, T$
    (1) Define $\hat{\mu}_{t,c}(a) = \frac{1}{I_c(t-1,a)} F_{t-1}^c(a)$, where $F_t^c(a) = \sum_{s=1}^t r_s(a)\mathbb{I}_{[h_{<s}]=c}\mathbb{I}_{a_s=a}$ and $I_c(t,a) = \sum_{s=1}^t \mathbb{I}_{[h_{<s}]=c}\mathbb{I}_{a_s=a}$.
    (2) Define $\tilde{\mu}_{t,c}(a) = \hat{\mu}_{t,c}(a) + \sqrt{\frac{\alpha\log(I_c(t))}{I_c(t-1,a)}}$.
    (3) Compute $c_t = \Phi(h_{<t})$.
    (4) Play $a_t \in \text{argmax}_{a\in\mathcal{A}} \tilde{\mu}_{t,c_t}(a)$
```

```
Initialization: Define $\forall a \in \mathcal{A}\quad \xi_1(a) = \frac{1}{A}$
For each round $t = 1, 2, \dots, T$
    (1) Play $a_t \sim \xi_t$, observe $r_t(a_t)$.
    (2) Define $\tilde{l}_t^c(a) = \frac{1-r_t(a_t)}{\xi_t(a)}\mathbb{I}_{a_t=a}\mathbb{I}_{c=[h_{<t}]}$.
    (3) Define $w_{t+1}^c(a) = \exp(-\eta\sum_{s=1}^t \tilde{l}_s^c(a))$.
    (4) Compute $c_{t+1} = \Phi(h_{<t+1})$.
    (5) Define $\xi_{t+1}(a) = \frac{w_{t+1}^{c_{t+1}}(a)}{\sum_a w_{t+1}^{c_{t+1}}(a)}$.
```

Figure 1: Φ-UCB (top) and Φ-Exp3 (down) algorithms.

2.3 Lower bounds on the Φ-regret

We now derive lower bounds on the Φ-regret to show that the previous upper bounds are tight.

Intuitively, on each class c, one may suffer a regret of order $\sqrt{I_c(T)A}$, where $I_c(T)$ is the number of times class c is visited. Now, since the way classes are "visited" depends on the structure of the game and the strategy of both the player and the opponent, those classes cannot be controlled by the player only. Thus we show that there always exist an environment such that whatever the strategy of the player is, a particular opponent will lead to visit all history-classes uniformly in expectation.

We consider here, for a given class function Φ, players that may depend on Φ and opponents that may depend both on Φ and on the player. Then we consider the worst opponent for the best player over the worst class-function Φ of given complexity (expressed in terms of number of classes C of $\mathcal{H}/\Phi$). The following result easily follows from [6].

Theorem 2 *Let* sup *represents the supremum taken over all Φ-constrained opponents and* inf *the infimum over all players, then the stochastic Φ-regret is lower-*

bounded as:

$$\sup_{\Phi;|\mathcal{H}/\Phi|=C} \inf_{algo} \sup_{\Phi-opp} R_T^\Phi \geq \frac{1}{20}\sqrt{TAC}.$$

Let sup *represents the supremum taken over all possible opponents, then the adversarial* Φ*-regret is lower-bounded as:*

$$\sup_{\Phi;|\mathcal{H}/\Phi|=C} \inf_{algo} \sup_{opp} \tilde{R}_T^\Phi \geq \frac{1}{20}\sqrt{TAC}.$$

3 Playing against an opponent using a pool of models

After this introductory section, we now turn to the main challenge of this paper. When playing against a given opponent, its model of constraints Φ may not be known. It is thus natural to consider several equivalence relations defined by a pool of class functions (models) $\Phi_\Theta = (\Phi_\theta)_{\theta\in\Theta}$, and that the opponent plays with some model induced by some Φ^*. We consider two cases: either $\Phi^* = \Phi_{\theta^*} \in \Phi_\Theta$, i.e. the opponent is a Φ_{θ^*}-constrained opponent with $\theta^* \in \Theta$, or the opponent is arbitrary, and we will compare our performance to that of the best model in Θ.

We define accordingly two notions of regret: If we consider a Φ^*-constrained opponent, where $\Phi^* \in \Phi_\Theta$, then one can define the so-called *stochastic* Φ_Θ*-regret* as:

$$R_T^\Theta = \mathbb{E}\Big(\sum_{t=1}^{T} \max_{a\in\mathcal{A}} \mu_{[h_{<t}]*}(a) - \mu_{[h_{<t}]}(a_t))\Big). \qquad (3)$$

where $[h_{<t}]*$ is the history-class used by the opponent.

Now, for an arbitrary opponent and a pool of equivalence classes Φ_Θ, we define a regret w.r.t. the best $\mathcal{H}/\Phi_\theta$-history-class-based strategy for the best model $\theta \in \Theta$, also called *adversarial* Φ_Θ*-regret*:

$$\tilde{R}_T^\Theta = \sup_{\theta\in\Theta} \sup_{g:\mathcal{H}/\Phi_\theta\to\mathcal{A}} \mathbb{E}\Big(\sum_{t=1}^{T}\Big[r_t(g([h_{<t}]_\theta)) - r_t(a_t)\Big]\Big), \qquad (4)$$

where the class $[h_{<t}]_\theta$ corresponds to the model θ.

Tractability This problem can be seen as a Sleeping bandits ([15, 14]) with stochastic availability and adversarial rewards. Indeed, by considering each class c in each model θ, we get a total of $C_\Theta = \sum_{\theta\in\Theta} C_\theta$ experts. Now at each time step, only one class per model is awake, and thus the best awake expert changes with time. Recasting this problem in a usual bandit setting where the best expert is constant over time requires considering the $C_\Theta!$ possible rankings (see [15]), each ranking being now seen as an expert. Running Exp4 algorithm on top of this new experts would give a sleeping-bandit regret (and thus a Φ_Θ-regret) of order $O(\sqrt{TA\log(C_\Theta!)}) = O(\sqrt{TAC_\Theta\log(C_\Theta)})$. Unfortunately this algorithm is intractable and the bound is very loose when the number of models is large. In [14], they proposed a (tractable) algorithm that would achieve in our setting a regret bounded by $O((TC_\Theta)^{4/5}\log(T))$.

We now describe tractable algorithms with regret upper-bounded by $O(T^{2/3}\log(|\Theta|)^{1/2})$ for both the stochastic and adversarial Φ_Θ-regret, which improves upon previous bounds for our setting.

EXP4/UCB and EXP4/EXP3 algorithms: A natural approach is to consider each model $\theta \in \Theta$ as one expert defined by a equivalence function Φ_θ and then run the Exp4 meta-algorithm (see [4]) to select an action based on the recommendations of all experts. More precisely, at each time t, the meta algorithm plays a_t according to a distribution $q_t(\cdot) = \sum_\theta p_t(\theta)\xi_t^\theta(\cdot)$ which is a mixture of distributions ξ_t^θ that each expert θ assigns to each action, weighted by a distribution $p_t(\theta)$ over the set of experts Θ. Figure 2 describes the Exp4 algorithm (see [4] for more details) using a mixing parameter $\gamma > 0$.

```
  Initialization: Define $\forall\theta \in \Theta$,  $p_1(\theta) = \frac{1}{|\Theta|}$.
For each round $t = 1, 2, \dots, T$,
(1) Define $q_t(a) = (1-\gamma)\sum_{\theta\in\Theta} p_t(\theta)\xi_t^\theta(a) + \frac{\gamma}{A}$.
(2) Draw $a_t \sim q_t$, and observe $r_t(a_t)$.
(3) Define $\tilde{l}_t(a) = \frac{1-r_t(a_t)}{q_t(a)}\mathbb{I}_{a_t=a}$.
(4) Define $g_t(\theta) = \sum_a \xi_t^\theta(a)\tilde{l}_t(a)$.
(5) Define $w_{t+1}(\theta) = \exp(-\gamma\sum_{s=1}^t g_t(\theta)/A)$.
(6) Define $p_{t+1}(\theta) = \frac{w_{t+1}(\theta)}{\sum_\theta w_{t+1}(\theta)}$.
```

Figure 2: The Exp4 meta algorithm

In [4] the authors relate the performance of the meta algorithm to that of any individual expert (see Theorem 7.1 in [4]). However, it is not obvious to build an algorithm for each individual expert Φ_θ that will minimize its Φ_θ-regret. Indeed, the actions played by the meta algorithm *differ* from the ones that would have been played by each specific expert θ. This means that for each expert, not only we have a limited (bandit) information w.r.t. the reward function, but also *each expert does not see the reward of its recommended action.* This results in individual expert algorithms with poorer regret bounds than in the single model case described in the previous section (for which one observes the reward of the chosen action).

We provide two algorithms based respectively on UCB and Exp3, that may be used by each individual expert θ:

- Φ_θ-UCB is defined as before (see Figure 1), except that instead of step (4) we define ξ_t^θ as a Dirac distribution at the recommended action $\operatorname{argmax}_{a\in\mathcal{A}} \tilde{\mu}_{t,c_t}(a)$.
- Φ_θ-Exp3 is defined as before (see Figure 1), except that in step (1), no action is drawn from ξ_t (since the meta algorithm chooses $a_t \sim q_t$), and step (2) is replaced by: $\tilde{l}_t^c(a) = \frac{1-r_t(a_t)}{q_t(a)}\mathbb{I}_{a_t=a}\mathbb{I}_{c=[h_{<t}]_\theta}$ (i.e. we re-weight by using the probability $q_t(a)$ of the meta algorithm instead of the probability $\xi_t^\theta(a)$ of the individual expert θ).

Regret bounds (proved in Appendix A.2,A.3) of the meta algorithm Exp4 combined respectively with individual algorithms UCB and Exp3 (called respectively Exp4/UCB and Exp4/Exp3) are given below.

Theorem 3 *Assume that we consider a Φ^*-constrained opponent with $\Phi^* \in \Phi_\Theta$, then the stochastic Φ_Θ-regret of Exp4/UCB is bounded as:*

$$R_T^\Theta = O\Big((TA)^{2/3}(\overline{C}\log(T))^{1/3}\log(|\Theta|)^{1/2}\Big),$$

where $\overline{C} = |\mathcal{H}/\Phi^*|$ is the number of classes of the model Φ^* of the opponent. Now, for any opponent, the adversarial Φ_Θ-regret of Exp4/Exp3 is bounded as

$$\tilde{R}_T^\Theta = O\Big(T^{2/3}(A\overline{C}\log(A))^{1/3}\log(|\Theta|)^{1/2}\Big),$$

where $\overline{C} = \max_{\theta\in\Theta}|\mathcal{H}/\Phi_\theta|$ is the maximum number of classes for models $\theta \in \Theta$.

Note that, like in EXP4, we obtain a logarithmic dependence on $|\Theta|$ since playing an action that has been chosen from a mixture of the probability distributions (over actions) of all models yields a reward which provides information about all the models.

4 Experiments

We illustrate our approach with three different adaptive opponents and compare the results of standard algorithms to the algorithms described here using two measures of performance: the Φ-regret, and the external regret.

We consider only two actions $\mathcal{A} = \{a, b\}$, and fix the time horizon at $T = 500$. The three considered opponents have finite short-term memory of length $m = 0, 1, 2$ respectively, i.e. are Φ_m-constrained opponents in the sense of Definition 1. More precisely, the reward distributions are Bernoulli, and the opponents are

- O_0 is a simple stochastic bandit (no memory). We choose $\mu(a) = 0.4$ and $\mu(b) = 0.7$
- O_1 provides a mean reward 0.8 when the action changes at each step, and 0.3 otherwise,
- O_2 provides a mean reward 0.8 when the action changes every two steps and 0.3 otherwise.

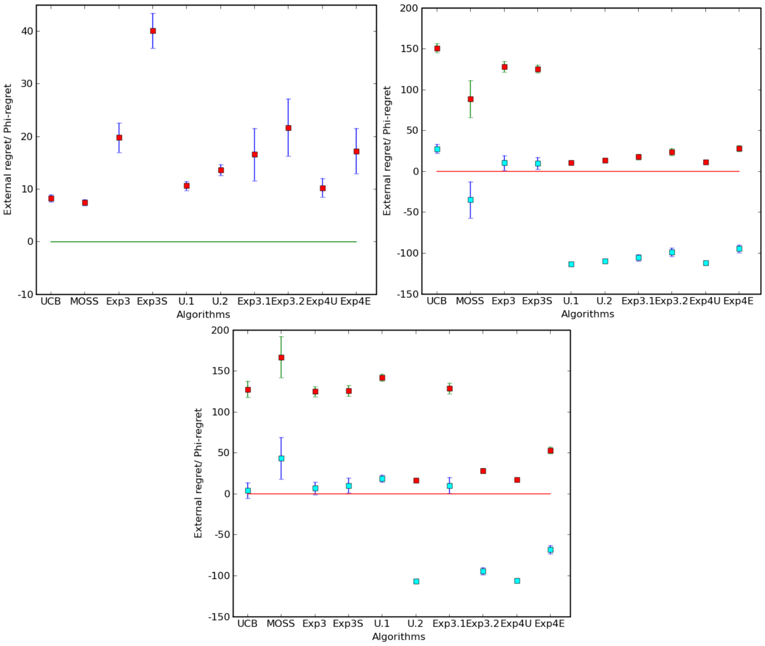

Figure 3: Regret w.r.t. the best history-dependent strategy (red) and best constant strategy (cyan) for 3 opponents. All experiments have been averaged over 50 trials.

Each plot of Figure 3 corresponds to one opponent (O_0 is left, O_1 right, and O_2 is bottom). In each plot, we represent the external regret (cyan) and Φ-regret (red) obtained for several algorithms. From left to right, the first four algorithms are UCB, MOSS, Exp3 and ShiftBand. The next four correspond respectively to Φ_1-UCB, Φ_2-UCB, Φ_1-Exp3, and Φ_2-Exp3 algorithms (i.e. versions of UCB and Exp3 with memory of length 1 and 2). Note that Φ_0-UCB (resp. Exp3) is just UCB (resp. Exp3). The last two algorithms correspond to the Exp4/UCB (resp. Exp4/Exp3), i.e. meta algorithm Exp4 run on top of Φ_m-UCB (resp Φ_m-Exp3) algorithms, for $m = 0, 1, 2$) as defined in Section 3.

The last two algorithms do not know the model of constraint corresponding to the opponent they are facing and still, they clearly outperform other standard algorithms (with frankly negative external regret) for the two adapting opponents (second and third). This clear improvement appears also when the model considered by the algorithm is *more* complex than that of the opponent (e.g. Φ_2-UCB facing opponent 1). On the other hand, the reverse is false (Φ_1-UCB and Φ_1-Exp facing opponent 2) since a algorithm using a piece of history of length 1 cannot play well against an opponent with memory 2.

Future works

We do not know whether in the case of a pool of models Φ_Θ, there exist tractable algorithms with Φ_Θ-regret better that $T^{2/3}$ with log dependency w.r.t. $|\Theta|$. Here we have used a meta Exp4 algorithm, but we could have used other meta algorithms using a mixture $q_t(a) = \sum_\theta p_t(\theta)\xi_t^\theta(a)$ (where the p_t are internal weights of the meta algorithm). However, when computing the approximation term of the best model θ^* by models $\theta \in \Theta$ (see the supplementary material), it seems that the Φ_Θ-regret cannot be strongly reduced without making further assumptions on the structure of the game, since in general the mixture distribution q_t may not converge to the distribution ξ_t^θ proposed by the best model $\theta \in \Theta$. However the question remains open.

A Some Technical proofs.

In order to relate the cumulative reward of the Exp4 algorithm to the one of the best expert on top of which it is run, we state the following simple Lemma. Note that in our case, since the ξ_t^θ are not fixed in advance but are random variables, we can not apply the original result of [4] for fixed expert advises, but need to adapt it. The proof easily follows from the original proof of Theorem 7.1 in [4], and is reported in the supplementary material.

Lemma 1 *For any $\gamma \in (0,1]$, for any family of experts which includes the uniform expert such that all expert advises are adapted to the filtration of the past, one has*

$$\max_\theta \sum_{t=1}^T \mathbb{E}_{a_1,\dots,a_{t-1}}(\mathbb{E}_{a\sim\xi_t^\theta}(r_t(a))) - \mathbb{E}_{a_1,\dots,a_T}(\sum_{t=1}^T r_t(a_t)) \le (e-1)\gamma T + \frac{A\log(|\Theta|)}{\gamma}.$$

A.1 The rebel bandit setting

We now introduce the setting of Rebel bandits that may have its own interest. It will be used to compute the model-based regret of the Exp4 algorithm. In this setting, we consider that at time t the player θ proposes a distribution of probability ξ_t^θ over the arms, but he actually receives the reward corresponding to an action drawn from another distribution, q_t, the distribution of probability proposed by the meta algorithm.

Following (4), we define the best model of the pool:

$$\theta^* = \operatorname*{argmax}_{\theta\in\Theta} \sup_{g:\mathcal{H}/\Phi\to\mathcal{A}} \mathbb{E}\Big(\sum_{t=1}^T \Big[r_t(g([h_{<t}])) - r_t(a_t)\Big]\Big).$$

We then define for any class $c \in \mathcal{H}/\Phi_{\theta^*}$, the action $a_c^* \overset{\text{def}}{=} \operatorname{argmax}_a \mu_c(a)$ that corresponds to the best history-class-based strategy. We now analyze the (Φ-constrained) Exp3 and UCB algorithms in this setting and bound the corresponding rebel-regret:

Definition 2 *(Rebel regret) The Rebel-regret of the algorithm that proposes at time t the distribution ξ_t^θ but in the game where the action $a_t \sim q_t$ is played instead is:*

$$\mathcal{R}_T^q(\theta) = \sum_{t=1}^T \mathbb{E}_{a_1,..,a_{t-1}}\Big(r_t(a^*_{[h_{<t}]_{\theta^*}}) - \mathbb{E}_{a\sim\xi_t^\theta}(r_t(a))\Big).$$

A.2 Φ-Exp3 in the Rebel bandit setting

We now consider using Exp4 on top of Φ-contrained algorithms. We first use the experts Φ_θ-Exp3 for $\theta\in\Theta$ with a slight modification on the definition of the function $\tilde{l}_t^c(a)$. Indeed since the action a_t are drawn according to the meta algorithm and not Φ_θ-Exp3, we redefine $\tilde{l}_t^c(a) = \frac{1-r_t(a)}{q_t(a)}\mathbb{I}_{a_t=a}\mathbb{I}_{[h_{<t}]_\theta=c}$ so as to get unbiased estimate of $r_t(a)$ for all a. We now provide a bound on the Rebel-regret of the Φ^*-Exp3 algorithm.

Theorem 4 *The Φ_{θ^*}-Exp3 algorithm in the Rebel bandit setting where $q_t(a) \ge \delta$ for all a, and choosing the parameter $\eta^\theta_{t_c^\theta(i)} = \sqrt{\frac{\delta\log(A)}{i}}$ satisfies $\mathcal{R}_T^q(\theta^*) \le 2\sqrt{\frac{T\overline{C}\log A}{\delta}}$.*

Proof: The proof is in six steps and mainly follows the proof in Section 2.1 of [6] that provides a bound on the regret of Exp3 algorithm.

Since we only consider the model θ^*, we will simply refer to it as θ and also write c_t for $[h_{<t}]_{\theta^*}$ to avoid cumbersome notations.

Step 1. Rewrite the regret term to make appear the actions a_t chosen by the meta algorithm at time t. By definition of $\tilde{l}^\theta_{t,c_\theta}(a)$ we have $\mathbb{E}_{a_t\sim q_t}(\tilde{l}^\theta_{t,c_t}(a)) = 1 - r_t(a)$, thus we get:

$$\mathcal{R}_T^q(\theta^*) = \sum_{t=1}^T \mathbb{E}_{a_1,..,a_{t-1}}\big[\mathbb{E}_{a_t\sim q_t}(\mathbb{E}_{a\sim\xi_t^\theta}(\tilde{l}^\theta_{t,c_t}(a))) - \tilde{l}^\theta_{t,c_t}(a^*_{c_t})\big].$$

Step 2. Decompose the term $\mathbb{E}_{a\sim\xi_t^\theta}(\tilde{l}^\theta_{t,c_t}(a))$ in order to use the definition of ξ_t^θ. Indeed, for $\phi(x) \overset{\text{def}}{=} \frac{1}{\eta_t^\theta}\log\mathbb{E}_{a\sim\xi_t^\theta}\exp(x)$, following the technique described in Section 2.1 of [6], we have:

$$\mathbb{E}_{a\sim\xi_t^\theta}(\tilde{l}^\theta_{t,c_t}(a)) = \phi\big(-\eta_t^\theta(\tilde{l}^\theta_{t,c_t}(a) - \mathbb{E}_{b\sim\xi_t^\theta}(\tilde{l}^\theta_{t,c_t}(b)))\big) - \phi(-\eta_t^\theta\tilde{l}^\theta_{t,c_t}(a))$$

Now using the fact that $\log x \leq x - 1$ and $\exp(-x) - 1 + x \leq x^2$, $\forall x \geq 0$, the first term on the right hand of (5) is bounded by: $\frac{\eta_t^\theta}{2}\mathbb{E}_{a\sim\xi_t^\theta}(\tilde{l}_{t,c_t}^\theta(a)^2)$

Thus, considering that $\xi_t^\theta(a) = \frac{\exp(-\eta_t^\theta \sum_{s=1}^{t-1} \tilde{l}_{s,c_t}^\theta(a))}{\sum_a \exp(-\eta_t^\theta \sum_{s=1}^{t-1} \tilde{l}_{s,c_t}^\theta(a)}$, we can introduce the quantity $\Psi_t^\theta(\eta, c) = \frac{1}{\eta}\log(\frac{1}{A}\sum_a \exp(-\eta \sum_{s=1}^{t} \tilde{l}_{s,c}^\theta(a)))$ so that the second right term of (5) is $\Psi_{t-1}^\theta(\eta_t^\theta, c_t) - \Psi_t^\theta(\eta_t^\theta, c_t)$. Thus we deduce that:

$$\mathcal{R}_T^q(\theta^*) \leq \sum_{t=1}^{T} \mathbb{E}_{a_1,..,a_{t-1}}\Big[\mathbb{E}_{a_t\sim q_t}(\frac{\eta_t^\theta}{2}(1 - r_t(a_t))^2 \frac{\xi_t^\theta(a_t)}{q_t^2(a_t)}) \\ + \mathbb{E}_{a_t}\big(\Psi_{t-1}^\theta(\eta_t^\theta, c_t) - \Psi_t^\theta(\eta_t^\theta, c_t)\big) - \mathbb{E}_{a_t}\tilde{l}_{t,c_t}^\theta(a_{c_t}^*)\Big],$$

where we have replace $\tilde{l}_{t,c_t}^\theta(a)$ by its definition.

Step 3. Now we consider the first term in the right hand side of previous equation, which is bounded by:

$$\mathbb{E}_{a_t\sim q_t}((1 - r_t(a_t))^2 \frac{\xi_t^\theta(a_t)}{q_t^2(a_t)}) \leq \sum_a \frac{\xi_t^\theta(a)}{q_t(a)} \leq \frac{1}{\delta}.$$

Step 4. Introduce the equivalence classes. We now consider the second term defined with Ψ functions. Let us introduce the notations $I_c^\theta(t) = \sum_{s=1}^{t} \mathbb{I}_{c=[h_{<s}]_\theta}$ and $t_c^\theta(i) = \min\{t; I_c^\theta(t) = i\}$. Thus we can write:

$$\sum_\theta \sum_{t=1}^{T}(\Psi_{t-1}^\theta(\eta_t^\theta, [h_{<t}]_\theta) - \Psi_t^\theta(\eta_t^\theta, [h_{<t}]_\theta)) = \\ \sum_\theta \sum_{c\in\theta} \sum_{i=1}^{I_c^\theta(T)} \Psi_{t_c^\theta(i)-1}^\theta(\eta_{t_c^\theta(i)}^\theta, c) - \Psi_{t_c^\theta(i)}^\theta(\eta_{t_c^\theta(i)}^\theta, c) = \\ \sum_\theta \sum_{c\in\theta} \Big(\sum_{i=1}^{I_c^\theta(T)-1} (\Psi_{t_c^\theta(i)}^\theta(\eta_{t_c^\theta(i)+1}^\theta, c) - \Psi_{t_c^\theta(i)}^\theta(\eta_{t_c^\theta(i)}^\theta, c))\Big) \\ -\Psi_{t_c^\theta(I_c^\theta(T))}^\theta(\eta_{t_c^\theta(I_c^\theta(T))}^\theta, c).$$

Now, by definition of Ψ_t^θ, we also have:

$$-\Psi_{t_c^\theta(I_c^\theta(T))}^\theta(\eta_{t_c^\theta(I_c^\theta(T))}^\theta, c) = \frac{\log A}{\eta_{t_c^\theta(I_c^\theta(T))}^\theta} - \\ \frac{1}{\eta_{t_c^\theta(I_c^\theta(T))}^\theta} \log(\frac{1}{A}\sum_a \exp(-\eta_{t_c^\theta(I_c^\theta(T))}^\theta \sum_{s=1}^{t_c^\theta(I_c^\theta(T))} \tilde{l}_{s,c}^\theta(a))),$$

which is less than $\frac{\log A}{\eta_{t_c^\theta(I_c^\theta(T))}^\theta} + \sum_{s=1}^{t_c^\theta(I_c^\theta(T))} \tilde{l}_{s,c}^\theta(a)$ for any given $a = a_c^{\theta^*}$ (we remind that $\theta = \theta^*$), in particular, we can use the optimal action a_c^* when $c = [h_{<t}]_{\theta^*}$.

Step 5. Remark that $\Psi_t^\theta(\cdot, c)$ is increasing for all θ, c. Indeed, we can show that

$$\frac{\partial}{\partial\eta}\Psi^\theta(\eta, c) = \frac{1}{\eta^2}KL(p_{t,c}^\eta, \pi),$$

where π is the uniform distribution over the arms, and $p_{t,c}^\eta(a) = \frac{\exp(-\eta \sum_{s=1}^{t-1} \tilde{l}_{s,c_t}^\theta(a))}{\sum_a \exp(-\eta \sum_{s=1}^{t-1} \tilde{l}_{s,c_t}^\theta(a)}$.

Step 6. Now since $\eta_{t_c^\theta(i)}^\theta \leq \eta_{t_c^\theta(i)+1}^\theta$, and $\Psi_{t_c^\theta(i)}^\theta(\cdot, c)$ is increasing, we combine the results of each step to deduce that:

$$\mathcal{R}_T^q(\theta^*) \leq \mathbb{E}\Big(\sum_c \sum_{i=1}^{I_c^\theta(T)} \frac{\eta_{t_c^\theta(i)}^\theta}{2\delta} + \frac{\log A}{\eta_{t_c^\theta(I_c^\theta(T))}^\theta}\Big).$$

Since we choose $\eta_{t_c^\theta(i)}^\theta = \sqrt{\frac{\delta\log(A)}{i}}$, we get:

$$\mathcal{R}_T^q(\theta^*) \leq 2\mathbb{E}\Big(\sum_c \sqrt{\frac{I_c^\theta(T)\log A}{\delta}}\Big) \leq 2\sqrt{\frac{T\overline{C}\log A}{\delta}}.$$

□

We now combine Lemma 1 and Theorem 4 using Exp4 meta algorithm with $\delta = \frac{\gamma}{A}$ to get the final bound:

Theorem 5 *For any opponent, the adversarial Φ_Θ-regret of Exp4/Exp3 is bounded as*

$$\tilde{R}_T^\Theta = O(T^{2/3}(A\overline{C}\log(A))^{1/3}\log(|\Theta|)^{1/2}),$$

where $\overline{C} = \max_{\theta\in\Theta}|\mathcal{H}/\Phi_\theta|$ is the maximum number of classes for models $\theta \in \Theta$.

Proof: By Lemma 1 and Theorem 4, we get

$$\tilde{R}_T^\Theta \quad \leq \quad 2\sqrt{\frac{TA\overline{C}\log A}{\gamma}} + 2\gamma T + \frac{A\log(|\Theta|)}{\gamma}.$$

□

A.3 Φ-UCB in the Rebel-bandit setting

Similarly, a bound on the Rebel-regret of the Φ^*-UCB algorithm can be derived assuming that we consider a Φ^*-constrained opponent with $\Phi^* = \Phi^{\theta^*} \in \Phi_\Theta$. The proof of the following statement is reported in the supplementary material.

Theorem 6 *The Φ_{θ^*}-UCB algorithm in the Rebel bandit setting where $q_t(a) \geq \delta$ for all a, and provided $\alpha > 1/2$, satisfies*

$$\mathcal{R}_T^q(\theta^*) \leq \sum_{c\in\mathcal{H}/\Phi^*} \sum_{a\neq a_c^*} \Delta_c(a)\Big[\frac{2\alpha\log(T)}{\Delta_c(a)^2\delta} + \sqrt{\frac{\pi\delta\Delta_c(a)^2}{32\alpha\log T}} + c_\alpha\Big]$$

We also have the distribution-free bound:

$$\mathcal{R}_T^q(\theta^*) \leq \sqrt{TC^*A}\sqrt{\frac{4\alpha\log(T)}{\delta} + c_\alpha + \sqrt{\frac{\pi\delta}{32\alpha\log(T)}}}.$$

This enables us to deduce the first part of Theorem 3, following the same method as Theorem 5 but for the stochastic Φ_Θ-regret of Exp4/UCB.

References

[1] J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *22nd annual conference on learning theory*, 2009.

[2] J.Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 2008.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning Journal*, 47:235–256, 2002.

[4] Peter Auer, Nicolò Cesa-bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32:2002, 2002.

[5] Avrim Blum, Yishay Mansour, and Ron Meir. From external to internal regret. In *In COLT*, pages 621–636, 2005.

[6] S. Bubeck. *Bandits Games and Clustering Foundations*. PhD thesis, Université Lille 1, 2010.

[7] Nicolò Cesa-Bianchi and Gábor Lugosi. Potential-based algorithms in on-line prediction and game theory. *Mach. Learn.*, 51(3):239–261, 2003.

[8] Damien Ernst, Guy-Bart Stan, Jorge Goncalves, and Louis Wehenkel. Clinical data based optimal sti strategies for hiv; a reinforcement learning approach. In *Machine Learning Conference of Belgium and The Netherlands (Benelearn)*, pages 65–72, 2006.

[9] Dean Foster and Rakesh Vohra. Asymptotic calibration. *Biometrika*, 85:379–390, 1996.

[10] Dean P. Foster and Rakesh Vohra. Regret in the on-line decision problem. *Games and Economic Behavior*, 29(1-2):7 – 35, 1999.

[11] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *EuroCOLT '95: Proceedings of the Second European Conference on Computational Learning Theory*, pages 23–37, London, UK, 1995. Springer-Verlag.

[12] Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

[13] Marcus Hutter. Feature reinforcement learning: Part I: Unstructured MDPs. *Journal of Artificial General Intelligence*, 1:3–24, 2009.

[14] Varun Kanade, H. Brendan McMahan, and Brent Bryan. Sleeping experts and bandits with stochastic action availability and adversarial rewards. In *AISTATS*.

[15] Robert D. Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. In *Conference on Learning Theory*, July 2008.

[16] Ronald Ortner. Online regret bounds for markov decision processes with deterministic transitions. In *ALT '08: Proceedings of the 19th international conference on Algorithmic Learning Theory*, pages 123–137, Berlin, Heidelberg, 2008. Springer-Verlag.

[17] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society*, 58:527–535, 1952.

[18] Daniil Ryabko and Marcus Hutter. On the possibility of learning in reactive environments with arbitrary dependence. *Theor. Comput. Sci.*, 405(3):274–284, 2008.

[19] Gilles Stoltz. Incomplete information and internal regret in prediction of individual sequences, 2005.