Supplementary material for the paper: "Adaptive Bandits: Towards the best history-dependent strategy"

Odalric-Ambrym Maillard
INRIA Lille Nord-Europe

Rémi Munos
INRIA Lille Nord-Europe

Abstract

In this document, we detail further some technical proofs not covered in the paper corresponding to this supplementary material.

1 Playing against an opponent using a known model

1.1 Regret upper bounds against the best history-class-based strategy

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 $\Delta_c(a) \stackrel{\text{def}}{=} \mu_c(a_c^*) - \mu_c(a)$, and the constant $c_\alpha = 1 + \frac{4}{\log(\alpha+1/2)}(\frac{\alpha+1/2}{\alpha-1/2})^2$. 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 during the run.

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

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

Proof: **Φ-UCB**: The distribution-dependent bound for Φ-UCB is a direct application of the result of [2] for the algorithm UCB about $\tau_a(t) \stackrel{\text{def}}{=} \sum_{s=1}^t \mathbb{I}_{a_s=a}$ where a_t is played by UCB, that states that $\mathbb{E}(\tau_a(t)) \leq \frac{4\alpha\log(t)}{\Delta_c^2(a)} + c_\alpha$. Indeed, we use the fact that $R_T^\Phi = \sum_{c\in\mathcal{H}/\Phi} R_T(c)$ and thus remark that when a class c is visited, then we play according to a UCB algorithm for this class.

Thus, for the distribution-free bound, we have:

$$\begin{aligned} R_T^\Phi &= \sum_c \sum_a \Delta_c(a)\mathbb{E}(\tau_a(I_c(T))) \\ &\leq \sum_c \sum_a \sqrt{\mathbb{E}(\tau_a(I_c(T)))}\sqrt{4\alpha\log(T)+c_\alpha} \\ &\leq \sum_c \sqrt{\mathbb{E}(I_c(T))}\sqrt{A}\sqrt{4\alpha\log(T)+c_\alpha} \\ &\leq \sqrt{T\overline{C}A}\sqrt{4\alpha\log(T)+c_\alpha}, \end{aligned}$$

where we used that $\sum_a \tau_a(s) = s$ for all s, and $\sum_c I_c(T) = T$, and the Cauchy-Schwartz inequality twice.

Φ-Exp3: The bound for Φ-Exp3 follows from the bound of the anytime version of the Exp3 algorithm. Indeed we have

$$\tilde{R}_T^\Phi \leq \sum_c \mathbb{E}(\frac{A}{2}\sum_{i=1}^{I_c(T)} \eta_i^c + \frac{\log(A)}{\eta_{I_c(T)}^c});$$

we deduce the bound by setting $\eta_i^c = \sqrt{\frac{2\log A}{Ai}}$. □

1.2 Lower bounds on the regret

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

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

Now, let sup *represents the supremum taken over all possible opponents and* inf *the infimum over all forecasters, then the adversarial Φ-regret is lower-bounded as:*

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

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

Proof: Let us fix the horizon T and the number of classes C. We consider the opponent defined using the specific class-function Φ such that each class c is periodically visited every C time steps, thus T/C times. Note that $T = \frac{T}{C}C$ and that this is intuitively the opponent that makes the algorithm switch between classes the most.

Now we define more precisely the rewards output by the opponent. Let us consider the stochastic bandits such that for each class c, one arm a_c is a Bernoulli $B((1+\epsilon_c)/2)$, and all others are $B((1-\epsilon_c)/2)$.

Then by application of Lemma 2.2 in [2], for ϵ_c of order $\sqrt{\frac{A}{s}}$, we have in the Bandit information setting the following inequality:

$$\sup_{a_c} \sum_{t=1}^{s} \mu_c(a_c) - \mu_c(a_t) \geq s\epsilon_c\Big(1 - \frac{1}{A} - \sqrt{\frac{s\epsilon_c}{2A}\log(\frac{1+\epsilon_c}{1-\epsilon_c})}\Big).$$

Thus with the notations $I_c(T) = \sum_{t=1}^{T} \mathbb{I}_{c=[h_{<t}]}$ and $t_c(i) = \min\{t; I_c(t) \geq i\}$, we deduce that:

$$\sup_{(a_c)_c} \sum_c \sum_{t=1}^{T} (\mu_c(a_c) - \mu_c(a_t))\mathbb{I}_{c=[h_{<t}]} =$$
$$\sum_c \sup_{a_c} \sum_{i=1}^{I_c(T)} (\mu_c(a_c) - \mu_c(a_{t_c(i)})) \geq$$
$$\sum_c I_c(T)\epsilon_c\Big(1 - \frac{1}{A} - \sqrt{\epsilon_c \log(\frac{1+\epsilon_c}{1-\epsilon_c})}\sqrt{\frac{I_c(T)}{2A}}\Big).$$

Since the a_c are chosen by the opponent such that each class is visited exactly $I_c(T) = T/C$ times, then we deduce that the Φ-pseudo-regret is lower-bounded as:

$$\sup_{(a_c)_c} \sum_c \sum_{t=1}^{T} (\mu_c(a_c) - \mu_c(a_t))\mathbb{I}_{c=[h_{<t}]}$$
$$\geq \sum_c \frac{T}{C}\epsilon_c\Big(1 - \frac{1}{A} - \sqrt{\epsilon_c \log(\frac{1+\epsilon_c}{1-\epsilon_c})}\sqrt{\frac{T}{2AC}}\Big).$$

Thus, after some tedious computations to optimize ϵ_c, we finally get a lower bound of order: $\frac{1}{20}\sum_c \sqrt{\frac{T}{C}A} = \frac{1}{20}\sqrt{TAC}$. Note that this is valid only if $\epsilon_c \sim \sqrt{\frac{A}{I_c(T)}}$ is small (less than 1), i.e. if the number of classes C is smaller than a constant times $\frac{T}{A}$ (and if this not the case, the lower bound becomes obviously of order T).

The second part of the Theorem can be proved using the same construction. □

2 Playing against an opponent using a pool of models

We first remind the result of [1] relating the cumulative reward of the Exp4 algorithm to the one of the best expert on top of which it is run. We have:

Lemma 1 *For any $\gamma \in (0,1]$, for any family of experts which includes the uniform expert, one has*

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

In our case, since the ξ_t^θ are not fixed in advance but are random variables, we can not apply this Lemma directly and need to adapt it. Based on the proof of [1], we can prove the following bound:

Lemma 2 *For any $\gamma \in (0,1]$, for any family of experts which includes the uniform expert such that all expert advices are adapted to the filtration of the past, one has*

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

Proof: Indeed, by construction of the algorithm, the beginning of the original proof from [1] applies and gives

$$\sum_{t=1}^{T} r_t(a_t) \geq (1-\gamma)\sum_{t=1}^{T} \mathbb{E}_{a\sim\xi_t^\theta}(\tilde{r}_t(a)) - \frac{A\log(|\Theta|)}{\gamma}$$
$$-(e-2)\frac{\gamma}{A}\sum_{t=1}^{T}\sum_{a\in\mathcal{A}} \tilde{r}_t(a),$$

where $\tilde{r}_t(a) = \frac{r_t(a_t)}{q_t(a)}\mathbb{I}_{a_t=a}$.

Now, we use the fact that $\xi_t^\theta(a)$ is adapted to the filtration of the past (which we denote $\mathcal{F}^{t-1}$) and the property that $\mathbb{E}(\tilde{r}_t(a)|\mathcal{F}^{t-1}) = \mathbb{E}(r_t(a)|\mathcal{F}^{t-1})$ to deduce that

$$\begin{aligned}
\mathbb{E}(\mathbb{E}_{a\sim\xi_t^\theta}(\tilde{r}_t(a))) &= \mathbb{E}(\sum_{a\in\mathcal{A}} \mathbb{E}(\tilde{r}_t(a)\xi_t^\theta(a)|\mathcal{F}^{t-1})) \\
&= \mathbb{E}(\sum_{a\in\mathcal{A}} \mathbb{E}(\tilde{r}_t(a)|\mathcal{F}^{t-1})\xi_t^\theta(a)) \\
&= \mathbb{E}(\sum_{a\in\mathcal{A}} \mathbb{E}(r_t(a)|\mathcal{F}^{t-1})\xi_t^\theta(a)) \\
&= \mathbb{E}(\sum_{a\in\mathcal{A}} \mathbb{E}(r_t(a)\xi_t^\theta(a)|\mathcal{F}^{t-1})) \\
&= \mathbb{E}(\mathbb{E}_{a\sim\xi_t^\theta}(r_t(a)))
\end{aligned}$$

On the other hand, since by assumption the uniform expert belongs to the set of considered experts, we also have

$$\begin{aligned}\frac{1}{A}\mathbb{E}(\sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\tilde{r}_t(a)) &= \sum_{t=1}^{T}\mathbb{E}(\mathbb{E}_{a\sim U(\mathcal{A})}(r_t(a)))\\ &\leq \max_{\theta}\sum_{t=1}^{T}\mathbb{E}(\mathbb{E}_{a\sim\xi_t^\theta}(r_t(a))),\end{aligned}$$

where $U(\mathcal{A})$ denotes the uniform distribution over the set of actions $\mathcal{A}$. This concludes the proof. □

We now define the best model of the pool θ^* to be

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

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. Thus we can also write $a_t^* = a^*_{[h_{<t}]_{\theta^*}}$.

2.1 The rebel bandit setting

We now introduce the setting of Rebel bandits that is interesting by itself. It will be used in order 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/she actually receives the reward corresponding to an action drawn with another distribution, say q_t. This will be the distribution of probability proposed by the meta algorithm. We now analyze the (Φ-constrained) Exp3 and UCB algorithms in this setting and bound the corresponding rebel-regret define by:

Definition 1 *(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:*

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

2.2 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 driven 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 3 *The Φ_{θ^*}-Exp3 algorithm in the Rebel bandit setting where $q_t(a)\geq\delta$ for all a, and choosing the parameter $\eta^\theta_{t_c^\theta(i)} = \sqrt{\frac{\delta\log(A)}{i}}$ satisfies*

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

The proof of this Theorem is reported in the main paper. We now combine Lemma 2 and Theorem 3 to get the final bound:

Theorem 4 *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: Indeed we can apply Theorem 3 using Exp4 meta algorithm with $\delta=\frac{\gamma}{A}$. Thus we get:

$$\begin{aligned}\tilde{R}_T^\Theta &= \sum_{t=1}^{T}\mathbb{E}_{a_1,\ldots,a_{t-1}}(r_t(a^*_{[h_{<t}]_{\theta^*}}) - \mathbb{E}_{a_t\sim q_t}r_t(a_t))\\ &\leq \mathcal{R}_T^q(\theta^*) + (e-1)\gamma T + \frac{A\log(|\Theta|)}{\gamma}\\ &\leq 2\sqrt{\frac{TA\overline{C}\log A}{\gamma}} + 2\gamma T + \frac{A\log(|\Theta|)}{\gamma}.\end{aligned}$$

We thus choose $\gamma = \frac{(A\overline{C}\log(A))^{1/3}\log(|\Theta|)^{1/2}}{(4T)^{1/3}}$ to conclude. □

2.3 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 $\Phi^* = \Phi^{\theta^*}\in\Phi_\Theta$.

Theorem 5 *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:

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

Proof: We write b_t the action proposed by the Φ-UCB algorithm at time t, and a_t the action effectively played according to distribution q_t. We introduce the notations: $I_c(T) = \sum_{t=1}^{T}\mathbb{I}_{[h_{<t}]=c}$, then $t_c(i) = \min\{t; I_c(t)=i\}$ and for all $a\in\mathcal{A}$, $I_c(T,a) =$

$\sum_{t=1}^{T} \mathbb{I}_{[h_{<t}]=c}\mathbb{I}_{a_t=a}$. The proof mainly follows the lines of [2]. Note that by definition, we want to bound the following term:

$$\mathcal{R}_T^q(\theta^*) = \sum_c \sum_a \Delta_c(a)\mathbb{E}\Big(\sum_{i=1}^{I_c^\theta(T)} \mathbb{I}_{b_t=a}\Big) \qquad (1)$$

Step one. Decompose the event $b_t = a$. Let us consider a time t for which $[h_{<t}] = c$. Then let us consider a sub-optimal arm a such that $\Delta_c(a) > 0$. Thus it appears that $b_t = a$ if one of the following conditions holds:

(1) $\tilde{\mu}_{t,c}(a_c) \leq \mu_c(a_c)$

(2) $\tilde{\mu}_{t,c}(a) > \mu_c(a)$

(3) $\Delta_c(a) < 2\sqrt{\frac{\alpha \log T}{I_c^\theta(t-1,a)}}$

Indeed, otherwise we would have

$$\begin{aligned}\tilde{\mu}_c(a_c) &> \mu_c(a_c) = \mu_c(a) + \Delta_c(a) \\ &\geq \mu_c(a) + 2\sqrt{\frac{\alpha \log I_c(t)}{I_c(t-1,a)}} \geq \tilde{\mu}_c(a).\end{aligned}$$

Thus we introduce the quantity $u_c(a) = \frac{4\alpha \log T}{\Delta_c(a)^2}$, and deduce that:

$$\mathbb{E}\Big(\sum_{i=1}^{I_c^\theta(T)} \mathbb{I}_{b_t=a}\Big) \leq \mathbb{E}\Big(\sum_{i=1}^{I_c^\theta(T)} \mathbb{I}_{(1)or(2)or I_c^\theta(t-1,a)<u_c(a)}\Big).$$

Step 2. Now since $I_c^\theta(.,a)$ is an increasing function of time (note though, that it does not increase by one each time b_t is proposed...), we can define the stopping time $\tau_c(a) = \min\{t; I_c^\theta(t,a) \geq u_c(a)\}$, or equivalently the stopping instant $i_c(a) = \min\{i; I_c^\theta(t_c^\theta(i),a) \geq u_c(a)\}$. Thus we deduce that:

$$\mathbb{E}\Big(\sum_{i=1}^{I_c^\theta(T)} \mathbb{I}_{b_t=a}\Big) \leq \mathbb{E}(i_c(a)) + \mathbb{E}\Big(\sum_{i=i_c(a)+1}^{I_c^\theta(T)} \mathbb{I}_{(1)or(2)}\Big) \quad (2)$$

Now we can bound the second term of (2) by a constant depending only on α, by an easy peeling argument (we refer to Section 2.2 of [2]):

$$\mathbb{E}\Big(\sum_{i=i_c(a)+1}^{I_c^\theta(T)} \mathbb{I}_{(1)or(2)}\Big) \leq 2\mathbb{E}\Big(\sum_{i=i_c(a)+1}^{I_c^\theta(T)} \Big(\frac{\log i}{\log 1/\beta}+1\Big)\frac{1}{i^{2\beta\alpha}}\Big) \qquad (3)$$

where $\beta = \frac{1}{\alpha+1/2}$.

Then, we also have, by integration by parts:

$$\begin{aligned}2\mathbb{E}\Big(\sum_{i=i_c(a)+1}^{I_c^\theta(T)} \Big(\frac{\log i}{\log 1/\beta}+1\Big)\frac{1}{i^{2\beta\alpha}}\Big) &\leq 2\int_1^\infty \Big(\frac{\log t}{\log 1/\beta}+1\Big)\frac{1}{t^{2\beta\alpha}})dt \\ &\leq \frac{4}{\log(1/\beta)(2\beta\alpha-1)^2}.\end{aligned}$$

Step 3. Thus we focus on the first term $\mathbb{E}(i_c(a))$ of (2). Since we know that $q_t(a) \geq \delta$ for all a, t, we thus deduce that:

$$\begin{aligned}&\mathbb{E}(i_c(a)) = \textstyle\sum_{l=0}^\infty \mathbb{P}(i_c(a) > l) \leq \\ &l_0 + \textstyle\sum_{l=l_0}^\infty \mathbb{P}\big(\forall j \leq l\ I_c^\theta(t_c^\theta(j),a) < u_c(a)\big) \leq \\ &l_0 + \textstyle\sum_{l=l_0}^\infty \mathbb{P}\big(\forall j \leq l\ \sum_{i=1}^j \mathbb{I}_{a_{t_c^\theta(i)}=a} - q_{t_c^\theta(i)}(a) < u_c(a) - \delta j\big).\end{aligned}$$

Now by property of martingale difference sequences, we deduce by setting $l_0 = \ulcorner \frac{u_c(a)}{\delta} \urcorner$, that:

$$\begin{aligned}\mathbb{E}(i_c(a)) &\leq l_0 + \sum_{l=l_0}^\infty \exp(-2(l-l_0)^2\delta^2 l) \\ &\leq l_0 + \sum_{l=l_0}^\infty \exp(-\frac{(l-l_0)^2}{2\sigma^2}),\end{aligned}$$

where we introduced the quantity $\sigma^2 = \frac{1}{4\delta^2 l_0}$. Thus we deduce that:

$$\mathbb{E}(i_c(a) \leq \ulcorner \frac{u_c(a)}{\delta} \urcorner + \sqrt{\frac{\pi}{8}}\sqrt{\frac{\delta}{u_c(a)}}. \qquad (4)$$

Step 4. Finally, by combining (3), (4) with (2) and (1), we deduce the following distribution-dependent bound on the rebel regret:

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

where $c_\alpha = 1 + \frac{4}{\log(\alpha+1/2)}(\frac{\alpha+1/2}{\alpha-1/2})^2$. We deduce the distribution-free bound by the same argument as for Theorem 1, remarking that $\sqrt{\frac{\pi}{8}}\sqrt{\frac{\delta\Delta_c(a)^2}{4\alpha\log T}} \leq \sqrt{\frac{\pi}{32\alpha\log(T)}} = c'_\alpha$. □

This enables us to deduce the following Theorem, that we prove using the same method as that of Theorem 4 but for the stochastic Φ_Θ-regret of Exp4/UCB.

Theorem 6 *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.

3 Approximation error of the models

The following result sheds light on a specific term that appears to be an approximation term of the true model θ^* by other models θ.

Theorem 7 *For any $(p_t(\theta))_{t,\theta} \in [0,1]$, thus for any meta algorithm run on top of Exp3 algorithm and defined with $q_t(a) = \sum_\theta p_t(\theta)\xi_t^\theta(a)$ and decreasing coefficient η_t^θ, the following holds true:*

$$\tilde{R}_T^\Theta \le \mathbb{E}\Big(\sum_\theta \sum_{c\in\theta} \sum_{i=1}^{I_c^\theta(T)} \frac{\eta^\theta_{t_c^\theta(i)} A}{2} p_{t_c^\theta(i)}(\theta) + \sum_\theta \sum_{c\in\theta} \frac{\log A}{\eta^\theta_{t_c^\theta(I_c^\theta(T))}} + \sum_\theta \sum_{c\in\theta} \inf_{a_c^\theta} \sum_{i=1}^{I_c^\theta(T)} (r_{t_c^\theta(i)}(a^*_{[h_{<t_c^\theta(i)}]}) - r_{t_c^\theta(i)}(a_c^\theta)) p_{t_c^\theta(i)}(\theta)\Big).$$

The term on the second line is actually a mixture of approximation errors of each model, and it seems it can not be reduced without further assumption on the quality of the models.

Proof: The proof is in four steps.

Step 1. Rewrite the regret to make appear the probabilities $\xi_t^\theta(a)$. We first introduce:

$$\begin{aligned} R_T &= \sum_{t=1}^T r_t(a_{[h_{<t}]_*}) - r_t(a_t) \\ &= \sum_\theta \sum_{t=1}^T \mathbb{E}_{a_t\sim q_t}\Big(\frac{r_t(a_t)p_t(\theta)}{q_t(a_t)} \mathbb{I}_{a_t = a_{[h_{<t}]_*}}\Big) \\ &\quad - \frac{r_t(a_t)\xi_t^\theta(a_t)p_t(\theta)}{q_t(a_t)}. \end{aligned}$$

Now we have : $\tilde{l}^\theta_{t,c_\theta}(a) = (1-r_t(a))\frac{p_t(\theta)}{q_t(a)}\mathbb{I}_{a_t=a}\mathbb{I}_{c_\theta=[h_{<t}]_\theta}$, thus taking the expectation over a_t for each time t, we have:

$$\begin{aligned} \tilde{R}_T &= \textstyle\sum_{t=1}^T \mathbb{E}_{a_r}(r_t(a_{[h_{<t}]_*} - r_t(a_t)) \\ &= \textstyle\sum_\theta \sum_{t=1}^T \mathbb{E}_{a_t}\Big(\frac{p_t(\theta)}{q_t(a_t)}\mathbb{I}_{a_t=a_{[h_{<t}]_*}} - \tilde{l}^\theta_{t,[h_{<t}]_\theta}(a_{[h_{<t}]_*})\Big)\Big) \\ &+ \textstyle\sum_\theta \sum_{t=1}^T \mathbb{E}_{a_t}\Big(\mathbb{E}_{a\sim\xi_t^\theta}(\tilde{l}^\theta_{t,[h_{<t}]_\theta}(a)) - \frac{p_t(\theta)\xi_t^\theta(a_t)}{q_t(a_t)}\Big). \end{aligned}$$

We can simplify the above expression since $\mathbb{E}_{a_t}(\frac{p_t(\theta)}{q_t(a_t)}\mathbb{I}_{a_t=a_{[h_{<t}]_*}}) = \mathbb{E}_{a_t}(\frac{p_t(\theta)\xi_t^\theta(a_t)}{q_t(a_t)}) = p_t(\theta)$.

Step 2. Decompose the term $\mathbb{E}_{a\sim\xi_t^\theta}(\tilde{l}^\theta_{t,[h_{<t}]_\theta}(a))$ in order to use the definition of ξ_t^θ. Indeed, one can upper bound this term by

$$\begin{aligned} \mathbb{E}_{a\sim\xi_t^\theta}(\tilde{l}^\theta_{t,[h_{<t}]_\theta}(a)) &\le \frac{\eta_t^\theta}{2}\mathbb{E}_{a\sim\xi_t^\theta}(\tilde{l}^\theta_{t,[h_{<t}]_\theta}(a)^2) \\ &\quad - \frac{1}{\eta_t^\theta}\log\Big(\sum_a \exp(-\eta_t^\theta \tilde{l}^\theta_{t,[h_{<t}]_\theta}(a)\xi_t^\theta(a))\Big). \end{aligned}$$

Thus, since by definition we have that $\xi_t^\theta(a) = \frac{\exp(-\eta_t^\theta \sum_{s=1}^{t-1}\tilde{l}^\theta_{s,[h_{<t}]_\theta}(a))}{\sum_a \exp(-\eta_t^\theta \sum_{s=1}^{t-1}\tilde{l}^\theta_{s,[h_{<t}]_\theta}(a))}$, we can introduce the quantity $\Psi_t^\theta(\eta,c) = \frac{1}{\eta}\log(\frac{1}{A}\sum_a \exp(-\eta\sum_{s=1}^t \tilde{l}^\theta_{s,c}(a)))$ so that the previous regret term writes:

$$\begin{aligned} \tilde{R}_T &\le \sum_\theta \sum_{t=1}^T \mathbb{E}_{a_t}\Big(\frac{\eta_t^\theta}{2}(1-r_t(a_t))^2\frac{p_t^2(\theta)\xi_t^\theta(a_t)}{q_t^2(a_t)}\Big) \\ &\quad + \sum_\theta \Big(\sum_{t=1}^T \mathbb{E}_{a_t}(\Psi^\theta_{t-1}(\eta_t^\theta,[h_{<t}]_\theta) - \Psi_t^\theta(\eta_t^\theta,[h_{<t}]_\theta)) \\ &\quad - \mathbb{E}_{a_t}(\tilde{l}^\theta_{t,[h_{<t}]_\theta}(a_{[h_{<t}]_*}))\Big). \end{aligned}$$

Step 3. Introduce the equivalence classes. We now consider the term in the right hand side of the above equation defined with Ψ functions. Note that we do not change the bound on the term $\tilde{R}_T$ by considering the sum over the θ such that $p_t(\theta) > 0$.

Let us introduce the following notations $I_c^\theta(t) = \sum_{s=1}^t \mathbb{I}_{c=[h_{<s}]_\theta}\mathbb{I}_{p_s(\theta)>0}$ and $t_c^\theta(i) = \min\{t; I_c^\theta(t) = i\}$. Thus we can write:

$$\begin{aligned} &\textstyle\sum_\theta \sum_{t=1}^T (\Psi^\theta_{t-1}(\eta_t^\theta,[h_{<t}]_\theta) - \Psi_t^\theta(\eta_t^\theta,[h_{<t}]_\theta))\mathbb{I}_{p_t(\theta)>0} \\ = &\textstyle\sum_\theta \sum_{c\in\theta} \sum_{i=1}^{I_c^\theta(T)} \Psi^\theta_{t_c^\theta(i)-1}(\eta^\theta_{t_c^\theta(i)},c) - \Psi^\theta_{t_c^\theta(i)}(\eta^\theta_{t_c^\theta(i)},c) \\ = &\textstyle\sum_\theta \sum_{c\in\theta} \sum_{i=1}^{I_c^\theta(T)-1} (\Psi^\theta_{t_c^\theta(i)}(\eta^\theta_{t_c^\theta(i)+1},c) - \Psi^\theta_{t_c^\theta(i)}(\eta^\theta_{t_c^\theta(i)},c) \\ &\qquad - \Psi^\theta_{t_c^\theta(I_c^\theta(T))}(\eta^\theta_{t_c^\theta(I_c^\theta(T))},c). \end{aligned}$$

Now, by definition of Ψ_t^θ, the last term of this sum, $-\Psi^\theta_{t_c^\theta(I_c^\theta(T))}(\eta^\theta_{t_c^\theta(I_c^\theta(T))},c)$, is equal to the following quantity

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

where $\eta = \eta^\theta_{t_c^\theta(I_c^\theta(T))}$, which can be upper bounded by $\frac{\log A}{\eta} + \sum_{s=1}^{t_c^\theta(I_c^\theta(T))} \tilde{l}^\theta_{s,c}(a)$ for any given $a = a_c^\theta$.

Step 4. Now since $\eta^\theta_{t_c^\theta(i)} \le \eta^\theta_{t_c^\theta(i)+1}$ and $\Psi^\theta_{t_c^\theta(i)}(\cdot,c)$ is increasing for all θ, c, we deduce from the previous equations that:

$$\begin{aligned} \tilde{R}_T &\le \sum_\theta \sum_{c\in\theta} \sum_{i=1}^{I_c^\theta(T)} \sum_a \frac{\eta^\theta_{t_c^\theta(i)}}{2}\frac{p^2_{t_c^\theta(i)}(\theta)\xi^\theta_{t_c^\theta(i)}(a)}{q_{t_c^\theta(i)}(a)} \\ &\quad + \sum_\theta \sum_{c\in\theta} \frac{\log A}{\eta^\theta_{t_c^\theta(I_c^\theta(T))}} \\ &\quad + \sum_\theta \sum_{c\in\theta} \inf_{a_c^\theta} \sum_{t=1}^{t_c^\theta(I_c^\theta(T))} \mathbb{E}_{a_t}(\tilde{l}^\theta_{t,c}(a_c^\theta) - \tilde{l}^\theta_{t,c}(a_{[h_{<t}]_*})). \end{aligned}$$

Now we conclude by taking the expectation, seeing that $p_{t_c^\theta(i)}(\theta)\xi_{t_c^\theta(i)} \le q_{t_c^\theta(i)}(a)$, and that by definition $\mathbb{E}_{a_t}(\tilde{l}^\theta_{t,c}(a_c^\theta)) = (1-r_t(a_c^\theta))p_t(\theta)\mathbb{I}_{c=[h_{<t}]_\theta}$.

□

References

[1] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32:48–77, January 2003.

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