Regret Bounds and Minimax Policies under Partial Monitoring

Jean-Yves Audibert, Sébastien Bubeck

Journal of Machine Learning Research 11 (2010) 2597-2648 Submitted 7/09; Revised 6/10; Published 10/10

Regret Bounds and Minimax Policies under Partial Monitoring

Jean-Yves Audibert* AUDIBERT@IMAGINE.ENPC.FR
Imagine, Université Paris Est
6 avenue Blaise Pascal
77455 Champs-sur-Marne, France

Sébastien Bubeck SEBASTIEN.BUBECK@INRIA.FR
SequeL Project, INRIA Lille
40 avenue Halley
59650 Villeneuve d'Ascq, France

Editor: Nicolò Cesa-Bianchi

Abstract

This work deals with four classical prediction settings, namely full information, bandit, label efficient and bandit label efficient as well as four different notions of regret: pseudo-regret, expected regret, high probability regret and tracking the best expert regret. We introduce a new forecaster, INF (Implicitly Normalized Forecaster) based on an arbitrary function ψ for which we propose a unified analysis of its pseudo-regret in the four games we consider. In particular, for $\psi(x) = \exp(\eta x) + \frac{\gamma}{K}$, INF reduces to the classical exponentially weighted average forecaster and our analysis of the pseudo-regret recovers known results while for the expected regret we slightly tighten the bounds. On the other hand with $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$, which defines a new forecaster, we are able to remove the extraneous logarithmic factor in the pseudo-regret bounds for bandits games, and thus fill in a long open gap in the characterization of the minimax rate for the pseudo-regret in the bandit game. We also provide high probability bounds depending on the cumulative reward of the optimal action.

Finally, we consider the stochastic bandit game, and prove that an appropriate modification of the upper confidence bound policy UCB1 (Auer et al., 2002a) achieves the distribution-free optimal rate while still having a distribution-dependent rate logarithmic in the number of plays.

Keywords: Bandits (adversarial and stochastic), regret bound, minimax rate, label efficient, upper confidence bound (UCB) policy, online learning, prediction with limited feedback.

1. Introduction

This section starts by defining the prediction tasks, the different regret notions that we will consider, and the different adversaries of the forecaster. We will then recap existing lower and upper regret bounds for the different settings, and give an overview of our contributions.

∗. Also at Willow, CNRS/ENS/INRIA—UMR 8548.

Parameters: the number of arms (or actions) K and the number of rounds n with $n \geq K \geq 2$.

For each round $t = 1, 2, \ldots, n$

(1) The forecaster chooses an arm $I_t \in \{1, \ldots, K\}$, possibly with the help of an external randomization.

(2) Simultaneously the adversary chooses a gain vector $g_t = (g_{1,t}, \ldots, g_{K,t}) \in [0,1]^K$ (see Section 8 for loss games or signed games).

(3) The forecaster receives the gain $g_{I_t,t}$ (without systematically observing it). He observes

- the reward vector $(g_{1,t}, \ldots, g_{K,t})$ in the **full information** game,
- the reward vector $(g_{1,t}, \ldots, g_{K,t})$ if he asks for it with the global constraint that he is not allowed to ask it more than m times for some fixed integer number $1 \leq m \leq n$. This prediction game is the **label efficient** game,
- only $g_{I_t,t}$ in the **bandit** game,
- only his obtained reward $g_{I_t,t}$ if he asks for it with the global constraint that he is not allowed to ask it more than m times for some fixed integer number $1 \leq m \leq n$. This prediction game is the **label efficient bandit** game.

Goal : The forecaster tries to maximize his cumulative gain $\sum_{t=1}^{n} g_{I_t,t}$.

Figure 1: The four prediction tasks considered in this work.

1.1 The Four Prediction Tasks

We consider a general prediction game where at each stage, a forecaster (or decision maker) chooses one action (or arm), and receives a reward from it. Then the forecaster receives a feedback about the rewards which he can use to make his choice at the next stage. His goal is to maximize his cumulative gain. In the simplest version, after choosing an arm the forecaster observes the rewards for all arms, this is the so called full information game. In the label efficient game, originally proposed by Helmbold and Panizza (1997), after choosing its action at a stage, the forecaster decides whether to ask for the rewards of the different actions at this stage, knowing that he is allowed to do it a limited number of times. Another classical setting is the bandit game where the forecaster only observes the reward of the arm he has chosen. In its original version (Robbins, 1952), this game was considered in a stochastic setting, that is, the nature draws the rewards from a fixed product-distribution. Later it was considered in an adversarial framework (Auer et al., 1995), where there is an adversary choosing the rewards on the arms. A combination of the two previous settings is the label efficient bandit game (György and Ottucsák, 2006), in which the only observed rewards are the ones obtained and asked by the forecaster, with again a limitation on the number of possible queries. These four games are described more precisely in Figure 1. Their Hannan consistency has been considered in Allenberg et al. (2006) in the case of

unbounded losses. Here we will focus on regret upper bounds and minimax policies for bounded losses.

1.2 Regret and Pseudo-regret

A natural way to assess the performance of a forecaster is to compute his *regret* with respect to the best action in hindsight (see Section 7 for a more general regret in which we compare to the best switching strategy having a fixed number of action-switches):

$$R_n = \max_{1\le i\le K} \sum_{t=1}^{n} \big(g_{i,t} - g_{I_t,t}\big).$$

A lot of attention has been drawn by the characterization of the minimax expected regret in the different games we have described. More precisely for a given game, let us write sup for the supremum over all allowed adversaries and inf for the infimum over all forecaster strategies for this game. We are interested in the quantity:

$$\inf \sup \mathbb{E} R_n,$$

where the expectation is with respect to the possible randomization of the forecaster and the adversary. Another related quantity which can be easier to handle is the *pseudo-regret*:

$$\overline{R}_n = \max_{1\le i\le K} \mathbb{E} \sum_{t=1}^{n} \big(g_{i,t} - g_{I_t,t}\big).$$

Note that, by Jensen's inequality, the pseudo-regret is always smaller than the expected regret. In Appendix D we discuss cases where the converse inequality holds (up to an additional term).

1.3 The Different Adversaries

The simplest adversary is the deterministic one. It is characterized by a fixed matrix of nK rewards corresponding to $(g_{i,t})_{1\le i\le K, 1\le t\le n}$. Another adversary is the "stochastic" one, in which the reward vectors are independent and have the same distribution.[1] This adversary is characterized by a distribution on $[0,1]^K$, corresponding to the common distribution of g_t, $t = 1, \ldots, n$. A more general adversary is the fully oblivious one, in which the reward vectors are independent. Here the adversary is characterized by n distributions on $[0,1]^K$ corresponding to the distributions of $g_1, \ldots, g_n$. Deterministic and stochastic adversaries are fully oblivious adversaries.

An even more general adversary is the oblivious one, in which the only constraint on the adversary is that the reward vectors are independent of the past decisions of the forecaster. The most general adversary is the one who may choose the reward vector g_t as a function of the past decisions $I_1, \ldots, I_{t-1}$ (non-oblivious adversary).

1. The term "stochastic" can be a bit misleading since the assumption is not just stochasticity but rather an i.i.d. assumption.

	$\inf \sup \overline{R}_n$		$\inf \sup \mathbb{E}R_n$	
	Lower bound	Upper bound	Lower bound	Upper bound
Full information game	$\sqrt{n \log K}$	$\sqrt{n \log K}$	$\sqrt{n \log K}$	$\sqrt{n \log K}$
Label efficient game	$n\sqrt{\frac{\log K}{m}}$	$n\sqrt{\frac{\log K}{m}}$	$n\sqrt{\frac{\log K}{m}}$	$n\sqrt{\frac{\log n}{m}}$
Bandit game	$\sqrt{nK}$	$\sqrt{nK \log K}$	$\sqrt{nK}$	$\sqrt{nK \log n}$
Bandit label efficient game	$n\sqrt{\frac{K}{m}}$	$n\sqrt{\frac{K \log K}{m}}$	$n\sqrt{\frac{K}{m}}$	$n\sqrt{\frac{K \log n}{m}}$

Table 1: Existing bounds (apart from the lower bounds in the last line which are proved in this paper) on the pseudo-regret and expected regret. Except for the full information game, there are logarithmic gaps between lower and upper bounds.

1.4 Known Regret Bounds

Table 1 recaps existing lower and upper bounds on the minimax pseudo-regret and the minimax expected regret for general adversaries (i.e., possibly non-oblivious ones). For the first three lines, we refer the reader to the book (Cesa-Bianchi and Lugosi, 2006) and references within, particularly Cesa-Bianchi et al. (1997) and Cesa-Bianchi (1999) for the full information game, Cesa-Bianchi et al. (2005) for the label efficient game, Auer et al. (2002b) for the bandit game and György and Ottucsák (2006) for the label efficient bandit game. The lower bounds in the last line do not appear in the existing litterature, but we prove them in this paper. Apart from the full information game, the upper bounds are usually proved on the pseudo-regret. The upper bounds on the expected regret are obtained by using high probability bounds on the regret. The parameters of the algorithm in the latter bounds usually depend on the confidence level δ that we want to obtain. Thus to derive bounds on the expected regret we can not integrate the deviations but rather we have to take δ of order $1/n$, which leads to the gaps involving $\log(n)$. Table 1 exhibits several logarithmic gaps between upper and lower bounds on the minimax rate, namely:

- $\sqrt{\log(K)}$ gap for the minimax pseudo-regret in the bandit game as well as the label efficient bandit game.
- $\sqrt{\log(n)}$ gap for the minimax expected regret in the bandit game as well as the label efficient bandit game.
- $\sqrt{\log(n)/\log(K)}$ gap for the minimax expected regret in the label efficient game,

1.5 Contributions of This Work

We reduce the above gaps by improving the upper bounds as shown by Table 2. Different proof techniques are used and new forecasting strategies are proposed. The most original contribution is the introduction of a new forecaster, INF (Implicitly Normalized Forecaster), for which we propose a unified analysis of its regret in the four games we consider. The analysis is original (it avoids the traditional but scope-limiting argument based on the simplification of a sum of logarithms of ratios), and allows to fill in the long open gap in the bandit problems with oblivious adversaries (and with general adversaries for the pseudo-regret notion). The analysis also applies to exponentially weighted average forecasters. It

allows to prove a regret bound of order $\sqrt{nKS\log(nK/S)}$ when the forecaster's strategy is compared to a strategy allowed to switch S times between arms, while the best known bound was $\sqrt{nKS\log(nK)}$ (Auer, 2002), and achieved for a different policy.

An "orthogonal" contribution is to propose a tuning of the parameters of the forecasting policies such that the high probability regret bounds holds for any confidence level (instead of holding just for a single confidence level as in previous works). Bounds on the expected regret that are deduced from these PAC ("probably approximately correct") regret bounds are better than previous bounds by a logarithmic factor in the games with limited information (see columns on $\inf\sup\mathbb{E}R_n$ in Tables 1 and 2). The arguments to obtain these bounds are not fundamentally new and rely essentially on a careful use of deviation inequalities for supermartingales. They can be used either in the standard analysis of exponentially weighted average forecasters or in the more general context of INF.

Another "orthogonal" contribution is the proposal of a new biased estimate of the rewards in bandit games, which allows to achieve high probability regret bounds depending on the performance of the optimal arm: in this new bound, the factor n is replaced by $G_{\max} = \max_{i=1,\dots,n}\sum_{t=1}^{n} g_{i,t}$. If the forecaster draws I_t according to the distribution $p_t = (p_{1,t},\dots,p_{K,t})$, then the new biased estimate of $g_{i,t}$ is $v_{i,t} = -\frac{\mathbb{1}_{I_t=i}}{\beta}\log\left(1-\frac{\beta g_{i,t}}{p_{i,t}}\right)$. This estimate should be compared to $v_{i,t} = g_{i,t}\frac{\mathbb{1}_{I_t=i}}{p_{i,t}}$, for which bounds in terms of $G_{\max}$ exists in expectations as shown in (Auer et al., 2002b, Section 3), and to $v_{i,t} = g_{i,t}\frac{\mathbb{1}_{I_t=i}}{p_{i,t}} + \frac{\beta}{p_{i,t}}$ for some $\beta > 0$ for which high probability bounds exist but they are expressed with the n factor, and not $G_{\max}$ (see Section 6 of Auer et al., 2002b, and Section 6.8 of Cesa-Bianchi and Lugosi, 2006).

We also propose a unified proof to obtain the lower bounds in Table 1. The contribution of this proof is two-fold. First it gives the first lower bound for the label efficient bandit game. Secondly in the case of the label efficient (full information) game it is a simpler proof than the one proposed in Cesa-Bianchi et al. (2005). Indeed in the latter proof, the authors use Birgé's version of Fano's lemma to prove the lower bound for deterministic forecasters. Then the extension to non-deterministic forecasters is done by a generalization of this information lemma and a decomposition of general forecasters into a convex combination of deterministic forecasters. The benefit from this proof technique is to be able to deal with the case $K = 2$ and $K = 3$ while the basic version of Fano's lemma does not give any information in this case. Here we propose to use Pinsker's inequality for the case $K = 2$ and $K = 3$. This allows us to use the basic version of Fano's lemma and to extend the result to non-deterministic forecasters with a simple application of Fubini's Theorem.

The last contribution of this work is also independent of the previous ones and concerns the stochastic bandit game (that is the bandit game with "stochastic" adversary). We prove that a modification of UCB1, Auer et al. (2002a), attains the optimal distribution-free rate $\sqrt{nK}$ as well as the logarithmic distribution-dependent rate. The key idea, compared to previous works, is to reduce exploration of sufficiently drawn arms.

1.6 Outline

In Section 2, we describe a new class of forecasters, called INF, for prediction games. Then we present a new forecaster inside this class, called Poly INF, for which we propose a general

	$\inf \sup \overline{R}_n$	$\inf \sup \mathbb{E} R_n$	High probability bound on R_n
Label efficient game		$n\sqrt{\frac{\log K}{m}}$	$n\sqrt{\frac{\log(K\delta^{-1})}{m}}$
Bandit game with fully oblivious adversary	$\sqrt{nK}$	$\sqrt{nK}$	$\sqrt{nK}\log(\delta^{-1})$
Bandit game with oblivious adversary	$\sqrt{nK}$	$\sqrt{nK}$	$\sqrt{\frac{nK}{\log K}}\log(K\delta^{-1})$
Bandit game with general adversary	$\sqrt{nK}$	$\sqrt{nK\log K}$	$\sqrt{\frac{nK}{\log K}}\log(K\delta^{-1})$
L.E. bandit with deterministic adversary	$n\sqrt{\frac{K}{m}}$	$n\sqrt{\frac{K}{m}}$	$n\sqrt{\frac{K}{m}}\log(\delta^{-1})$
L.E. bandit with oblivious adversary	$n\sqrt{\frac{K}{m}}$	$n\sqrt{\frac{K}{m}}$	$n\sqrt{\frac{K}{m\log K}}\log(K\delta^{-1})$
L.E. bandit with general adversary	$n\sqrt{\frac{K}{m}}$	$n\sqrt{\frac{K\log K}{m}}$	$n\sqrt{\frac{K}{m\log K}}\log(K\delta^{-1})$

Table 2: New regret upper bounds proposed in this work. The high probability bounds are for a policy of the forecaster that does not depend on the confidence level δ (unlike previously known high probability bounds).

theorem bounding its regret. A more general statement on the regret of any INF can be found in Appendix A. Exponentially weighted average forecasters are a special case of INF as shown in Section 3. In Section 4, we prove that our forecasters and analysis recover the known results for the full information game.

Section 5 contains the core contributions of the paper, namely all the regret bounds for the limited feedback games. The interest of Poly INF appears in the bandit games where it satisfies a regret bound without a logarithmic factor, unlike exponentially weighted average forecasters. Section 6 provides high probability bounds in the bandit games that depends on the cumulative reward of the optimal arm: the factor n is replaced by $\max_{1\leq i\leq K}\sum_{t=1}^{n} g_{i,t}$. In Section 7, we consider a stronger notion of regret, when we compare ourselves to a strategy allowed to switch between arms a fixed number of times. Section 8 shows how to generalize our results when one considers losses rather than gains, or signed games.

Section 9 considers a framework fundamentally different from the previous sections, namely the stochastic multi-armed bandit problem. There we propose a new forecaster, MOSS, for which we prove an optimal distribution-free rate as well as a logarithmic distribution-dependent rate.

Appendix A contains a general regret upper bound for INF and two useful technical lemmas. Appendix B contains the unified proof of the lower bounds. Appendix C contains the proofs that have not been detailed in the main body of the paper. Finally, Appendix D gathers the different results we have obtained regarding the relation between the expected regret and the pseudo-regret.

2. The Implicitly Normalized Forecaster

In this section, we define a new class of randomized policies for the general prediction game. Let us consider a continuously differentiable function $\psi : \mathbb{R}_-^* \to \mathbb{R}_+^*$ satisfying

$$\psi' > 0, \qquad \lim_{x\to-\infty} \psi(x) < 1/K, \qquad \lim_{x\to 0} \psi(x) \geq 1. \tag{1}$$

Lemma 1 *There exists a continuously differentiable function $C : \mathbb{R}_+^K \to \mathbb{R}$ satisfying for any $x = (x_1, \ldots, x_K) \in \mathbb{R}_+^K$,*

$$\max_{1 \le i \le K} x_i < C(x) \le \max_{1 \le i \le K} x_i - \psi^{-1}(1/K), \tag{2}$$

and

$$\sum_{i=1}^{K} \psi(x_i - C(x)) = 1. \tag{3}$$

Proof Consider a fixed $x = (x_1, \ldots, x_K)$. The decreasing function $\phi : c \mapsto \sum_{i=1}^K \psi(x_i - c)$ satisfies

$$\lim_{c \to \max_{1 \le i \le K} x_i} \phi(c) > 1 \qquad \text{and} \qquad \lim_{c \to +\infty} \phi(c) < 1.$$

From the intermediate value theorem, there is a unique $C(x)$ satisfying $\phi(C(x)) = 1$. From the implicit function theorem, the mapping $x \mapsto C(x)$ is continuously differentiable. ■

INF (Implicitly Normalized Forecaster):

Parameters:

- the continuously differentiable function $\psi : \mathbb{R}_-^* \to \mathbb{R}_+^*$ satisfying (1)
- the estimates $v_{i,t}$ of $g_{i,t}$ based on the (drawn arms and) observed rewards at time t (and before time t)

Let p_1 be the uniform distribution over $\{1, \ldots, K\}$.

For each round $t = 1, 2, \ldots,$

(1) Draw an arm I_t from the probability distribution p_t.

(2) Use the observed reward(s) to build the estimate $v_t = (v_{1,t}, \ldots, v_{K,t})$ of $(g_{1,t}, \ldots, g_{K,t})$ and let: $V_t = \sum_{s=1}^t v_s = (V_{1,t}, \ldots, V_{K,t})$.

(3) Compute the normalization constant $C_t = C(V_t)$.

(4) Compute the new probability distribution $p_{t+1} = (p_{1,t+1}, \ldots, p_{K,t+1})$ where

$$p_{i,t+1} = \psi(V_{i,t} - C_t).$$

Figure 2: The proposed policy for the general prediction game.

The implicitly normalized forecaster (INF) is defined in Figure 2. Equality (3) makes the fourth step in Figure 2 legitimate. From (2), $C(V_t)$ is roughly equal to $\max_{1 \le i \le K} V_{i,t}$. Recall that $V_{i,t}$ is an estimate of the cumulative gain at time t for arm i. This means that INF chooses the probability assigned to arm i as a function of the (estimated) regret. Note that, in spirit, it is similar to the traditional weighted average forecaster, see for example Section 2.1 of Cesa-Bianchi and Lugosi (2006), where the probabilities are proportional to a function of the difference between the (estimated) cumulative reward of arm i and the

cumulative reward of the policy, which should be, for a well-performing policy, of order $C(V_t)$.

The interesting feature of the implicit normalization is the following argument, which allows to recover the results concerning the exponentially weighted average forecasters, and more interestingly to propose a policy having a regret of order $\sqrt{nK}$ in the bandit game with oblivious adversary. First note that $\sum_{t=1}^n \sum_{i=1}^K p_{i,t} v_{i,t}$ roughly evaluates the cumulative reward $\sum_{t=1}^n g_{I_t,t}$ of the policy. In fact, it is exactly the cumulative gain in the bandit game when $v_{i,t} = g_{i,t} \frac{\mathbb{1}_{I_t=i}}{p_{i,t}}$, and its expectation is exactly the expected cumulative reward in the full information game when $v_{i,t} = g_{i,t}$. The argument starts with an Abel transformation and consequently is "orthogonal" to the usual argument given in the beginning of Section C.2. Letting $V_0 = 0 \in \mathbb{R}^K$. We have

$$\begin{aligned}
\sum_{t=1}^n g_{I_t,t} &\approx \sum_{t=1}^n \sum_{i=1}^K p_{i,t} v_{i,t} \\
&= \sum_{t=1}^n \sum_{i=1}^K p_{i,t} (V_{i,t} - V_{i,t-1}) \\
&= \sum_{i=1}^K p_{i,n+1} V_{i,n} + \sum_{i=1}^K \sum_{t=1}^n V_{i,t} (p_{i,t} - p_{i,t+1}) \\
&= \sum_{i=1}^K p_{i,n+1} \left(\psi^{-1}(p_{i,n+1}) + C_n \right) + \sum_{i=1}^K \sum_{t=1}^n (\psi^{-1}(p_{i,t+1}) + C_t)(p_{i,t} - p_{i,t+1}) \\
&= C_n + \sum_{i=1}^K p_{i,n+1} \psi^{-1}(p_{i,n+1}) + \sum_{i=1}^K \sum_{t=1}^n \psi^{-1}(p_{i,t+1})(p_{i,t} - p_{i,t+1}),
\end{aligned}$$

where the remarkable simplification in the last step is closely linked to our specific class of randomized algorithms. The equality is interesting since, from (2), C_n approximates the maximum estimated cumulative reward $\max_{1 \le i \le K} V_{i,n}$, which should be close to the cumulative reward of the optimal arm $\max_{1 \le i \le K} G_{i,n}$, where $G_{i,n} = \sum_{t=1}^n g_{i,t}$. Since the last term in the right-hand side is

$$\sum_{i=1}^K \sum_{t=1}^n \psi^{-1}(p_{i,t+1})(p_{i,t} - p_{i,t+1}) \approx \sum_{i=1}^K \sum_{t=1}^n \int_{p_{i,t}}^{p_{i,t+1}} \psi^{-1}(u) du = \sum_{i=1}^K \int_{1/K}^{p_{i,n+1}} \psi^{-1}(u) du, \quad (4)$$

we obtain

$$\max_{1 \le i \le K} G_{i,n} - \sum_{t=1}^n g_{I_t,t} \lessapprox - \sum_{i=1}^K p_{i,n+1} \psi^{-1}(p_{i,n+1}) + \sum_{i=1}^K \int_{1/K}^{p_{i,n+1}} \psi^{-1}(u) du. \quad (5)$$

The right-hand side is easy to study: it depends only on the final probability vector and has simple upper bounds for adequate choices of ψ. For instance, for $\psi(x) = \exp(\eta x) + \frac{\gamma}{K}$ with $\eta > 0$ and $\gamma \in [0,1)$, which corresponds to exponentially weighted average forecasters as we will explain in Section 3, the right-hand side is smaller than $\frac{1-\gamma}{\eta} \log\left(\frac{K}{1-\gamma}\right) + \gamma C_n$. For $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$ with $\eta > 0$, $q > 1$ and $\gamma \in [0,1)$, which will appear to be a fruitful

choice, it is smaller than $\frac{q}{q-1}\eta K^{1/q} + \gamma C_n$. For sake of simplicity, we have been hiding the residual terms of (4) coming from the Taylor expansions of the primitive function of ψ^{-1}. However, these terms when added together (nK terms!) are not that small, and in fact constrain the choice of the parameters γ and η if one wishes to get the tightest bound.

The rigorous formulation of (5) is given in Theorem 27, which has been put in Appendix A for lack of readability. We propose here its specialization to the function $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$ with $\eta > 0$, $q > 1$ and $\gamma \in [0, 1)$. This function obviously satisfies conditions (1). We will refer to the associated forecasting strategy as "Poly INF". Here the (normalizing) function C has no closed form expression (this is a consequence of Abel's impossibility theorem). Actually this remark holds in general, hence the name of the general policy. However this does not lead to a major computational issue since, in the interval given by (2), $C(x)$ is the unique solution of $\phi(c) = 1$, where $\phi : c \mapsto \sum_{i=1}^K \psi(x_i - c)$ is a decreasing function. We will prove that Poly INF forecaster generates nicer probability updates than the exponentially weighted average forecasteras as, for bandits games (label efficient or not), it allows to remove the extraneous $\log K$ factor in the pseudo-regret bounds and some regret bounds.

Theorem 2 (General regret bound for Poly INF) *Let $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$ with $q > 1$, $\eta > 0$ and $\gamma \in [0, 1)$. Let $(v_{i,t})_{1\le i\le K,\, 1\le t\le n}$ be a sequence of nonnegative real numbers,*

$$B_t = \max_{1\le i\le K} v_{i,t}, \quad \text{and} \quad B = \max_t B_t.$$

If $\gamma = 0$ then INF satisfies:

$$\left(\max_{1\le i\le K} \sum_{t=1}^n v_{i,t}\right) - \sum_{t=1}^n \sum_{i=1}^K p_{i,t} v_{i,t} \le \frac{q}{q-1}\eta K^{1/q} + \frac{q}{2\eta} \exp\left(2\frac{q+1}{\eta}B\right) \sum_{t=1}^n B_t^2, \tag{6}$$

and

$$\left(\max_{1\le i\le K} \sum_{t=1}^n v_{i,t}\right) - \sum_{t=1}^n \sum_{i=1}^K p_{i,t} v_{i,t} \le \frac{q}{q-1}\eta K^{1/q} + \frac{qB}{\eta} \exp\left(\frac{8qB}{\eta}\right) \sum_{t=1}^n \sum_{i=1}^K p_{i,t} v_{i,t}. \tag{7}$$

For $\gamma > 0$, if we have $v_{i,t} = \frac{c_t}{p_{i,t}} \mathbb{1}_{i=I_t}$ for some random variable c_t taking values in $[0, c]$ with $0 < c < q\eta\left(\frac{\gamma}{(q-1)K}\right)^{(q-1)/q}$, then

$$(1-\gamma)\left(\max_{1\le i\le K} \sum_{t=1}^n v_{i,t}\right) - (1+\gamma\zeta) \sum_{t=1}^n \sum_{i=1}^K p_{i,t} v_{i,t} \le \frac{q}{q-1}\eta K^{\frac{1}{q}}, \tag{8}$$

where

$$\zeta = \frac{1}{(q-1)K}\left(\frac{(q-1)cK\mu(1+\mu)}{2\gamma\eta}\right)^q,$$

with

$$\mu = \exp\left\{\frac{2(q+1)c}{\eta}\left(\frac{K}{\gamma}\right)^{(q-1)/q}\left(1 - \frac{c}{q\eta}\left(\frac{(q-1)K}{\gamma}\right)^{(q-1)/q}\right)^{-q}\right\}.$$

In all this work, the parameters η, q and γ will be chosen such that ζ and μ act as numerical constants. To derive concrete bounds from the above theorem, most of the work lies in relating the left-hand side with the different notions of regret we consider. This task is trivial for the pseudo-regret. To derive high probability regret bounds, deviation inequalities for supermartingales are used on top of (6) and (8) (which hold with probability one). Finally, the expected regret bounds are obtained by integration of the high probability bounds.

As long as numerical constants do not matter, one can use (7) to recover the bounds obtained from (6). The advantage of (7) over (6) is that it allows to get regret bounds where the factor n is replaced by $G_{\max} = \max_{i=1,\dots,n} G_{i,n}$.

3. Exponentially Weighted Average Forecasters

The normalization by division that weighted average forecasters perform is different from the normalization by shift of the real axis that INF performs. Nonetheless, we can recover exactly the exponentially weighted average forecasters because of the special relation of the exponential with the addition and the multiplication.

Let $\psi(x) = \exp(\eta x) + \frac{\gamma}{K}$ with $\eta > 0$ and $\gamma \in [0,1)$. Then conditions (1) are clearly satisfied and (3) is equivalent to $\exp(-\eta C(x)) = \frac{1-\gamma}{\sum_{i=1}^K \exp(\eta x_i)}$, which implies

$$p_{i,t+1} = (1-\gamma)\frac{\exp(\eta V_{i,t})}{\sum_{j=1}^K \exp(\eta V_{i,t})} + \frac{\gamma}{K}.$$

In other words, for the full information case (label efficient or not), we recover the exponentially weighted average forecaster (with $\gamma = 0$) while for the bandit game we recover EXP3. For the label efficient bandit game, it does not give us the GREEN policy proposed in Allenberg et al. (2006) but rather the straightforward modification of the exponentially weighted average forecaster to this game (György and Ottucsák, 2006). Theorem 3 below gives a unified view on this algorithm for these four games. In the following, we will refer to this algorithm as the "exponentially weighted average forecaster" whatever the game is.

Theorem 3 (General regret bound for the exponentially weighted average forecaster) *Let $\psi(x) = \exp(\eta x) + \frac{\gamma}{K}$ with $\eta > 0$ and $\gamma \in [0,1)$. Let $(v_{i,t})_{1\le i\le K,\, 1\le t\le n}$ be a sequence of nonnegative real numbers,*

$$B_t = \max_{1\le i\le K} v_{i,t}, \quad \text{and} \quad B = \max_{1\le t\le n} B_t.$$

Consider the increasing function $\Theta : u \mapsto \frac{e^u-1-u}{u^2}$ equal to $1/2$ by continuity at zero. If $\gamma = 0$ then INF satisfies:

$$\left(\max_{1\le i\le K} \sum_{t=1}^n v_{i,t}\right) - \sum_{t=1}^n \sum_{i=1}^K p_{i,t} v_{i,t} \le \frac{\log K}{\eta} + \frac{\eta}{8}\sum_{t=1}^n B_t^2, \tag{9}$$

and

$$\left(\max_{1\le i\le K} \sum_{t=1}^n v_{i,t}\right) - \sum_{t=1}^n \sum_{i=1}^K p_{i,t} v_{i,t} \le \frac{\log K}{\eta} + \eta B \Theta(\eta B) \sum_{t=1}^n \sum_{i=1}^K p_{i,t} v_{i,t}. \tag{10}$$

If we have

$$\gamma \geq K\eta\Theta(\eta B)\max_{i,t} p_{i,t}v_{i,t}, \tag{11}$$

then INF satisfies:

$$(1-\gamma)\left(\max_{1\leq i\leq K}\sum_{t=1}^{n} v_{i,t}\right) - \sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t}v_{i,t} \leq (1-\gamma)\frac{\log K}{\eta}. \tag{12}$$

We have the same discussion about (9) and (10) than about (6) and (7): Inequality (10) allows to prove bounds where the factor n is replaced by $G_{\max} = \max_{i=1,\dots,n} G_{i,n}$, but at the price of worsened numerical constants, when compared to (9). We illustrate this point in Theorem 4, where (13) and (14) respectively comes from (9) and (10) .

The above theorem relies on the standard argument based on the cancellation of terms in a sum of logarithms of ratios (see Section C.2). For sake of comparison, we have applied our general result for INF forecasters, that is Theorem 27 (see Appendix A). This leads to the same result with worsened constants. Precisely, $\frac{\eta}{8}$ becomes $\frac{\eta}{2}\exp(2\eta B)$ in (9) while $\Theta(\eta B)$ becomes $\frac{\exp(2B\eta)[1+\exp(2B\eta)]}{2}$ in (11). This seems to be the price for having a theorem applying to a large class of forecasters.

4. The Full Information (FI) Game

The purpose of this section is to illustrate the general regret bounds given in Theorems 2 and 3 in the simplest case, when we set $v_{i,t} = g_{i,t}$, which is possible since the rewards for all arms are observed in the full information setting. The next theorem is given explicitly to show an easy application of Inequalities (9) and (10).

Theorem 4 (Exponentially weighted average forecaster in the FI game) *Let* $\psi(x) = \exp(\eta x)$ *with* $\eta > 0$. *Let* $v_{i,t} = g_{i,t}$. *Then in the full information game, INF satisfies*

$$\max_{1\leq i\leq K}\sum_{i=1}^{n} g_{i,t} - \sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t}g_{i,t} \leq \frac{\log K}{\eta} + \frac{\eta n}{8}. \tag{13}$$

and

$$\max_{1\leq i\leq K}\sum_{i=1}^{n} g_{i,t} - \sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t}g_{i,t} \leq \frac{\log K}{\eta} + \eta\Theta(\eta)\sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t}g_{i,t}. \tag{14}$$

In particular with $\eta = \sqrt{\frac{8\log K}{n}}$, *we get* $\mathbb{E}R_n \leq \sqrt{\frac{n}{2}\log K}$, *and there exists* $\eta > 0$ *such that*

$$\mathbb{E}R_n \leq \sqrt{2\mathbb{E}G_{\max}\log K}.$$

Proof It comes from (9) and (10) since we have $B \leq 1$ and $\sum_{t=1}^{n} B_t^2 \leq n$. The only nontrivial result is the last inequality. It obviously holds for any η when $\mathbb{E}G_{\max} = 0$, and is achieved for $\eta = \log\left(1 + \sqrt{2(\log K)/\mathbb{E}G_{\max}}\right)$, when $\mathbb{E}G_{\max} > 0$. Indeed, by taking the

expectation in (14), we get

$$\mathbb{E}\sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t}g_{i,t} \geq \frac{\eta\mathbb{E}G_{\max} - \log K}{\exp(\eta)-1} = \log\left(1+\sqrt{\frac{2\log K}{\mathbb{E}G_{\max}}}\right)\sqrt{\frac{(\mathbb{E}G_{\max})^3}{2\log K}} - \sqrt{\frac{\mathbb{E}G_{\max}\log K}{2}}$$
$$\geq \mathbb{E}G_{\max} - 2\sqrt{\frac{\mathbb{E}G_{\max}\log K}{2}},$$

where we use $\log(1+x) \geq x - \frac{x^2}{2}$ for any $x \geq 0$ in the last inequality. ■

Now we consider a new algorithm for the FI game, that is INF with $\psi(x) = \left(\frac{\eta}{-x}\right)^q$ and $v_{i,t} = g_{i,t}$.

Theorem 5 (Poly INF in the FI game) *Let $\psi(x) = \left(\frac{\eta}{-x}\right)^q$ with $\eta > 0$ and $q > 1$. Let $v_{i,t} = g_{i,t}$. Then in the full information game, INF satisfies:*

$$\max_{1\leq i\leq K}\sum_{i=1}^{n} g_{i,t} - \sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t}g_{i,t} \leq \frac{q}{q-1}\eta K^{1/q} + \exp\left(\frac{4q}{\eta}\right)\frac{qn}{2\eta}. \tag{15}$$

In particular with $q = 3\log K$ and $\eta = 1.8\sqrt{n\log K}$ we get

$$\mathbb{E}R_n \leq 7\sqrt{n\log K}.$$

Proof It comes from (6), $q+1 \leq 2q$ and $\sum_{t=1}^{n} B_t^2 \leq n$. ■

Remark 6 By using the Hoeffding-Azuma inequality (see, e.g., Lemma A.7 of Cesa-Bianchi and Lugosi, 2006), one can derive high probability bounds from (13) and (15): for instance, from (15), for any $\delta > 0$, with probability at least $1-\delta$, Poly INF satisfies:

$$R_n \leq \frac{q}{q-1}\eta K^{1/q} + \exp\left(\frac{4q}{\eta}\right)\frac{qn}{2\eta} + \sqrt{\frac{n\log(\delta^{-1})}{2}}.$$

5. The Limited Feedback Games

This section provides regret bounds for three limited feedback games: the label efficient game, the bandit game, and the mixed game, that is the label efficient bandit game.

5.1 Label Efficient Game (LE)

The variants of the LE game consider that the number of queried reward vectors is constrained either strictly or just in expectation. This section considers successively these two cases.

5.1.1 Constraint on the Expected Number of Queried Reward Vectors

As in Section 4, the purpose of this section is to show how to use INF in order to recover known minimax bounds (up to constant factors) in a slight modification of the LE game: the simple LE game, in which the requirement is that the *expected* number of queried reward vectors should be less or equal to m.

Let us consider the following policy. At each round, we draw a Bernoulli random variable Z_t, with parameter $\varepsilon = m/n$, to decide whether we ask for the gains or not. Note that we do not fulfill exactly the requirement of the LE game as we might ask a bit more than m reward vectors, but we fulfill the one of the simple LE game. We do so in order to avoid technical details and focus on the main argument of the proof. The exact LE game will be addressed in Section 5.1.2, where, in addition, we will prove bounds on the expected regret $\mathbb{E}R_n$ instead of just the pseudo-regret $\overline{R}_n$.

In this section, the estimate of $g_{i,t}$ is $v_{i,t} = \frac{g_{i,t}}{\varepsilon} Z_t$, which is observable since the rewards at time t for all arms are observed when $Z_t = 1$.

Theorem 7 (Exponentially weighted average forecaster in the simple LE game) *Let $\psi(x) = \exp(\eta x)$ with $\eta = \frac{\sqrt{8m \log K}}{n}$. Let $v_{i,t} = \frac{g_{i,t}}{\varepsilon} Z_t$ with $\varepsilon = \frac{m}{n}$. Then in the simple LE game, INF satisfies*

$$\overline{R}_n \leq n \sqrt{\frac{\log K}{2m}}.$$

Proof The first inequality comes from (9). Since we have $B_t \leq Z_t/\varepsilon$ and $v_{i,t} = \frac{g_{i,t}}{\varepsilon} Z_t$, we obtain

$$\left(\max_{1 \leq i \leq K} \sum_{t=1}^{n} g_{i,t} \frac{Z_t}{\varepsilon} \right) - \sum_{t=1}^{n} \sum_{i=1}^{K} p_{i,t} g_{i,t} \frac{Z_t}{\varepsilon} \leq \frac{\log K}{\eta} + \frac{\eta}{8\varepsilon^2} \sum_{t=1}^{n} Z_t,$$

hence, by taking the expectation of both sides,

$$\overline{R}_n = \left(\max_{1 \leq i \leq K} \mathbb{E} \sum_{t=1}^{n} g_{i,t} \frac{Z_t}{\varepsilon} \right) - \mathbb{E} \sum_{t=1}^{n} \sum_{i=1}^{K} p_{i,t} g_{i,t} \frac{Z_t}{\varepsilon} \leq \frac{\log K}{\eta} + \frac{n\eta}{8\varepsilon} = \frac{\log K}{\eta} + \frac{n^2 \eta}{8m}.$$

Straightforward computations conclude the proof. ■

A similar result can be proved for the INF forecaster with $\psi(x) = \left(\frac{\eta}{-x}\right)^q$, $\eta > 0$ and q of order $\log K$. We do not state it since we will prove a stronger result in the next section.

5.1.2 Hard Constraint on the Number of Queried Reward Vectors

The goal of this section is to push the idea that by using high probability bounds as an intermediate step, one can control the expected regret $\mathbb{E}R_n = \mathbb{E} \max_{1 \leq i \leq K} \sum_{t=1}^{n} \left(g_{i,t} - g_{I_t,t} \right)$ instead of just the pseudo-regret $\overline{R}_n = \max_{1 \leq i \leq K} \mathbb{E} \sum_{t=1}^{n} \left(g_{i,t} - g_{I_t,t} \right)$. Most previous works have obtained results for $\overline{R}_n$. These results are interesting for oblivious opponents, that is when the adversary's choices of the rewards do not depend on the past draws and obtained rewards, since in this case Proposition 33 in Appendix D shows that one can extend bounds on the pseudo-regret $\overline{R}_n$ to the expected regret $\mathbb{E}R_n$. For non-oblivious opponents, upper bounds on $\overline{R}_n$ are rather weak statements and high probability bounds on R_n or bounds on

$\mathbb{E}R_n$ are desirable. In Auer (2002) and Cesa-Bianchi and Lugosi (2006), high probability bounds on R_n have been given. Unfortunately, the policies proposed there are depending on the confidence level of the bound. As a consequence, the resulting best bound on $\mathbb{E}R_n$, obtained by choosing the policies with confidence level parameter of order $1/n$, has an extraneous $\log n$ term. Specifically, from Theorem 6.2 of Cesa-Bianchi and Lugosi (2006), one can immediately derive $\mathbb{E}R_n \leq 8n\sqrt{\frac{\log(4K)+\log(n)}{m}} + 1$. The theorems of this section essentially show that the $\log n$ term can be removed.

As in Section 5.1.1, we still use a draw of a Bernoulli random variable Z_t to decide whether we ask for the gains or not. The difference is that, if $\sum_{s=1}^{t-1} Z_s \geq m$, we do not ask for the gains (as we are not allowed to do so). To avoid that this last constraint interferes in the analysis, the parameter of the Bernoulli random variable is set to $\varepsilon = \frac{3m}{4n}$ and the probability of the event $\sum_{t=1}^{n} Z_t > m$ is upper bounded. The estimate of $g_{i,t}$ remains $v_{i,t} = \frac{g_{i,t}}{\varepsilon} Z_t$.

Theorem 8 (Exponentially weighted average forecaster in the LE game) *Let $\psi(x) = \exp(\eta x)$ with $\eta = \frac{\sqrt{m \log K}}{n}$. Let $v_{i,t} = \frac{g_{i,t}}{\varepsilon} Z_t$ with $\varepsilon = \frac{3m}{4n}$. Then in the LE game, for any $\delta > 0$, with probability at least $1 - \delta$, INF satisfies:*

$$R_n \leq n\sqrt{\frac{27 \log(2K\delta^{-1})}{m}},$$

and

$$\mathbb{E}R_n \leq n\sqrt{\frac{27 \log(6K)}{m}}.$$

Theorem 9 (Poly INF in the LE game) *Let $\psi(x) = \left(\frac{\eta}{-x}\right)^q$ with $q = 3\log(2K)$ and $\eta = 2n\sqrt{\frac{\log(2K)}{m}}$. Let $v_{i,t} = \frac{g_{i,t}}{\varepsilon} Z_t$ with $\varepsilon = \frac{3m}{4n}$. Then in the LE game, for any $\delta > 0$, with probability at least $1 - \delta$, INF satisfies:*

$$R_n \leq \left(8 - \sqrt{27}\right) n\sqrt{\frac{\log(2K)}{m}} + n\sqrt{\frac{27 \log(2K\delta^{-1})}{m}},$$

and

$$\mathbb{E}R_n \leq 8n\sqrt{\frac{\log(6K)}{m}}.$$

5.2 Bandit Game

This section is cut into two parts. In the first one, from Theorem 2 and Theorem 3, we derive upper bounds on the pseudo-regret $\overline{R}_n = \max_{1 \leq i \leq K} \mathbb{E} \sum_{t=1}^{n} \left(g_{i,t} - g_{I_t,t}\right)$. To bound the expected regret $\mathbb{E}R_n = \mathbb{E} \max_{1 \leq i \leq K} \sum_{t=1}^{n} \left(g_{i,t} - g_{I_t,t}\right)$, we will then use high probability bounds on top of the use of these theorems. Since this makes the proofs more intricate, we have chosen to provide the less general results, but easier to obtain, in Section 5.2.1 and the more general ones in Section 5.2.2.

The main results here are that, by using the INF with a polynomial function ψ, we obtain an upper bound of order $\sqrt{nK}$ for $\overline{R}_n$, which imply a bound of order $\sqrt{nK}$ on $\mathbb{E}R_n$ for oblivious adversaries (Proposition 33 in Appendix D). In the general case (containing

the non-oblivious opponent), we show an upper bound of order $\sqrt{nK\log K}$ on $\mathbb{E}R_n$. We conjecture that this bound cannot be improved, that is the opponent may take advantage of the past to make the player pay a regret with the extra logarithmic factor (see Remark 14).

5.2.1 Bounds on the Pseudo-regret

In this section, the estimate of $g_{i,t}$ is $v_{i,t} = \frac{g_{i,t}}{p_{i,t}}\mathbb{1}_{I_t=i}$, which is observable since the reward $g_{I_t,t}$ is revealed at time t.

Theorem 10 (Exponentially weighted average forecaster in the bandit game) *Let $\psi(x) = \exp(\eta x) + \frac{\gamma}{K}$ with $1 > \gamma \geq \frac{4\eta K}{5} > 0$. Let $v_{i,t} = \frac{g_{i,t}}{p_{i,t}}\mathbb{1}_{I_t=i}$. Then in the bandit game, INF satisfies:*

$$\overline{R}_n \leq \frac{\log K}{\eta} + \gamma \max_{1\leq i\leq K} \mathbb{E}G_{i,n}.$$

In particular, for $\gamma = \min\left(\frac{1}{2}, \sqrt{\frac{4K\log K}{5n}}\right)$ and $\eta = \sqrt{\frac{5\log K}{4nK}}$, we have

$$\overline{R}_n \leq \sqrt{\frac{16}{5}nK\log K}.$$

Proof One simply needs to note that for $5\gamma \geq 4K\eta$, (11) is satisfied (since $B = K/\gamma$), and thus (12) can be rewritten into

$$(1-\gamma)\left(\max_{1\leq i\leq K}\sum_{t=1}^{n}\frac{g_{i,t}}{p_{i,t}}\mathbb{1}_{I_t=i}\right) - \sum_{t=1}^{n} g_{I_t,t} \leq (1-\gamma)\frac{\log K}{\eta}.$$

By taking the expectation, we get

$$\overline{R}_n \leq (1-\gamma)\frac{\log K}{\eta} + \gamma \max_{1\leq i\leq K} \mathbb{E}G_{i,n}.$$

For the numerical application, since $\overline{R}_n \leq n$, the bound is trivial $\sqrt{(4K\log K)/(5n)} < \frac{1}{2}$. Otherwise, it is a direct application of the general bound. ■

Theorem 11 (Poly INF in the bandit game) *Consider $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$ with $\gamma = \min\left(\frac{1}{2}, \sqrt{\frac{3K}{n}}\right)$, $\eta = \sqrt{5n}$ and $q = 2$. Let $v_{i,t} = \frac{g_{i,t}}{p_{i,t}}\mathbb{1}_{I_t=i}$. Then in the bandit game, INF satisfies:*

$$\overline{R}_n \leq 8\sqrt{nK}.$$

Proof The bound is trivial when $\frac{1}{2} \leq \sqrt{\frac{3K}{n}}$. So we consider hereafter that $\gamma = \sqrt{\frac{3K}{n}} < \frac{1}{2}$. By taking the expectation in (8) and letting $\bar{G}_{\max} = \max_{1\leq i\leq K}\mathbb{E}G_{i,n}$, we obtain that for $\gamma > (q-1)K(q\eta)^{q/(1-q)} > 0$ (condition coming from the condition on c for (8)),

$$(1-\gamma)\bar{G}_{\max} - (1+\gamma\zeta)\mathbb{E}\sum_{t=1}^{n} g_{I_t,t} \leq \frac{q}{q-1}\eta K^{\frac{1}{q}},$$

with

$$\zeta = \frac{1}{(q-1)K}\Big(\frac{(q-1)K\mu(1+\mu)}{2\gamma\eta}\Big)^q,$$

and

$$\mu = \exp\left\{\frac{2(q+1)}{\eta}\left(\frac{K}{\gamma}\right)^{(q-1)/q}\left(1-\frac{1}{q\eta}\left(\frac{(q-1)K}{\gamma}\right)^{(q-1)/q}\right)^{-q}\right\}.$$

We thus have

$$\overline{R}_n \leq \gamma(1+\zeta)\bar{G}_{\max} + \frac{q}{q-1}\eta K^{\frac{1}{q}} \leq \gamma(1+\zeta)n + \frac{q}{q-1}\eta K^{\frac{1}{q}}.$$

The desired inequality is trivial when $\sqrt{K/n} \geq 1/8$. So we now consider that $\sqrt{K/n} < 1/8$. For $\gamma = \sqrt{3K/n}$, $\eta = \sqrt{5n}$ and $q = 2$, the condition on γ is satisfied (since $\sqrt{K/n} < 1/8$), and we have $\frac{1}{\eta}(\frac{K}{\gamma})^{(q-1)/q} \leq 0.121$, hence $\mu \leq 2.3$, $\zeta \leq 1$ and $\overline{R}_n \leq 8\sqrt{nK}$. ■

We have arbitrarily chosen $q = 2$ to provide an explicit upper bound. More generally, it is easy to check from the proof of Theorem 11 that for any real number $q > 1$, we obtain the convergence rate $\sqrt{nK}$, provided that γ and η are respectively taken of order $\sqrt{K/n}$ and $\sqrt{nK}/K^{1/q}$.

5.2.2 High Probability Bounds and Bounds on the Expected Regret

Theorems 10 and 11 provide upper bounds on $\overline{R}_n = \max_{1\leq i\leq K}\mathbb{E}\sum_{t=1}^n\big(g_{i,t} - g_{I_t,t}\big)$. To bound $\mathbb{E}R_n = \mathbb{E}\max_{1\leq i\leq K}\sum_{t=1}^n\big(g_{i,t} - g_{I_t,t}\big)$, we will use high probability bounds. First we need to modify the estimates of $g_{i,t}$ by considering $v_{i,t} = \frac{g_{i,t}}{p_{i,t}}\mathbb{1}_{I_t=i} + \frac{\beta}{p_{i,t}}$ with $0 < \beta \leq 1$, as was proposed in Auer (2002),[2] or $v_{i,t} = -\frac{\mathbb{1}_{I_t=i}}{\beta}\log\big(1 - \frac{\beta g_{i,t}}{p_{i,t}}\big)$ as we propose here.

Theorem 12 (Exponentially weighted average forecaster in the bandit game) *Consider* $\psi(x) = \exp(\eta x) + \frac{\gamma}{K}$ *with* $\gamma = \min\left(\frac{2}{3}, 2\sqrt{\frac{K\log(3K)}{n}}\right)$ *and* $\eta = 2\sqrt{\frac{\log(3K)}{Kn}}$. *Let* $v_{i,t} = -\frac{\mathbb{1}_{I_t=i}}{\beta}\log\big(1 - \frac{\beta g_{i,t}}{p_{i,t}}\big)$ *with* $\beta = \sqrt{\frac{\log(3K)}{2Kn}}$. *Then in the bandit game, against any adversary (possibly a non-oblivious one), for any* $\delta > 0$, *with probability at least* $1-\delta$, *INF satisfies:*

$$R_n \leq 3\sqrt{nK\log(3K)} + \sqrt{\frac{2nK}{\log(3K)}}\log(K\delta^{-1}),$$

and also $\mathbb{E}R_n \leq (3+\sqrt{2})\sqrt{nK\log(3K)}$.

This theorem is similar to Theorem 6.10 of Cesa-Bianchi and Lugosi (2006). The main difference here is that the high probability bound holds for any confidence level, and not

2. The technical reason for this modification, which may appear surprising as it introduces a bias in the estimate of $g_{i,t}$, is that it allows to have high probability upper bounds with the correct rate on the difference $\sum_{t=1}^n g_{i,t} - \sum_{t=1}^n v_{i,t}$. A second reason for this modification (but useless for this particular section) is that it allows to track the best expert (see Section 7).

only for a confidence level depending on the algorithm. As a consequence, our algorithm, unlike the one proposed in previous works, satisfies both a high probability bound and an expected regret bound of order $\sqrt{nK\log(K)}$.

Theorem 13 (Poly INF in the bandit game) *Let $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$ with $\eta = 2\sqrt{n}$, $q = 2$ and $\gamma = \min\left(\frac{1}{2}, 3\sqrt{\frac{K}{n}}\right)$. Consider $v_{i,t} = -\frac{\mathbb{1}_{I_t=i}}{\beta}\log\left(1 - \frac{\beta g_{i,t}}{p_{i,t}}\right)$ with $\beta = \frac{1}{\sqrt{2Kn}}$. Then in the bandit game, against a deterministic adversary, for any $\delta > 0$, with probability at least $1-\delta$, INF satisfies:*

$$R_n \leq 9\sqrt{nK} + \sqrt{2nK}\log(\delta^{-1}). \tag{16}$$

Against an oblivious adversary, it satisfies

$$\mathbb{E}R_n \leq 10\sqrt{nK}. \tag{17}$$

Moreover in the general case (containing the non-oblivious opponent), with the following parameters $q = 2$, $\gamma = \min\left(\frac{1}{2}, 3\sqrt{\frac{K\log(3K)}{n}}\right)$, $\eta = 2\sqrt{\frac{n}{\log(3K)}}$ and $\beta = \sqrt{\frac{\log(3K)}{2nK}}$, it satisfies with probability at least $1-\delta$,

$$R_n \leq 9\sqrt{nK\log(3K)} + \sqrt{\frac{2nK}{\log(3K)}}\log(\delta^{-1}),$$

and

$$\mathbb{E}R_n \leq 9\sqrt{nK\log(3K)}.$$

Remark 14 *We conjecture that the order $\sqrt{nK\log K}$ of the bound on $\mathbb{E}R_n$ cannot be improved in the general case containing the non-oblivious opponent. Here is the main argument to support our conjecture. Consider an adversary choosing all rewards to be equal to one until time $n/2$ (say n is even to simplify). Then, let $\hat{k}$ denote the arm for which the estimate $V_{i,n/2} = \sum_{1\leq t\leq n/2} v_{i,t}$ of the cumulative reward of arm i is the smallest. After time $n/2$, all rewards are chosen to be equal to zero except for arm $\hat{k}$ for which the rewards are still chosen to be equal to 1. Since we believe that with high probability, $\max_{1\leq i\leq K} V_{i,n/2} - \min_{j\in\{1,\dots,K\}} V_{j,n/2} \geq \kappa\sqrt{nK\log K}$ for some small enough $\kappa > 0$, it seems that the INF algorithm achieving a bound of order $\sqrt{nK}$ on $\mathbb{E}R_n$ in the oblivious setting will suffer an expected regret of order at least $\sqrt{nK\log K}$. While this does not prove the conjecture as one can design other algorithms, it makes the conjecture likely to hold.*

5.3 Label Efficient Bandit Game (LE Bandit)

The following theorems concern the simple LE bandit game, in which the requirement is that the *expected* number of queried rewards should be less or equal to m. We consider the following policy. At each round, we draw a Bernoulli random variable Z_t, with parameter $\varepsilon = m/n$, to decide whether the gain of the chosen arm is revealed or not. Note that this policy does not fulfil exactly the requirement of the LE bandit game as we might ask a bit more than m rewards, but, as was argued in Section 5.1.2, it can be modified in order to fulfil the hard constraint of the game. The theoretical guarantees are then the same (up to numerical constant factors).

Theorem 15 (Exponentially weighted average forecaster in the simple LE bandit game) *Let* $\psi(x) = \exp(\eta x) + \frac{\gamma}{K}$ *with* $\gamma = \min\left(\frac{1}{2}, \sqrt{\frac{4K\log K}{5m}}\right)$ *and* $\eta = \frac{1}{n}\sqrt{\frac{5m\log K}{4K}}$. *Let* $v_{i,t} = g_{i,t}\frac{\mathbb{1}_{I_t=i}}{p_{i,t}}\frac{Z_t}{\varepsilon}$ *with* $\varepsilon = \frac{m}{n}$. *Then in the simple LE bandit game, INF satisfies:*

$$\overline{R}_n \leq n\sqrt{\frac{16K\log K}{5m}}.$$

Proof One simply needs to note that for $5\gamma \geq \frac{4K\eta}{\varepsilon}$, (11) is satisfied, and thus by taking the expectation in (12), we get

$$\overline{R}_n \leq (1-\gamma)\frac{\log K}{\eta} + \gamma\mathbb{E}\max_{1\leq i\leq K} V_{i,n} \leq (1-\gamma)\frac{\log K}{\eta} + \gamma n.$$

We thus have

$$\overline{R}_n \leq \frac{n}{m}\left((1-\gamma)\frac{\log K}{\eta/\varepsilon} + \gamma m\right).$$

The numerical application for the term in parenthesis is then exactly the same as the one proposed in the proof of Theorem 10 (with n and η respectively replaced by m and η/ε). ■

Theorem 16 (Poly INF in the simple LE bandit game) *Let* $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$ *with* $\gamma = \min\left(\frac{1}{2}, \sqrt{\frac{3K}{m}}\right)$, $\eta = n\sqrt{\frac{5}{m}}$ *and* $q = 2$. *Let* $v_{i,t} = g_{i,t}\frac{\mathbb{1}_{I_t=i}}{p_{i,t}}\frac{Z_t}{\varepsilon}$ *with* $\varepsilon = \frac{m}{n}$. *Then in the simple LE bandit game, INF satisfies:*

$$\overline{R}_n \leq 8n\sqrt{\frac{K}{m}}.$$

Proof By taking the expectation in (8) and letting $\bar{G}_{\max} = \max_{1\leq i\leq K}\mathbb{E}G_{i,n}$, we obtain that for $\gamma > (q-1)K(q\eta\varepsilon)^{q/(1-q)} > 0$ (condition coming from the condition on c for (8)),

$$(1-\gamma)\bar{G}_{\max} - (1+\gamma\zeta)\mathbb{E}\sum_{t=1}^{n} g_{I_t,t} \leq \frac{q}{q-1}\eta K^{\frac{1}{q}},$$

with

$$\zeta = \frac{1}{(q-1)K}\Big(\frac{(q-1)K\mu(1+\mu)}{2\gamma\eta\varepsilon}\Big)^q,$$

and

$$\mu = \exp\left\{\frac{2(q+1)}{\eta\varepsilon}\left(\frac{K}{\gamma}\right)^{(q-1)/q}\left(1 - \frac{1}{q\eta}\left(\frac{(q-1)K}{\gamma}\right)^{(q-1)/q}\right)^{-q}\right\}.$$

We thus have

$$\overline{R}_n \leq \frac{n}{m}\left(\gamma(1+\zeta)m + \frac{q}{q-1}(\eta\varepsilon)K^{\frac{1}{q}}\right).$$

The numerical application for the term in parenthesis is exactly the same than the one proposed in the proof of Theorem 11 (with n and η respectively replaced by m and $\eta\varepsilon$). ■

Both previous theorems only consider the pseudo-regret. By estimating $g_{i,t}$ differently, we obtain the following result for the regret.

Theorem 17 (Poly INF in the simple LE bandit game) *Let $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$ with $\eta = 2n/\sqrt{m}$, $q = 2$ and $\gamma = \min\left(\frac{1}{2}, 3\sqrt{\frac{K}{m}}\right)$. Consider $v_{i,t} = -\frac{\mathbb{1}_{I_t=i} Z_t}{\beta} \log\left(1 - \frac{\beta g_{i,t}}{\varepsilon p_{i,t}}\right)$ with $\beta = \frac{1}{n}\sqrt{\frac{m}{2K}}$. Then in the simple LE bandit game, against a deterministic adversary, for any $\delta > 0$, with probability at least $1 - \delta$, INF satisfies:*

$$R_n \leq 10.7n\sqrt{\frac{K}{m}} + 3.1n\sqrt{\frac{K}{m}}\log(\delta^{-1}).$$

Against an oblivious adversary, it satisfies

$$\mathbb{E}R_n \leq 13\sqrt{nK}.$$

Moreover in the general case (containing the non-oblivious opponent), with the following parameters $q = 2$, $\gamma = \min\left(\frac{1}{2}, 3\sqrt{\frac{K\log(3K)}{m}}\right)$, $\eta = 2n/\sqrt{m\log(3K)}$ and $\beta = \frac{1}{n}\sqrt{\frac{m\log(3K)}{2K}}$, it satisfies with probability at least $1 - \delta$,

$$R_n \leq 10n\sqrt{\frac{K\log(3K)}{m}} + 3.5n\sqrt{\frac{K}{m\log(3K)}}\log(\delta^{-1}),$$

and

$$\mathbb{E}R_n \leq 13n\sqrt{\frac{\log(3K)}{m}}.$$

A similar result can be obtained for Exp INF, at the price of an additional logarithmic term in K against oblivious (deterministic or not) adversaries. We omit the details.

6. Regret Bounds Scaling with the Optimal Arm Rewards

In this section, we provide regret bounds for bandit games depending on the performance of the optimal arm: in these bounds, the factor n is essentially replaced by

$$G_{\max} = \max_{i=1,\ldots,n} G_{i,n},$$

where $G_{i,n} = \sum_{t=1}^{n} g_{i,t}$. Such a bound has been proved on the expected regret for deterministic adversaries in the seminal work of Auer et al. (2002b). Here, by using a new biased estimate of $g_{i,t}$, that is $v_{i,t} = -\frac{\mathbb{1}_{I_t=i}}{\beta}\log\left(1 - \frac{\beta g_{i,t}}{p_{i,t}}\right)$, we obtain a bound holding with high probability and we also consider its extension to any adversary.

The bounds presented here are especially interesting when $G_{\max} \ll n$: this typically occurs in online advertizing where the different arms are the ads that can be put on the website and where the probability that a user clicks on an ad banner (and thus induces a reward to the webpage owner) is very low. For deterministic adversaries, as in the bandit game, the $\log K$ factor appearing in the exponentially weighted average forecaster regret bound disappears in the Poly INF regret bound as follows.

Theorem 18 *Let G_0 be a real number such that $G_0 \geq 81K$. Let $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$ with $\eta = 2\sqrt{G_0}$, $q = 2$ and $\gamma = 3\sqrt{\frac{K}{G_0}}$. Let $v_{i,t} = -\frac{\mathbb{1}_{I_t=i}}{\beta}\log\left(1 - \frac{\beta g_{i,t}}{p_{i,t}}\right)$ with $\beta = \frac{1}{\sqrt{2KG_0}}$. Then in the bandit game, against a deterministic adversary, for any $\delta > 0$, with probability at least $1-\delta$, INF satisfies:*

$$R_n \leq 4.5\sqrt{K\frac{G_{\max}^2}{G_0}} + 4\sqrt{KG_0} + \sqrt{2KG_0}\log(\delta^{-1}). \tag{18}$$

For fully oblivious adversaries, for any $\delta > 0$, with probability at least $1-\delta$, INF satisfies:

$$R_n \leq 4.5\sqrt{K\frac{G_{\max}^2}{G_0}} + 4\sqrt{KG_0} + \sqrt{2KG_0}\log(2\delta^{-1}) + \sqrt{8\log(2K\delta^{-1})G_{\max}}. \tag{19}$$

For the choice $G_0 = n$, the high probability upper bounds are of the order of $\sqrt{nK} + \sqrt{nK}\log(\delta^{-1})$. The interest of the theorem is to provide a policy which, for small $G_{\max}$, leads to smaller regret bounds, as long as G_0 is taken much smaller than n and but not much smaller than $G_{\max}$. For deterministic adversaries, $G_{\max}$ is nonrandom, and provided that we know its order, one has interest of taking G_0 of this order. Precisely, we have the following corollary for deterministic adversaries.

Corollary 19 *Let $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$ with $\eta = 2\sqrt{G_{\max}}$, $q = 2$ and $\gamma = \min\left(\frac{1}{2}, 3\sqrt{\frac{K}{G_{\max}}}\right)$. Consider $v_{i,t} = -\frac{\mathbb{1}_{I_t=i}}{\beta}\log\left(1 - \frac{\beta g_{i,t}}{p_{i,t}}\right)$ with $\beta = \frac{1}{\sqrt{2KG_{\max}}}$. Then in the bandit game, against a deterministic adversary, for any $\delta > 0$, with probability at least $1-\delta$, INF satisfies:*

$$R_n \leq 9\sqrt{KG_{\max}} + \sqrt{2KG_{\max}}\log(\delta^{-1}), \tag{20}$$

and

$$\mathbb{E}R_n \leq 10\sqrt{KG_{\max}}. \tag{21}$$

For more general adversaries than fully oblivious ones, we have the following result in which the $\log K$ factor reappears.

Theorem 20 *Let $G_0 \geq 81K\log(3K)$. Let $\psi(x) = \left(\frac{\eta}{-x}\right)^q + \frac{\gamma}{K}$ with $q = 2$, $\gamma = 3\sqrt{\frac{K\log(3K)}{G_0}}$ and $\eta = 2\sqrt{\frac{G_0}{\log(3K)}}$. Let $v_{i,t} = -\frac{\mathbb{1}_{I_t=i}}{\beta}\log\left(1 - \frac{\beta g_{i,t}}{p_{i,t}}\right)$ with $\beta = \sqrt{\frac{\log(3K)}{2KG_0}}$. Then in the bandit game, against any adversary (possibly a non-oblivious one), for any $\delta > 0$, with probability at least $1-\delta$, INF satisfies:*

$$R_n \leq \frac{9}{2}\sqrt{\frac{G_{\max}^2}{G_0}K\log(3K)} + 4\sqrt{\frac{KG_0}{\log(3K)}} + \sqrt{\frac{2KG_0}{\log(3K)}}\log(K\delta^{-1}).$$

This last result concerning Poly INF is similar to the following one concerning the exponentially weighted average forecaster: the advantage of Poly INF only appears when it allows to remove the $\log K$ factor.

Theorem 21 *Let $G_0 > 4K\log(3K)$. Let $\psi(x) = \exp(\eta x) + \frac{\gamma}{K}$ with $\gamma = 2\sqrt{\frac{K\log(3K)}{G_0}}$ and $\eta = 2\sqrt{\frac{\log(3K)}{KG_0}}$. Let $v_{i,t} = -\frac{\mathbb{1}_{I_t=i}}{\beta}\log\left(1-\frac{\beta g_{i,t}}{p_{i,t}}\right)$ with $\beta = \sqrt{\frac{\log(3K)}{2KG_0}}$. Then in the bandit game, against any adversary (possibly a non-oblivious one), for any $\delta > 0$, with probability at least $1-\delta$, INF satisfies:*

$$R_n \leq \frac{5}{2}\sqrt{\frac{G_{\max}^2}{G_0}K\log(3K)} + \frac{1}{2}\sqrt{KG_0\log(3K)} + \sqrt{\frac{2KG_0}{\log(3K)}}\log(K\delta^{-1}).$$

7. Tracking the Best Expert in the Bandit Game

In the previous sections, the cumulative gain of the forecaster was compared to the cumulative gain of the best single expert. Here, it will be compared to more flexible strategies that are allowed to switch actions. We will use

$$v_{i,t} = g_{i,t}\frac{\mathbb{1}_{I_t=i}}{p_{i,t}} + \frac{\beta}{p_{i,t}},$$

with $0 < \beta \leq 1$. The β term introduces a bias in the estimate of $g_{i,t}$, that constrains the differences $\max_{1\leq i\leq K} V_{i,t} - \min_{1\leq j\leq K} V_{j,t}$ to be relatively small. This is the key property in order to track the best switching strategy, provided that the number of switches is not too large. A switching strategy is defined by a vector $(i_1,\ldots,i_n) \in \{1,\ldots,K\}^n$. Its size is defined by

$$\mathcal{S}(i_1,\ldots,i_n) = \sum_{t=1}^{n-1}\mathbb{1}_{i_{t+1}\neq i_t},$$

and its cumulative gain is

$$G_{(i_1,\ldots,i_n)} = \sum_{t=1}^{n} g_{i_t,t}.$$

The regret of a forecaster with respect to the best switching strategy with S switches is then given by:

$$R_n^S = \max_{(i_1,\ldots,i_n):\,\mathcal{S}(i_1,\ldots,i_n)\leq S} G_{(i_1,\ldots,i_n)} - \sum_{t=1}^{n} g_{I_t,t}.$$

Theorem 22 (INF for tracking the best expert in the bandit game) *Let $s = S\log\left(\frac{3nK}{S}\right) + 2\log K$ with the natural convention $S\log(3nK/S) = 0$ for $S = 0$. Let $v_{i,t} = g_{i,t}\frac{\mathbb{1}_{I_t=i}}{p_{i,t}} + \frac{\beta}{p_{i,t}}$ with $\beta = 3\sqrt{\frac{s}{nK}}$. Let $\psi(x) = \exp(\eta x) + \frac{\gamma}{K}$ with $\gamma = \min\left(\frac{1}{2}, \sqrt{\frac{Ks}{2n}}\right)$ and $\eta = \frac{1}{5}\sqrt{\frac{s}{nK}}$. Then in the bandit game, for any $0 \leq S \leq n-1$, for any $\delta > 0$, with probability at least $1-\delta$, INF satisfies:*

$$R_n^S \leq 7\sqrt{nKs} + \sqrt{\frac{nK}{s}}\log(\delta^{-1}),$$

and

$$\mathbb{E}R_n^S \leq 7\sqrt{nKs}.$$

Note that for $S = 0$, we have $R_n^S = R_n$, and we recover an expected regret bound of order $\sqrt{nK \log K}$ similar to the one of Theorem 12.

Remark 23 Up to constant factors, the same bounds as the ones of Theorem 22 can be obtained (via a tedious proof not requiring new arguments than the ones presented in this work) for the INF forecaster using $\psi(x) = \frac{c_1}{K}\left(\frac{\sqrt{snK}}{-x}\right)^{c_3 s} + c_2\sqrt{\frac{s}{nK}}$, with $s = S\log\left(\frac{enK}{S}\right) + \log(2K)$ and appropriate constants c_1, c_2 and c_3.

8. Gains vs Losses, Unsigned Games vs Signed Games

To simplify, we have considered so far that the rewards were in $[0,1]$. Here is a trivial argument which shows how to transfer our analysis to loss games (i.e., games with only non-positive rewards), and more generally to signed games (i.e., games in which the rewards can be positive and negative). If the rewards, denoted now $g'_{i,t}$, are in some interval $[a,b]$ potentially containing zero, we set $g_{i,t} = \frac{g'_{i,t}-a}{b-a} \in [0,1]$. Then we can apply our analysis to:

$$\max_{i\in\{1,\dots,K\}} \sum_{t=1}^n g_{i,t} - \sum_{t=1}^n g_{I_t,t} = \frac{1}{b-a}\left(\max_{i\in\{1,\dots,K\}} \sum_{t=1}^n g'_{i,t} - \sum_{t=1}^n g'_{I_t,t}\right).$$

A less straightforward analysis can be done by looking at the INF algorithm directly applied to the observed rewards (and not to the renormalized rewards). In this case, as it was already noted in Remark 6.5 of Cesa-Bianchi and Lugosi (2006), the behavior of the algorithm may be very different for loss and gain games. However it can be proved that our analysis still holds up to constant factors (one has to go over the proofs and make appropriate modifications since for simplicity, we have presented the general results concerning INF under the assumptions that the estimates $v_{i,t}$ are nonnegative). In Section 6, we provide regret bounds scaling with the cumulative reward of the optimal arm. For this kind of results, renormalizing will not lead to regret bounds scaling with the cumulative reward before renormalization of the optimal arm, and consequently, the study of INF directly applied to the observed rewards is necessary. In particular, obtaining low regret bounds when the optimal arm has small cumulative loss would require appropriate modifications in the proof.

9. Stochastic Bandit Game

By considering the deterministic case when the rewards are $g_{i,t} = 1$ if $i = 1$ and $g_{i,t} = 0$ otherwise, it can be proved that the INF policies considered in Theorem 10 and Theorem 11 have a pseudo-regret lower bounded by $\sqrt{nK}$. In this simple setting, and more generally in most of the stochastic multi-armed bandit problems, one would like to suffer a much smaller regret.

We recall that in the stochastic bandit considered in this section, the rewards $g_{i,1}, \dots, g_{i,n}$ are independent and drawn from a fixed distribution ν_i on $[0,1]$ for each arm i, and the reward vectors $g_1, \dots, g_n$ are independent.[3] The suboptimality of an arm i is then measured

3. Note that we do not assume independence of $g_{1,t}, \dots, g_{K,t}$ for each t. This assumption is usually made in the literature, but is often useless. In our work, assuming it would just have improved Proposition 36 by a constant factor, and would not have improved the constant in Theorem 24.

by $\Delta_i = \max_{1\leq j\leq K} \mu_j - \mu_i$ where μ_i is the mean of ν_i. We provide now a strategy achieving a $\sqrt{nK}$ regret in the worst case, and a much smaller regret as soon as the Δ_i of the suboptimal arms are much larger than $\sqrt{K/n}$.

Let $\widehat{\mu}_{i,s}$ be the empirical mean of arm i after s draws of this arm. Let $T_i(t)$ denote the number of times we have drawn arm i on the first t rounds. In this section, we propose a policy, called MOSS (Minimax Optimal Strategy in the Stochastic case), inspired by the UCB1 policy (Auer et al., 2002a). As in UCB1, each arm has an index measuring its performance, and at each round, we choose the arm having the highest index. The only difference with UCB1 is to use $\log\left(\frac{n}{Ks}\right)$ instead of $\log(t)$ at time t (see Figure 3). As a consequence, an arm that has been drawn more than n/K times has an index equal to the empirical mean of the rewards obtained from the arm, and when it has been drawn close to n/K times, the logarithmic term is much smaller than the one of UCB1, implying less exploration of this already intensively drawn arm.

MOSS (Minimax Optimal Strategy in the Stochastic case):

For an arm i, define its index $B_{i,s}$ by

$$B_{i,s} = \widehat{\mu}_{i,s} + \sqrt{\frac{\max\left(\log(\frac{n}{Ks}), 0\right)}{s}} \ .$$

for $s \geq 1$ and $B_{i,0} = +\infty$.

At time t, draw an arm maximizing $B_{i,T_i(t-1)}$.

Figure 3: The proposed policy for the stochastic bandit game.

Theorem 24 *Introduce* $\Delta = \min_{i\in\{1,\ldots,K\}:\Delta_i>0} \Delta_i$. *MOSS satisfies*

$$\overline{R}_n \leq \frac{23K}{\Delta} \log\left(\max\left(\frac{110n\Delta^2}{K}, 10^4\right)\right), \tag{22}$$

and

$$\mathbb{E}R_n \leq 25\sqrt{nK}. \tag{23}$$

Besides, if there exists a unique arm with $\Delta_i = 0$, *we also have*

$$\mathbb{E}R_n \leq \frac{23K}{\Delta} \log\left(\max\left(\frac{140n\Delta^2}{K}, 10^4\right)\right). \tag{24}$$

The distribution-dependent bounds Inequalities (22) and (24) show the desired logarithmic dependence in n, while the distribution-free regret bound (23) has the minimax rate $\sqrt{nK}$.

Remark 25 The uniqueness of the optimal arm is really needed to have the logarithmic (in n) bound on the expected regret. This can be easily seen by considering a two-armed bandit in which both reward distributions are identical (and non degenerated). In this case, the pseudo-regret is equal to zero while the expected regret is of order $\sqrt{n}$. This reveals a fundamental difference between the expected regret and the pseudo-regret.

Remark 26 A careful tuning of the constants in front and inside the logarithmic term of $B_{i,s}$ and of the thresholds used in the proof leads to smaller numerical constants in the previous theorem, and in particular to $\sup \mathbb{E}R_n \leq 6\sqrt{nK}$. However, it makes the proof more intricate. So we will only prove (23).

Acknowledgments

Thanks to Gilles Stoltz for pointing us out Proposition 33. This work has been supported by the French National Research Agency (ANR) through the COSINUS program (ANR-08-COSI-004: EXPLO-RA project).

Appendix A. The General Regret Upper Bound of INF

Theorem 27 (INF regret upper bound) *For any nonnegative real numbers $v_{i,t}$, where $i \in \{1, \ldots, K\}$ and $t \in \mathbb{N}^*$, we still use $v_t = (v_{1,t}, \ldots, v_{K,t})$ and $V_t = \sum_{s=1}^{t} v_t$. Define $[V_{t-1}, V_t] = \{\lambda V_{t-1} + (1-\lambda)V_t : \lambda \in [0,1]\}$. Let*

$$B_t = \max_{1 \leq i \leq K} v_{i,t},$$

$$\rho = \max_{1 \leq t \leq n} \max_{v,w \in [V_{t-1}, V_t],\, 1 \leq i \leq K} \frac{\psi'(v_i - C(v))}{\psi'(w_i - C(w))},$$

and

$$A_t = \min\left(B_t^2 \sum_{i=1}^{K} \psi' \circ \psi^{-1}(p_{i,t}), (1+\rho^2) \sum_{i=1}^{K} \psi' \circ \psi^{-1}(p_{i,t}) v_{i,t}^2 \right).$$

Then the INF forecaster based on ψ satisfies:

$$\max_{1 \leq i \leq K} V_{i,n} \leq C_n \leq \sum_{t=1}^{n} \sum_{i=1}^{K} p_{i,t} v_{i,t} - \sum_{i=1}^{K} \left(p_{i,n+1} \psi^{-1}(p_{i,n+1}) + \int_{p_{i,n+1}}^{1/K} \psi^{-1}(u) du \right) + \frac{\rho^2}{2} \sum_{t=1}^{n} A_t. \tag{25}$$

Proof In the following we set $V_0 = 0 \in \mathbb{R}_+^K$ and $C_0 = C(V_0)$. The proof is divided into four steps.

First step: Rewriting $\sum_{t=1}^{n} \sum_{i=1}^{K} p_{i,t} v_{i,t}$.

We start with a simple Abel transformation:

$$\begin{aligned}\sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t}v_{i,t} &= \sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t}(V_{i,t}-V_{i,t-1}) \\ &= \sum_{i=1}^{K} p_{i,n+1}V_{i,n} + \sum_{i=1}^{K}\sum_{t=1}^{n} V_{i,t}(p_{i,t}-p_{i,t+1}) \\ &= \sum_{i=1}^{K} p_{i,n+1}\left(\psi^{-1}(p_{i,n+1})+C_n\right) + \sum_{i=1}^{K}\sum_{t=1}^{n}(\psi^{-1}(p_{i,t+1})+C_t)(p_{i,t}-p_{i,t+1}) \\ &= C_n + \sum_{i=1}^{K} p_{i,n+1}\psi^{-1}(p_{i,n+1}) + \sum_{i=1}^{K}\sum_{t=1}^{n}\psi^{-1}(p_{i,t+1})(p_{i,t}-p_{i,t+1})\end{aligned}$$

where the last step comes from the fact that $\sum_{i=1}^{K} p_{i,t} = 1$.

Second step: A Taylor-Lagrange expansion.

For $x \in [0,1]$ we define $f(x) = \int_0^x \psi^{-1}(u)du$. Remark that $f'(x) = \psi^{-1}(x)$ and $f''(x) = 1/\psi'(\psi^{-1}(x))$. Then by the Taylor-Lagrange formula, we know that for any i, there exists $\tilde{p}_{i,t+1} \in [p_{i,t}, p_{i,t+1}]$ (with the convention $[a,b] = [b,a]$ when $a > b$) such that

$$f(p_{i,t}) = f(p_{i,t+1}) + (p_{i,t}-p_{i,t+1})f'(p_{i,t+1}) + \frac{(p_{i,t}-p_{i,t+1})^2}{2}f''(\tilde{p}_{i,t+1}),$$

or, in other words:

$$(p_{i,t}-p_{i,t+1})\psi^{-1}(p_{i,t+1}) = \int_{p_{i,t+1}}^{p_{i,t}} \psi^{-1}(u)du - \frac{(p_{i,t}-p_{i,t+1})^2}{2\psi'(\psi^{-1}(\tilde{p}_{i,t+1}))}.$$

Now by summing over t the first term on the right-hand side becomes $\int_{p_{i,n+1}}^{1/K} \psi^{-1}(u)du$. Moreover, since $x \to \psi(x - C(x))$ is continuous, there exists $W^{(i,t)} \in [V_t, V_{t+1}] \subset \mathbb{R}^K$ such that $\psi\left(W_i^{(i,t)} - C(W^{(i,t)})\right) = \tilde{p}_{i,t+1}$. Thus we have

$$\sum_{i=1}^{K}\sum_{t=1}^{n}\psi^{-1}(p_{i,t+1})(p_{i,t}-p_{i,t+1}) = \sum_{i=1}^{K}\int_{p_{i,n+1}}^{1/K}\psi^{-1}(u)du - \sum_{i=1}^{K}\sum_{t=1}^{n}\frac{(p_{i,t}-p_{i,t+1})^2}{2\psi'\left(W_i^{(i,t)} - C(W^{(i,t)})\right)}.$$

From the equality obtained in the first step, it gives

$$C_n - \sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t}v_{i,t} = -\sum_{i=1}^{K}\left(p_{i,n+1}\psi^{-1}(p_{i,n+1}) + \int_{p_{i,n+1}}^{1/K}\psi^{-1}(u)du\right) + \sum_{i=1}^{K}\sum_{t=1}^{n}\frac{(p_{i,t}-p_{i,t+1})^2}{2\psi'\left(W_i^{(i,t)} - C(W^{(i,t)})\right)}.$$

Third step: The mean value theorem to compute $(p_{i,t+1} - p_{i,t})^2$.

It is now convenient to consider the functions f_i and h_i defined for any $x \in \mathbb{R}_+^K$ by

$$f_i(x) = \psi(x_i - C(x)) \quad \text{and} \quad h_i(x) = \psi'(x_i - C(x)).$$

We are going to bound $p_{i,t+1} - p_{i,t} = f_i(V_t) - f_i(V_{t-1})$ by using the mean value theorem. To do so we need to compute the gradient of f_i. First, we have

$$\frac{\partial f_i}{\partial x_j}(x) = \left(\mathbb{I}_{i=j} - \frac{\partial C}{\partial x_j}(x) \right) h_i(x).$$

Now, by definition of C, we have $\sum_{k=1}^K f_k(x) = 1$ and thus $\sum_{k=1}^K \frac{\partial f_k}{\partial x_j}(x) = 0$, which implies

$$\frac{\partial C}{\partial x_j}(x) = \frac{h_j(x)}{\sum_{k=1}^K h_k(x)} \qquad \text{and} \qquad \frac{\partial f_i}{\partial x_j}(x) = \left(\mathbb{I}_{i=j} - \frac{h_j(x)}{\sum_{k=1}^K h_k(x)} \right) h_i(x).$$

Now the mean value theorem says that there exists $V^{(i,t)} \in [V_{t-1}, V_t]$ such that

$$f_i(V_t) - f_i(V_{t-1}) = \sum_{j=1}^K v_{j,t} \frac{\partial f_i}{\partial x_j}(V^{(i,t)}).$$

Thus we have

$$\begin{aligned}(p_{i,t} - p_{i,t+1})^2 &= \left(\sum_{j=1}^K v_{j,t} \left(\mathbb{I}_{i=j} - \frac{h_j(V^{(i,t)})}{\sum_{k=1}^K h_k(V^{(i,t)})} \right) h_i(V^{(i,t)}) \right)^2 \\ &= h_i(V^{(i,t)})^2 \left(v_{i,t} - \frac{\sum_{j=1}^K v_{j,t} h_j(V^{(i,t)})}{\sum_{k=1}^K h_k(V^{(i,t)})} \right)^2 .\end{aligned}$$

Fourth step: An almost variance term.

We introduce $\rho = \max_{v,w\in[V_{t-1},V_t],\, 1\le t\le n,\, 1\le i\le K} \frac{h_i(v)}{h_i(w)}$. Thus we have

$$\begin{aligned}\sum_{i=1}^{K}\sum_{t=1}^{n}\frac{(p_{i,t}-p_{i,t+1})^2}{2\psi'\left(W_i^{(i,t)}-C(W^{(i,t)})\right)} &= \sum_{i=1}^{K}\sum_{t=1}^{n}\frac{h_i(V^{(i,t)})^2}{2h_i(W^{(i,t)})}\left(v_{i,t}-\frac{\sum_{j=1}^K v_{j,t}h_j(V^{(i,t)})}{\sum_{k=1}^K h_k(V^{(i,t)})}\right)^2 \\ &\le \frac{\rho^2}{2}\sum_{t=1}^{n}\sum_{i=1}^{K}h_i(V_{t-1})\left(v_{i,t}-\frac{\sum_{j=1}^K v_{j,t}h_j(V^{(i,t)})}{\sum_{k=1}^K h_k(V^{(i,t)})}\right)^2.\end{aligned}$$

Now we need to control the term $\left(v_{i,t}-\frac{\sum_{j=1}^K v_{j,t}h_j(V^{(i,t)})}{\sum_{k=1}^K h_k(V^{(i,t)})}\right)^2$. Remark that since the function ψ is increasing we know that $h_i(x)\ge 0, \forall x$. Now since we have $0\le v_{i,t}\le B_t$, we can simply bound this last term by B_t^2. A different bound can be obtained by using $(a-b)^2\le a^2+b^2$ when a and b have the same sign:

$$\begin{aligned}\left(v_{i,t}-\frac{\sum_{j=1}^K v_{j,t}h_j(V^{(i,t)})}{\sum_{k=1}^K h_k(V^{(i,t)})}\right)^2 &\le v_{i,t}^2+\left(\frac{\sum_{j=1}^K v_{j,t}h_j(V^{(i,t)})}{\sum_{k=1}^K h_k(V^{(i,t)})}\right)^2 \\ &\le v_{i,t}^2+\frac{\sum_{j=1}^K v_{j,t}^2h_j(V^{(i,t)})}{\sum_{k=1}^K h_k(V^{(i,t)})} \\ &\le v_{i,t}^2+\rho^2\frac{\sum_{j=1}^K v_{j,t}^2h_j(V_{t-1})}{\sum_{k=1}^K h_k(V_{t-1})},\end{aligned}$$

where the first inequality comes from the fact that both terms are nonnegative and the second inequality comes from Jensen's inequality. As a consequence, we have

$$\begin{aligned}\sum_{i=1}^{K}h_i(V_{t-1})\left(v_{i,t}-\frac{\sum_{j=1}^K v_{j,t}h_j(V^{(i,t)})}{\sum_{k=1}^K h_k(V^{(i,t)})}\right)^2 &\le \sum_{i=1}^{K}h_i(V_{t-1})v_{i,t}^2+\rho^2\sum_{j=1}^{K}h_j(V_{t-1})v_{j,t}^2 \\ &\le (1+\rho^2)\sum_{i=1}^{K}h_i(V_{t-1})v_{i,t}^2.\end{aligned}$$

We have so far proved

$$C_n-\sum_{t=1}^{n}\sum_{i=1}^{K}p_{i,t}v_{i,t} \le -\sum_{i=1}^{K}\left(p_{i,n+1}\psi^{-1}(p_{i,n+1})+\int_{p_{i,n+1}}^{1/K}\psi^{-1}(u)du\right)+\frac{\rho^2}{2}\sum_{t=1}^{n}A_t.$$

The announced result is then obtained by using Inequality (2). ■

To apply successfully Theorem 27 (page 2620), we need to have tight upper bounds on ρ. The two following lemmas provide such bounds.

Lemma 28 (A simple bound on the quantity ρ of Theorem 27) *Let ψ be a convex function satisfying* (1) *and assume that there exists $B > 0$ such that $\forall i, j, t \; |v_{i,t} - v_{j,t}| \leq B$. Then:*

$$\rho = \max_{1 \leq t \leq n} \max_{v,w \in [V_{t-1}, V_t],\, 1 \leq i \leq K} \frac{\psi'(v_i - C(v))}{\psi'(w_i - C(w))} \leq \sup_{x \in (-\infty, \psi^{-1}(1)]} \exp\left(B \frac{\psi''}{\psi'}(x) \right).$$

Proof Let $h_i(x) = \psi'(x_i - C(x)), m_i(x) = \psi''(x_i - C(x))$. For $\alpha \in [0,1]$ we note

$$\varphi(\alpha) = \log\left\{ h_i(V_{t-1} + \alpha(V_t - V_{t-1})) \right\}.$$

Remark that we should rather note this function $\varphi_{i,t}(\alpha)$ but for sake of simplicity we omit this dependency. With these notations we have $\rho = \max_{\alpha,\beta \in [0,1];\, 1 \leq t \leq n,\, 1 \leq i \leq K} \exp(\varphi(\alpha) - \varphi(\beta))$. By the mean value theorem for any $\alpha, \beta \in [0,1]$ there exists $\xi \in [0,1]$ such that $\varphi(\alpha) - \varphi(\beta) = (\alpha - \beta)\varphi'(\xi)$. Now with the calculus done in the third step of the proof of Theorem 27 and using the notations $h_i := h_i(V_{t-1} + \xi(V_t - V_{t-1}))$, $m_i := m_i(V_{t-1} + \xi(V_t - V_{t-1}))$ we obtain

$$\varphi'(\xi) = \sum_{j=1}^{K} (V_{j,t} - V_{j,t-1}) \left(\mathbb{1}_{i=j} - \frac{h_j}{\sum_{k=1}^K h_k} \right) \frac{m_i}{h_i} = \sum_{j=1}^{K} \frac{(v_{i,t} - v_{j,t}) h_j}{\sum_{k=1}^K h_k} \frac{m_i}{h_i}.$$

Thus we get

$$|\varphi'(\xi)| \leq \max_{1 \leq i,j \leq K} |v_{i,t} - v_{j,t}| \sup_{v \in [V_{t-1}, V_t]} \frac{\psi''}{\psi'}\left(v_i - C(v)\right).$$

Moreover, using that $x \to \psi(x - C(x))$ is continuous we know that there exists $\tilde{p}_{i,t+1} \in [p_{i,t}, p_{i,t+1}]$ such that $\tilde{p}_{i,t+1} = \psi(v_i - C(v))$ and thus $v_i - C(v) = \psi^{-1}(\tilde{p}_{i,t+1})$. This concludes the proof. ∎

Lemma 29 (An other bound on the quantity ρ of Theorem 27) *Let ψ be a function satisfying* (1) *and assume that there exists $c > 0$ such that $0 \leq v_{i,t} \leq \frac{c}{p_{i,t}} \mathbb{1}_{i = I_t}$. We also assume that ψ'/ψ is a nondecreasing function and that there exists $a > 1$ such that $\psi\left(x + \frac{c}{\psi(x)}\right) \leq a\psi(x)$. Then:*

$$\rho \leq \sup_{x \in (-\infty, \psi^{-1}(1)]} \exp\left(ac \frac{\psi''}{\psi \times \psi'}(x) \right).$$

Proof We extract from the previous proof that $\rho \leq \max_{\xi \in [0,1];\, 1 \leq t \leq n,\, 1 \leq i \leq K} \exp(|\varphi'(\xi)|)$ where

$$\varphi'(\xi) = \sum_{j=1}^{K} \frac{(v_{i,t} - v_{j,t}) h_j}{\sum_{k=1}^K h_k} \frac{m_i}{h_i}.$$

Note that, since the functions ψ and ψ'/ψ are nondecreasing, the function ψ is convex, hence $\psi'' \geq 0$ and $m_i \geq 0$. Now using our assumption on $v_{i,t}$ and since $p_{i,t} = f_i(V_{t-1})$, if $i \neq I_t$ we have:

$$|\varphi'(\xi)| \leq \frac{c \frac{h_{I_t}}{f_{I_t}(V_{t-1})}}{\sum_{k=1}^K h_k} \frac{m_i}{h_i} \leq c \frac{f_{I_t}(V_{t-1} + \xi(V_t - V_{t-1}))}{f_{I_t}(V_{t-1})} \times \frac{h_{I_t}}{f_{I_t}(V_{t-1} + \xi(V_t - V_{t-1}))} \times \frac{m_i}{h_i} \times \frac{1}{h_{I_t} + h_i}.$$

Noticing that for any x, y in $\mathbb{R}_-^*$, $\frac{\frac{\psi'}{\psi}(x)\times\frac{\psi''}{\psi'}(y)}{\psi'(x)+\psi'(y)} \leq \frac{\psi''(y)}{\psi'(y)\psi(y)}$, we obtain

$$|\varphi'(\xi)| \leq c\frac{f_{I_t}(V_{t-1}+\xi(V_t-V_{t-1}))}{f_{I_t}(V_{t-1})}\frac{m_i}{h_i \times f_i(V_{t-1}+\xi(V_t-V_{t-1}))}.$$

On the other hand if $i = I_t$ then

$$|\varphi'(\xi)| \leq \frac{c}{f_i(V_{t-1})}\frac{m_i}{h_i}.$$

To finish we only have to prove that $f_{I_t}(V_{t-1}+\xi(V_t-V_{t-1})) \leq a f_{I_t}(V_{t-1})$. Since ψ is increasing it is enough to prove that $f_{I_t}(V_t) \leq a f_{I_t}(V_{t-1})$ which is equivalent to

$$\psi(V_{I_t,t-1}+v_{I_t,t}-C_t) \leq a\psi(V_{I_t,t-1}-C_{t-1}).$$

Since $0 \leq v_{i,t} \leq \frac{c}{p_{i,t}}\mathbb{I}_{i=I_t}$ and C is an increasing function in each of its argument it is enough to prove

$$\psi\left(V_{I_t,t-1}-C_{t-1}+\frac{c}{\psi(V_{I_t,t-1}-C_{t-1})}\right) \leq a\psi(V_{I_t,t-1}-C_{t-1})$$

which is true by hypothesis on ψ. ■

Appendix B. Lower Bounds

In this section we propose a simple unified proof to derive lower bounds on the pseudo-regret in the four problems that we consider.

Theorem 30 *Let $m \geq K$. Let* sup *represents the supremum taken over all oblivious adversaries and* inf *the infimum taken over all forecasters, then the following holds true in the label efficient game:*[4]

$$\inf \sup \overline{R}_n \geq 0.03n\sqrt{\frac{\log(K)}{m}}.$$

and in the label efficient bandit game we have:

$$\inf \sup \overline{R}_n \geq 0.04n\sqrt{\frac{K}{m}}.$$

Proof First step: Definitions.

We consider a set of K oblivious adversaries. The i^{th} adversary selects its gain vectors as follows: For any $t \in \{1,\dots,n\}, g_{i,t} \sim Ber\left(\frac{1+\varepsilon}{2}\right)$ and for $j \neq i$, $g_{j,t} \sim Ber\left(\frac{1-\varepsilon}{2}\right)$. We note $\mathbb{E}_i$ when we integrate with respect to the reward generation process of the i^{th} adversary. We focus on the label efficient versions of the full information and bandits games since by

4. Slightly better numerical constants can be obtained with a more careful optimization in step four of the proof.

taking $m = n$ we recover the traditional games.

Until the fifth step we consider a deterministic forecaster, that is he does not have access to an external randomization. Let $q_n = (q_{1,n}, \ldots, q_{K,n})$ be the empirical distribution of plays over the arms defined by:

$$q_{i,n} = \frac{\sum_{t=1}^n \mathbb{1}_{I_t=i}}{n}.$$

Let J_n be drawn according to q_n. We note $\mathbb{P}_i$ the law of J_n when the forecaster plays against the i^{th} adversary. Remark that we have $\mathbb{P}_i(J_n = j) = \mathbb{E}_i \frac{1}{n} \sum_{t=1}^n \mathbb{1}_{I_t=j}$, hence, against the i^{th} adversary we have:

$$\overline{R}_n = \mathbb{E}_i \sum_{t=1}^n (g_{i,t} - g_{I_t,t}) = \varepsilon n \sum_{j \neq i} \mathbb{P}_i(J_n = j) = \varepsilon n (1 - \mathbb{P}_i(J_n = i)),$$

which implies (since a maximum is larger than a mean)

$$\sup \overline{R}_n \geq \varepsilon n \left(1 - \frac{1}{K} \sum_{i=1}^K \mathbb{P}_i(J_n = i) \right). \tag{26}$$

Second step: Information inequality.

Let $\mathbb{P}_0$ (respectively $\mathbb{P}_{K+1}$) be the law of J_n against the adversary drawing all its losses from the Bernoulli of parameter $\frac{1-\varepsilon}{2}$ (respectively $\frac{1-\varepsilon}{2} + \frac{\varepsilon}{K}$), we call it the 0^{th} adversary (respectively the $(K+1)^{th}$ adversary). Now we use either Pinsker's inequality which gives:

$$\mathbb{P}_i(J_n = i) \leq \mathbb{P}_0(J_n = i) + \sqrt{\frac{1}{2} \mathrm{KL}(\mathbb{P}_0, \mathbb{P}_i)},$$

and thus (thanks to the concavity of the square root)

$$\frac{1}{K} \sum_{i=1}^K \mathbb{P}_i(J_n = i) \leq \frac{1}{K} + \sqrt{\frac{1}{2K} \sum_{i=1}^K \mathrm{KL}(\mathbb{P}_0, \mathbb{P}_i)}; \tag{27}$$

or Fano's lemma:

$$\frac{1}{K} \sum_{i=1}^K \mathbb{P}_i(J_n = i) \leq \frac{\log 2 + \frac{1}{K} \sum_{i=1}^K \mathrm{KL}(\mathbb{P}_i, \mathbb{P}_{K+1})}{\log(K-1)}. \tag{28}$$

We will use (28) for the full information games when $K > 3$ and (27) the bandits games and the full information games with $K \in \{2, 3\}$.

Third step: Computation of $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_i)$ and $\mathrm{KL}(\mathbb{P}_i, \mathbb{P}_{K+1})$ with the Chain rule for Kullback-Leibler divergence.

Remark that since the forecaster is deterministic, the sequence of observed rewards (up to time n) W_n ($W_n \in \{0,1\}^{mK}$ for the full information label efficient game and $W_n \in \{0,1\}^m$ for the label efficient bandit game) uniquely determines the empirical distribution of plays q_n, and in particular the law of J_n conditionally to W_n is the same for any adversary. Thus, if for $i \in \{0,\dots,K+1\}$ we note $\mathbb{P}_i^n$ the law of W_n when the forecaster plays against the i^{th} adversary, then one can easily prove that

$$\mathrm{KL}(\mathbb{P}_0,\mathbb{P}_i) \leq \mathrm{KL}(\mathbb{P}_0^n,\mathbb{P}_i^n), \text{ and } \mathrm{KL}(\mathbb{P}_i,\mathbb{P}_{K+1}) \leq \mathrm{KL}(\mathbb{P}_i^n,\mathbb{P}_{K+1}^n).$$

Now we use the Chain rule for Kullback-Leibler divergence iteratively to introduce the laws $\mathbb{P}_i^t$ of the observed rewards W_t up to time t. We also note $Z_t = 1$ if some rewards are revealed at the end of round t and $Z_t = 0$ otherwise. With these notations we have in the full information games, for $K > 3$,

$$\begin{aligned}
&\mathrm{KL}(\mathbb{P}_i^n,\mathbb{P}_{K+1}^n)\\
&= \mathrm{KL}(\mathbb{P}_i^1,\mathbb{P}_{K+1}^1) + \sum_{t=2}^{n}\sum_{w_{t-1}} \mathbb{P}_i^{t-1}(w_{t-1})\mathrm{KL}(\mathbb{P}_i^t(.|w_{t-1}),\mathbb{P}_{K+1}^t(.|w_{t-1}))\\
&= \mathrm{KL}(\mathbb{P}_i^1,\mathbb{P}_{K+1}^1)\\
&\quad + \sum_{t=2}^{n}\left\{\sum_{w_{t-1}:Z_t=1}\mathbb{P}_i^{t-1}(w_{t-1})\left[\mathrm{KL}\left(\frac{1+\varepsilon}{2},\frac{1-\varepsilon}{2}+\frac{\varepsilon}{K}\right) + (K-1)\mathrm{KL}\left(\frac{1-\varepsilon}{2},\frac{1-\varepsilon}{2}+\frac{\varepsilon}{K}\right)\right]\right\}\\
&= \left[\mathrm{KL}\left(\frac{1+\varepsilon}{2},\frac{1-\varepsilon}{2}+\frac{\varepsilon}{K}\right) + (K-1)\mathrm{KL}\left(\frac{1-\varepsilon}{2},\frac{1-\varepsilon}{2}+\frac{\varepsilon}{K}\right)\right]\mathbb{E}_i\sum_{t=1}^{n} Z_t\\
&\leq m\left[\mathrm{KL}\left(\frac{1+\varepsilon}{2},\frac{1-\varepsilon}{2}+\frac{\varepsilon}{K}\right) + (K-1)\mathrm{KL}\left(\frac{1-\varepsilon}{2},\frac{1-\varepsilon}{2}+\frac{\varepsilon}{K}\right)\right].
\end{aligned}$$

Summing and plugging this into (28) we obtain for the full information games:

$$\frac{1}{K}\sum_{i=1}^{K}\mathbb{P}_i(J_n=i) \leq \frac{\log 2 + m\mathrm{KL}\left(\frac{1+\varepsilon}{2},\frac{1-\varepsilon}{2}+\frac{\varepsilon}{K}\right) + m(K-1)\mathrm{KL}\left(\frac{1-\varepsilon}{2},\frac{1-\varepsilon}{2}+\frac{\varepsilon}{K}\right)}{\log(K-1)}. \quad (29)$$

In the bandits games we have:

$$\begin{aligned}
&\mathrm{KL}(\mathbb{P}_0^n,\mathbb{P}_i^n)\\
&= \mathrm{KL}(\mathbb{P}_0^1,\mathbb{P}_i^1) + \sum_{t=2}^{n}\sum_{w_{t-1}} \mathbb{P}_0^{t-1}(w_{t-1})\mathrm{KL}(\mathbb{P}_0^t(.|w_{t-1}),\mathbb{P}_i^t(.|w_{t-1}))\\
&= \mathrm{KL}(\mathbb{P}_0^1,\mathbb{P}_i^1) + \sum_{t=2}^{n}\sum_{w_{t-1}:Z_t=1,I_t=i}\mathbb{P}_0^{t-1}(w_{t-1})\mathrm{KL}\left(\frac{1-\varepsilon}{2},\frac{1+\varepsilon}{2}\right)\\
&= \mathrm{KL}\left(\frac{1-\varepsilon}{2},\frac{1+\varepsilon}{2}\right)\mathbb{E}_0\sum_{t=1}^{n}\mathbb{1}_{Z_t=1,I_t=i}.
\end{aligned}$$

Summing and plugging this into (27) we obtain for the bandits games:

$$\frac{1}{K}\sum_{i=1}^{K}\mathbb{P}_i(J_n=i) \leq \frac{1}{K} + \sqrt{\frac{m}{2K}\mathrm{KL}\left(\frac{1-\varepsilon}{2},\frac{1+\varepsilon}{2}\right)}. \quad (30)$$

Note that with the same reasoning we obtain for the full information games:

$$\frac{1}{K}\sum_{i=1}^{K}\mathbb{P}_i(J_n=i)\leq\frac{1}{K}+\sqrt{\frac{m}{2}\mathrm{KL}\left(\frac{1-\varepsilon}{2},\frac{1+\varepsilon}{2}\right)}. \tag{31}$$

Fourth step: Conclusion for deterministic forecasters.
To conclude the proof for deterministic forecaster one needs to plug in (29) (for the full information games with $K>3$) or (31) (for the full information games with $K\in\{2,3\}$) or (30) (for the bandits games) in (26) along with straightforward computations and the following simple formula:

$$\mathrm{KL}(p,q)\leq\frac{(p-q)^2}{q(1-q)}.$$

Fifth step: Fubini's Theorem to handle non-deterministic forecasters.

Now let us consider a randomized forecaster. Denote by $\mathbb{E}_{reward,i}$ the expectation with respect to the reward generation process of the i^{th} adversary, $\mathbb{E}_{rand}$ the expectation with respect to the randomization of the forecaster and $\mathbb{E}_i$ the expectation with respect to both processes. Then one has (thanks to Fubini's Theorem),

$$\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_i\sum_{t=1}^{n}(g_{i,t}-g_{I_t,t})=\mathbb{E}_{rand}\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{reward,i}\sum_{t=1}^{n}(g_{i,t}-g_{I_t,t}).$$

Now remark that if we fix the realization of the forecaster's randomization then the results of the previous steps apply and in particular we can lower bound $\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{reward,i}\sum_{t=1}^{n}(g_{i,t}-g_{I_t,t})$ as before. ■

Appendix C. Proofs

This section gathers the proofs that have not been provided so far.

C.1 Proof of Theorem 2 (page 2605)

The proof relies on combining Theorem 27 (page 2620) with Lemma 28 (page 2624) for $\gamma=0$, and with Lemma 29 (page 2624) for $\gamma>0$.

We make use of Theorem 27 and start with straightforward computations to bound the first sum in (25). We have $\psi^{-1}(x)=-\eta(x-\gamma/K)^{-1/q}$ which admits as a primitive $\int\psi^{-1}(u)du=\frac{-\eta}{1-1/q}(u-\gamma/K)^{1-1/q}$. Thus one immediately gets

$$\int_{p_{i,n+1}}^{1/K}(-\psi^{-1})(u)du\leq\frac{\eta}{1-1/q}\frac{1}{K^{1-1/q}}-\eta(p_{i,n+1}-\gamma/K)^{1-1/q}$$

and

$$p_{i,n+1}(-\psi^{-1})(p_{i,n+1})=-\frac{\gamma}{K}\psi^{-1}(p_{i,n+1})+\eta(p_{i,n+1}-\gamma/K)^{1-1/q}.$$

Summing over i proves that

$$-\sum_{i=1}^{K}\left(p_{i,n+1}\psi^{-1}(p_{i,n+1})+\int_{p_{i,n+1}}^{1/K}\psi^{-1}(u)du\right)\leq\frac{q}{q-1}\eta K^{1/q}-\frac{\gamma}{K}\sum_{i=1}^{K}\psi^{-1}(p_{i,n+1}).$$

With the notations of Theorem 27, we need now to bound ρ and A_t. First we deal with the case $\gamma=0$. Lemma 28 (page 2624) implies $\rho\leq\exp(B(q+1)/\eta)$ since we have $\frac{\psi''}{\psi'}(x)=\frac{q+1}{-x}=\frac{q+1}{\eta}\psi(x)^{1/q}$. The proof of (6) is concluded by $\psi'=\frac{q}{\eta}\psi^{(q+1)/q}$, and

$$A_t\leq B_t^2\sum_{i=1}^{K}\psi'\circ\psi^{-1}(p_{i,t})=B_t^2\sum_{i=1}^{K}\frac{q}{\eta}p_{i,t}^{(q+1)/q}\leq\frac{q}{\eta}B_t^2.$$

For (7), the term A_t is controlled differently:

$$A_t\leq(1+\rho^2)\sum_{i=1}^{K}\psi'\circ\psi^{-1}(p_{i,t})v_{i,t}^2=\frac{q}{\eta}(1+\rho^2)\sum_{i=1}^{K}p_{i,t}^{(q+1)/q}v_{i,t}^2\leq\frac{qB}{\eta}(1+\rho^2)\sum_{i=1}^{K}p_{i,t}v_{i,t}.$$

Now we have already seen that $\rho\leq\exp(B(q+1)/\eta)$, hence $\rho^2(1+\rho^2)\leq 2\exp(8Bq/\eta)$, which leads to (7).

The case $\gamma>0$ is more intricate. This is why we restrict ourselves to a specific form for the estimates $v_{i,t}$, see the assumption in Theorem 2. We start by using Lemma 29 (page 2624) to prove that $\rho\leq\mu$. First we have $\frac{\psi''}{\psi'}=\frac{q+1}{\eta}(\psi-\gamma/K)^{1/q}\leq\frac{q+1}{\eta}\psi^{1/q}$. Besides, for any $a\geq b\geq d$ we have $\frac{a}{b}\leq\frac{a-d}{b-d}$ and thus for any $x<0$, we have

$$\frac{\psi\left(x+\frac{c}{\psi(x)}\right)}{\psi(x)}\leq\frac{\psi\left(x+\frac{c}{\psi(x)}\right)-\frac{\gamma}{K}}{\psi(x)-\frac{\gamma}{K}}=\left(1-\frac{c}{-x\psi(x)}\right)^{-q}\leq\left(1-\frac{c}{q\eta}\left(\frac{(q-1)K}{\gamma}\right)^{(q-1)/q}\right)^{-q}.$$

Thus Lemma 29 gives us

$$\rho^2\leq\exp\left\{\frac{2(q+1)c}{\eta}\left(\frac{K}{\gamma}\right)^{(q-1)/q}\left(1-\frac{c}{q\eta}\left(\frac{(q-1)K}{\gamma}\right)^{(q-1)/q}\right)^{-q}\right\}=\mu.$$

Next we use $\psi'=\frac{q}{\eta}(\psi-\gamma/K)^{(q+1)/q}$ and the form of $v_{i,t}$ to get

$$A_t\leq(1+\rho^2)\sum_{i=1}^{K}\psi'\circ\psi^{-1}(p_{i,t})v_{i,t}^2\leq\frac{q(1+\mu)}{\eta}\sum_{i=1}^{K}p_{i,t}^{(q+1)/q}v_{i,t}^2=\frac{q(1+\mu)}{\eta}p_{I_t,t}^{(1-q)/q}c_t^2.$$

Let $\zeta' = \frac{qc\mu(1+\mu)}{2\eta}$. From Theorem 27, we get

$$\begin{aligned}
C_n - \sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t}v_{i,t} &\leq \frac{q}{q-1}\eta K^{1/q} + \gamma C_n + \frac{\rho^2}{2}\sum_{t=1}^{n} A_t - \frac{\gamma}{K}\sum_{t=1}^{n}\sum_{i=1}^{K} v_{i,t} \\
&\leq \frac{q}{q-1}\eta K^{1/q} + \gamma C_n + \sum_{t=1}^{n} c_t\left(\zeta' p_{I_t,t}^{(1-q)/q} - \frac{\gamma}{K p_{I_t,t}}\right) \\
&\leq \frac{q}{q-1}\eta K^{1/q} + \gamma C_n + \max_{u>0}\left(\zeta' u^{(1-q)/q} - \frac{\gamma}{Ku}\right)\sum_{t=1}^{n} c_t \\
&= \frac{q}{q-1}\eta K^{1/q} + \gamma C_n + \frac{\gamma}{(q-1)K}\left(\frac{(q-1)\zeta' K}{q\gamma}\right)^q \sum_{t=1}^{n} c_t \\
&= \frac{q}{q-1}\eta K^{1/q} + \gamma C_n + \gamma\zeta\sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t}v_{i,t}.
\end{aligned}$$

The proof of (8) is concluded by using Inequality (2).

C.2 Proof of Theorem 3 (page 2606)

We have

$$\begin{aligned}
\sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t}v_{i,t} &= \sum_{t=1}^{n} \mathbb{E}_{k\sim p_t} v_{k,t} \\
&= \frac{1-\gamma}{\eta}\sum_{t=1}^{n}\left(\log \mathbb{E}_{i\sim q_t}\exp(\eta v_{i,t}) - \log\left[\exp\left(-\frac{\eta}{1-\gamma}\mathbb{E}_{k\sim p_t} v_{k,t}\right)\mathbb{E}_{i\sim q_t}\exp(\eta v_{i,t})\right]\right) \\
&= \frac{1-\gamma}{\eta}\left(S - \sum_{t=1}^{n}\log(D_t)\right),
\end{aligned}$$

where

$$\begin{aligned}
S &= \sum_{t=1}^{n}\log \mathbb{E}_{i\sim q_t}\exp(\eta v_{i,t}) \\
&= \sum_{t=1}^{n}\log\left(\frac{\sum_{i=1}^{K}\exp(\eta V_{i,t})}{\sum_{i=1}^{K}\exp(\eta V_{i,t-1})}\right) = \log\left(\frac{\sum_{i=1}^{K}\exp(\eta V_{i,n})}{K}\right) \geq \eta \max_{1\leq i\leq K} V_{i,n} - \log K
\end{aligned}$$

and

$$D_t = \exp\left(-\frac{\eta}{1-\gamma}\mathbb{E}_{k\sim p_t} v_{k,t}\right)\mathbb{E}_{i\sim q_t}\exp(\eta v_{i,t})$$

When $\gamma = 0$, since $0 \leq v_{i,t} \leq B_t$, by applying Hoeffding's inequality, we get $\log(D_t) \leq \frac{\eta^2 B_t^2}{8}$, hence Inequality (9). For $\gamma = 0$, we can also use Lemma 35 and obtain $\log(D_t) \leq$

$\eta^2 B\Theta(\eta B)\mathbb{E}_{i\sim p_t} v_{i,t}$, hence Inequality (10). For γ satisfying (11), we have

$$\begin{aligned} D_t &\le \exp\left(-\frac{\eta}{1-\gamma}\mathbb{E}_{k\sim p_t} v_{k,t}\right)\mathbb{E}_{i\sim q_t}\left(1+\eta v_{i,t}+\Theta(\eta B)\eta^2 v_{i,t}^2\right) \qquad (32)\\ &= \exp\left(-\frac{\eta}{1-\gamma}\mathbb{E}_{k\sim p_t} v_{k,t}\right)\left(1+\frac{\eta}{1-\gamma}\mathbb{E}_{i\sim p_t} v_{i,t}-\eta\frac{\gamma\sum_{i=1}^K v_{i,t}}{K(1-\gamma)}+\Theta(\eta B)\eta^2\mathbb{E}_{i\sim q_t} v_{i,t}^2\right)\\ &\le \exp\left(-\frac{\eta}{1-\gamma}\mathbb{E}_{k\sim p_t} v_{k,t}\right)\left(1+\frac{\eta}{1-\gamma}\mathbb{E}_{i\sim p_t} v_{i,t}\right) \qquad (33)\\ &\le 1. \end{aligned}$$

To get (32), we used that Θ is an increasing function and that $\eta v_{i,t}\le \eta B$. To get (33), we noticed that it is trivial when $\max_{i,t} p_{i,t}v_{i,t}=0$, and that otherwise, we have

$$\frac{\gamma\sum_{i=1}^K v_{i,t}}{K(1-\gamma)} \ge \frac{\gamma\sum_{i=1}^K p_{i,t}v_{i,t}^2}{K(1-\gamma)\max_{i,t} p_{i,t}v_{i,t}} \ge \frac{\gamma}{K\max_{i,t} p_{i,t}v_{i,t}}\mathbb{E}_{i\sim q_t} v_{i,t}^2 \ge \eta\Theta(\eta B)\mathbb{E}_{i\sim q_t} v_{i,t}^2,$$

where the last inequality uses (11). We have thus proved

$$\sum_{t=1}^n\sum_{i=1}^K p_{i,t}v_{i,t} \ge \frac{1-\gamma}{\eta}\log\mathbb{E}_{i\sim p_1}\exp(\eta V_{i,n}) \ge (1-\gamma)\left(\max_{1\le i\le K} V_{i,n}-\frac{\log K}{\eta}\right),$$

hence the announced result.

C.3 Recovering Theorem 3 from Theorem 27

We start with straightforward computations to bound the first sum in (25). We have $\psi^{-1}(x)=\frac{1}{\eta}\log(x-\gamma/K)$ which admits as a primitive $\int\psi^{-1}(u)du=\frac{1}{\eta}\left[(u-\gamma/K)\log(u-\gamma/K)-u\right]$. Thus one immediately gets

$$\begin{aligned} &-\int_{p_{i,n+1}}^{1/K}\psi^{-1}(u)du - p_{i,n+1}\psi^{-1}(p_{i,n+1})\\ &\qquad = \frac{1}{\eta}\left(\frac{1}{K}-\frac{1-\gamma}{K}\log\left(\frac{1-\gamma}{K}\right)-p_{i,n+1}-\frac{\gamma}{K}\log\left(p_{i,n+1}-\frac{\gamma}{K}\right)\right). \end{aligned}$$

Summing over i proves that

$$-\sum_{i=1}^K\left(p_{i,n+1}\psi^{-1}(p_{i,n+1})+\int_{p_{i,n+1}}^{1/K}\psi^{-1}(u)du\right) = \frac{1-\gamma}{\eta}\log\left(\frac{K}{1-\gamma}\right)-\frac{\gamma}{K}\sum_{i=1}^K\psi^{-1}(p_{i,n+1}).$$

With the notations of Theorem 27, we need now to bound ρ and A_t. For the former, we use Lemma 28 (page 2624) which directly shows $\rho\le\exp(\eta B)$. For the latter we distinguish two cases. If $\gamma=0$ we use

$$A_t \le B_t^2\sum_{i=1}^K\psi'\circ\psi^{-1}(p_{i,t}) = \eta B_t^2,$$

which concludes the proof of the weakened version of (9) with $\frac{\eta}{8}$ replaced by $\frac{\eta}{2}\exp(2\eta B)$. On the other hand if $\gamma > 0$ we use

$$A_t \leq (1+\rho^2)\sum_{i=1}^{K} \psi' \circ \psi^{-1}(p_{i,t}) v_{i,t}^2 \leq (1+\rho^2)\eta \sum_{i=1}^{K} p_{i,t} v_{i,t}^2.$$

From Theorem 27, when the weakened version of (11) holds, that is when

$$\gamma \geq K \frac{\eta \exp(2B\eta)\left[1+\exp(2B\eta)\right]}{2} \max_{i,t} p_{i,t} v_{i,t},$$

we have

$$\begin{aligned}
&C_n - \sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t} v_{i,t} \\
&\leq \frac{1-\gamma}{\eta}\log\left(\frac{K}{1-\gamma}\right) - \frac{\gamma}{K}\sum_{i=1}^{K}\left(\sum_{t=1}^{n} v_{i,t} - C_n\right) + \frac{\eta\exp(2B\eta)\left[1+\exp(2B\eta)\right]}{2}\sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t} v_{i,t}^2 \\
&\leq \frac{1-\gamma}{\eta}\log\left(\frac{K}{1-\gamma}\right) + \gamma C_n,
\end{aligned}$$

hence

$$(1-\gamma)\left(C_n + \frac{\log(1-\gamma)}{\eta}\right) - \sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t} v_{i,t} \leq (1-\gamma)\frac{\log K}{\eta}.$$

This gives the desired result since we have

$$C_n + \frac{\log(1-\gamma)}{\eta} = \frac{1}{\eta}\log\left(\sum_{j=1}^{K}\exp(\eta V_{i,n})\right) \geq \max_{1\leq i\leq K} V_{i,n}.$$

C.4 Proof of Theorem 8 (page 2610)

We will use the following version of Bernstein's inequality for martingales.

Theorem 31 *Let $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n$ be a filtration, and $X_1, \ldots, X_n$ random variables such that $|X_t| \leq b$ for some $b > 0$, X_t is $\mathcal{F}_t$-measurable, $\mathbb{E}(X_t|\mathcal{F}_{t-1}) = 0$ and $\mathbb{E}(X_t^2|\mathcal{F}_{t-1}) \leq v$ for some $v > 0$. Then, for any $t > 0$, we have*

$$\mathbb{P}\Big(\sum_{t=1}^{n} X_t \geq t\Big) \leq \exp\left(-\frac{t^2}{2nv + 2bt/3}\right), \tag{34}$$

and for any $\delta > 0$, with probability at least $1-\delta$, we have

$$\sum_{t=1}^{n} X_t \leq \sqrt{2nv\log(\delta^{-1})} + \frac{b\log(\delta^{-1})}{3}.$$

Proof of Theorem 31 Both inequalities come from Result (1.6) of Freedman (1975). The first inequality then uses $(1+x)\log(1+x)-x \geq \frac{x^2}{2+2x/3}$, while the other uses Inequality (45) of Audibert et al. (2009). This last inequality allows to remove the $\sqrt{2}$ factor appearing in Lemma A.8 of Cesa-Bianchi and Lugosi (2006). ■

We start the proof of Theorem 8 by noting that, since $R_n \leq n$, the result is trivial for $\delta \leq 2K\exp(-m/27)$ so that we assume hereafter that $\delta \geq 2K\exp(-m/27)$, or equivalently $\frac{\log(2K\delta^{-1})}{m} \leq \frac{1}{27}$. We consider the event $\mathcal{E}$ on which we simultaneously have

$$\sum_{t=1}^{n} Z_t \leq m, \tag{35}$$

$$-\sum_{t=1}^{n} g_{I_t,t} \leq -\sum_{t=1}^{n}\sum_{k=1}^{K} p_{k,t}g_{k,t} + \sqrt{\frac{n\log(4\delta^{-1})}{2}}, \tag{36}$$

and

$$\max_{1\leq i\leq K}\sum_{t=1}^{n}\left(g_{i,t}-\frac{\eta}{8\varepsilon}-\sum_{k=1}^{K}p_{k,t}g_{k,t}\right)\left(1-\frac{Z_t}{\varepsilon}\right) \leq 2\sqrt{\frac{n\log(2K\delta^{-1})}{\varepsilon}}+\frac{\log(2K\delta^{-1})}{2\varepsilon}. \tag{37}$$

Let us first prove that this event holds with probability at least $1-\delta$. From (34), we have

$$\mathbb{P}\left(\sum_{t=1}^{n} Z_t > m\right) \leq \exp\left(-\frac{m^2/16}{3m/2+m/6}\right) \leq \exp\left(-\frac{m}{27}\right) \leq \frac{\delta}{4}$$

So (35) holds with probability at least $1-\delta/4$. From the concentration of martingales with bounded differences (Hoeffding, 1963; Azuma, 1967), (36) holds with probability at least $1-\delta/4$. For $\eta/(8\varepsilon) \leq \sqrt{2}-1$ (which is true for our particular η), we can apply Theorem 31 with $b=\sqrt{2}/\varepsilon$ and $v=2/\varepsilon$ to the random variables $\left(g_{i,t}-\frac{\eta}{8\varepsilon}-\sum_{k=1}^{K}p_{k,t}g_{k,t}\right)\left(1-\frac{Z_t}{\varepsilon}\right)$. We get that for a fixed $i\in\{1,\dots,K\}$, with probability at least $1-\delta/(2K)$, we have

$$\sum_{t=1}^{n}\left(g_{i,t}-\frac{\eta}{8\varepsilon}-\sum_{k=1}^{K}p_{k,t}g_{k,t}\right)\left(1-\frac{Z_t}{\varepsilon}\right) \leq 2\sqrt{\frac{n\log(2K\delta^{-1})}{\varepsilon}}+\frac{\log(2K\delta^{-1})}{2\varepsilon}.$$

From a union bound, we get that (37) holds with probability at least $1-\delta/2$. Using again a union bound, we thus have proved that the event $\mathcal{E}$ holds with probability at least $1-\delta$.

Now, on the event $\mathcal{E}$, by combining (36) and (37), we obtain

$$\begin{aligned}
R_n = \max_{1\leq i\leq K}\sum_{t=1}^{n} g_{i,t} - \sum_{t=1}^{n} g_{I_t,t} &\leq \sqrt{\frac{n\log(4\delta^{-1})}{2}} + \max_{1\leq i\leq K}\sum_{t=1}^{n} g_{i,t} - \sum_{t=1}^{n}\sum_{k=1}^{K} p_{k,t}g_{k,t} \\
&\leq \sqrt{\frac{n\log(4\delta^{-1})}{2}} + \max_{1\leq i\leq K}\sum_{t=1}^{n} g_{i,t}\frac{Z_t}{\varepsilon} - \sum_{t=1}^{n}\sum_{k=1}^{K} p_{k,t}g_{k,t}\frac{Z_t}{\varepsilon} \\
&\quad + \sum_{t=1}^{n}\frac{\eta}{8\varepsilon}\left(1-\frac{Z_t}{\varepsilon}\right) + 2\sqrt{\frac{n\log(2K\delta^{-1})}{\varepsilon}} + \frac{\log(2K\delta^{-1})}{2\varepsilon}
\end{aligned}$$

Since we have $\sum_{t=1}^{n} Z_t \leq m$, the rewards received by the forecaster are equal to the rewards which would receive the forecaster that uses Z_t to decide whether he asks for the gains or not, whatever $\sum_{s=1}^{t-1} Z_s$ is. This enables us to use (9) (which holds with probability one). We obtain

$$\begin{aligned} R_n &\leq \sqrt{\frac{n\log(4\delta^{-1})}{2}} + \frac{\log K}{\eta} + \frac{n\eta}{8\varepsilon} + 2\sqrt{\frac{n\log(2K\delta^{-1})}{\varepsilon}} + \frac{\log(2K\delta^{-1})}{2\varepsilon} \\ &\leq \sqrt{\frac{n\log(4\delta^{-1})}{2}} + \frac{\log K}{\eta} + \frac{n^2\eta}{6m} + 4n\sqrt{\frac{\log(2K\delta^{-1})}{3m}} + \frac{2n\log(2K\delta^{-1})}{3m}. \end{aligned}$$

From the inequalities $m \leq n$, $K \geq 2$ and $\frac{\log(2K\delta^{-1})}{m} \leq \frac{1}{27}$, this implies

$$R_n \leq \frac{10n}{3}\sqrt{\frac{\log(2K\delta^{-1})}{m}} + \frac{\log K}{\eta} + \frac{\eta n^2}{6m}.$$

The first inequality of the theorem is then obtained by plugging $\eta = \frac{\sqrt{m\log K}}{n}$. The second inequality is derived by integrating the deviations using the standard formula $\mathbb{E}W \leq \int_0^1 \frac{1}{\delta}\mathbb{P}(W > \log(\delta^{-1}))d\delta$.

C.5 Proof of Theorem 9 (page 2610)

The proof goes exactly like for Theorem 8. We just use (6) instead of (9).

C.6 Proof of Theorem 12 (page 2612)

The bound is trivial for $9K\log(3K) \geq n$. So we consider hereafter that $\gamma = 2\sqrt{\frac{K\log(3K)}{n}} < \frac{2}{3}$. The result is then a direct consequence of Theorem 21 with $G_0 = n$. The inequality on $\mathbb{E}R_n$ comes by integrating the deviations.

C.7 Proof of Theorem 13 (page 2613)

First note that (16) holds for $9\sqrt{nK} \geq n$ since we trivially have $R_n \leq n$. For $9\sqrt{nK} < n$, we apply (18) with $G_0 = n > 81K$, and obtain $R_n \leq 8.5\sqrt{nK} + \sqrt{2nK}\log(\delta^{-1})$. This implies (16) and also (17) by using Proposition 33.

For the last assertions, we proceed similarly. They trivially hold for $9\sqrt{nK\log(3K)} \geq n$. For $n > 9\sqrt{nK\log(3K)}$, we apply Theorem 20 with $G_0 = n$, and obtain

$$R_n \leq \frac{9}{2}\sqrt{nK\log(3K)} + 4\sqrt{\frac{nK}{\log(3K)}} + \sqrt{\frac{2nK}{\log(3K)}}\log(K\delta^{-1}).$$

By using $\frac{1}{\sqrt{\log(3K)}} \leq \frac{1}{\log(6)}\sqrt{\log(3K)}$, this independently implies

$$R_n \leq 9\sqrt{nK\log(3K)} + \sqrt{\frac{2nK}{\log(3K)}}\log(\delta^{-1}),$$

and by integration,

$$\mathbb{E}R_n \leq 9\sqrt{nK\log(3K)},$$

hence the desired inequalities.

C.8 Proof of Theorem 17 (page 2615)

The proof follows the scheme described in Section 5.1.2. In particular let

$$\nu = \frac{\gamma\varepsilon}{\beta K} + \frac{1}{\log\left(1 - \frac{\beta K}{\gamma\varepsilon}\right)}.$$

Then we have

$$\begin{aligned} -\sum_{i=1}^{K} p_{i,t} v_{i,t} &= Z_t \frac{p_{I_t,t}}{\beta} \log\left(1 - \frac{\beta g_{I_t,t}}{\varepsilon p_{I_t,t}}\right) \\ &\geq Z_t \left(-\frac{g_{I_t,t}}{\varepsilon} + \frac{\nu g_{I_t,t}}{\varepsilon} \log\left(1 - \frac{\beta g_{I_t,t}}{\varepsilon p_{I_t,t}}\right)\right) \\ &\geq -g_{I_t,t}\frac{Z_t}{\varepsilon} - \frac{\nu\beta}{\varepsilon}\sum_{i=1}^{K} v_{i,t}. \end{aligned}$$

Moreover with a simple application of Theorem 31 we have that with probability at least $1-\delta$,

$$-\sum_{t=1}^{n} g_{I_t,t} \leq -\sum_{t=1}^{n} g_{I_t,t}\frac{Z_t}{\varepsilon} + \sqrt{\frac{2n\log\delta^{-1}}{\varepsilon}} + \frac{\log\delta^{-1}}{3\varepsilon}.$$

Now note that the sequence $W_t = \exp\left(\beta G_{i,t} - \beta V_{i,t}\right)$, $t = 1, \ldots, n$, is a supermartingale over the filtration generated by (g_t, I_t, Z_t), $t = 1, \ldots, n$. Indeed, we have for any $t \in \{1, \ldots, n\}$,

$$\mathbb{E}_{I_t \sim p_t, Z_t} \exp(-\beta v_{i,t}) = 1 - \varepsilon + \varepsilon\left(1 - \frac{\beta g_{i,t}}{\varepsilon}\right) = 1 - \beta g_{i,t} \leq \exp(-\beta g_{i,t}).$$

Thus, with probability at least $1-\delta$, we have against deterministic adversaries

$$\max_{1\leq i\leq K} V_{i,n} \geq G_{\max} - \frac{\log\delta^{-1}}{\beta},$$

and against general adversaries

$$\max_{1\leq i\leq K} V_{i,n} \geq G_{\max} - \frac{\log(K\delta^{-1})}{\beta}.$$

Now we apply (8) of Theorem 2. Let $c = -\frac{\gamma}{\beta K}\log\left(1 - \frac{\beta K}{\gamma\varepsilon}\right)$. If $c < q\eta\left(\frac{\gamma}{(q-1)K}\right)^{(q-1)/q}$ then we have

$$(1-\gamma)\left(\max_{1\leq i\leq K} V_{i,n}\right) - (1+\gamma\zeta)\sum_{t=1}^{n}\sum_{i=1}^{K} p_{i,t} v_{i,t} \leq \frac{q}{q-1}\eta K^{\frac{1}{q}},$$

where

$$\zeta = \frac{1}{(q-1)K}\left(\frac{(q-1)cK\mu(1+\mu)}{2\gamma\eta}\right)^{q},$$

with

$$\mu = \exp\left\{\frac{2(q+1)c}{\eta}\left(\frac{K}{\gamma}\right)^{(q-1)/q}\left(1-\frac{c}{q\eta}\left(\frac{(q-1)K}{\gamma}\right)^{(q-1)/q}\right)^{-q}\right\}.$$

Thus, in the case of a deterministic adversaries, we obtain

$$\begin{aligned}
&G_{\max} - \sum_{t=1}^{n} g_{I_t,t} \\
&\leq \big(\gamma(1+\zeta) + \frac{\nu\beta K}{\varepsilon}\big)n + \frac{q}{q-1}\eta K^{\frac{1}{q}} + \frac{\log(2\delta^{-1})}{\beta} + n\sqrt{\frac{2\log(2\delta^{-1})}{m}} + n\frac{\log(2\delta^{-1})}{3m} \\
&= \frac{n}{m}\left(\big(\gamma(1+\zeta) + \nu\beta' K\big)m + \frac{q}{q-1}\eta' K^{\frac{1}{q}} + \frac{\log(2\delta^{-1})}{\beta'}\right) + n\sqrt{\frac{2\log(2\delta^{-1})}{m}} + n\frac{\log(2\delta^{-1})}{3m},
\end{aligned} \tag{38}$$

where $\beta' = \beta/\varepsilon$ and $\eta' = \eta\varepsilon$. One can see that the term into parenthesis in (38) is exactly the same than the right hand side of (43), up to the relabelling of β and η into β' and η'. This allows us to use the same numerical application as in Section C.9 (up to the additional terms outside of the parenthesis in (43)). One can apply the same technique in the case of a general adversary.

C.9 Proof of Theorem 18 (page 2616), Corollary 19 and Theorem 20 (page 2616)

Consider parameters $q > 1$, $0 < \gamma < 1$, $\eta > 0$ and $\beta > 0$ such that $\beta K < \gamma$ and such that the real number $c = -\frac{\gamma}{\beta K}\log\big(1 - \frac{\beta K}{\gamma}\big)$ satisfies $c < q\eta\big(\frac{\gamma}{(q-1)K}\big)^{(q-1)/q}$. From (8) of Theorem 2, since we have $v_{i,t} = \frac{c_t}{p_{i,t}}\mathbb{1}_{I_t=i}$ with $c_t = -\frac{p_{i,t}}{\beta}\log\big(1 - \frac{\beta g_{i,t}}{p_{i,t}}\big) \leq -\frac{\gamma}{\beta K}\log\big(1 - \frac{\beta K}{\gamma}\big) = c$, we have

$$(1-\gamma)\left(\max_{1\leq i\leq K} V_{i,n}\right) + \frac{1+\gamma\zeta}{\beta}\sum_{t=1}^{n} p_{I_t,t}\log\left(1 - \frac{\beta g_{I_t,t}}{p_{I_t,t}}\right) \leq \frac{q}{q-1}\eta K^{\frac{1}{q}}, \tag{39}$$

where

$$\zeta = \frac{1}{(q-1)K}\Big(\frac{(q-1)cK\mu(1+\mu)}{2\gamma\eta}\Big)^{q},$$

with

$$\mu = \exp\left\{\frac{2(q+1)c}{\eta}\left(\frac{K}{\gamma}\right)^{(q-1)/q}\left(1-\frac{c}{q\eta}\left(\frac{(q-1)K}{\gamma}\right)^{(q-1)/q}\right)^{-q}\right\}.$$

Let

$$\nu = \frac{\gamma}{\beta K} + \frac{1}{\log(1-\beta K/\gamma)}.$$

The function $x \mapsto \frac{1}{x} + \frac{1}{\log(1-x)}$ is increasing on $(0, +\infty)$. So we have

$$\frac{1}{\log(1 - \beta g_{I_t,t}/p_{I_t,t})} + \frac{p_{I_t,t}}{\beta g_{I_t,t}} \leq \nu,$$

hence

$$\frac{p_{I_t,t}}{\beta}\log\left(1-\frac{\beta g_{I_t,t}}{p_{I_t,t}}\right) \geq -g_{I_t,t} + \nu g_{I_t,t}\log\left(1-\frac{\beta g_{I_t,t}}{p_{I_t,t}}\right) \geq -g_{I_t,t} - \nu\beta\sum_{i=1}^{K} v_{i,t}. \tag{40}$$

Inequality (39) thus implies

$$(1-\gamma)\left(\max_{1\leq i\leq K} V_{i,n}\right) - (1+\gamma\zeta)\sum_{t=1}^{n} g_{I_t,t} - (1+\gamma\zeta)\nu\beta\sum_{i=1}^{K} V_{i,n} \leq \frac{q}{q-1}\eta K^{\frac{1}{q}},$$

which leads to

$$\big(1-\gamma-(1+\gamma\zeta)\nu\beta K\big)\left(\max_{1\leq i\leq K} V_{i,n}\right) - (1+\gamma\zeta)\sum_{t=1}^{n} g_{I_t,t} \leq \frac{q}{q-1}\eta K^{\frac{1}{q}}, \tag{41}$$

We now provide a high probability lower bound of the left-hand side. The technical tool (essentially deviation inequalities for supermartingales) comes from Section 6.8 of Cesa-Bianchi and Lugosi (2006).

High probability lower bound on $\max_{1\leq i\leq K} V_{i,n}$**.**

For any $t \in \{1,\dots,n\}$, we have

$$\mathbb{E}_{I_t\sim p_t}\exp(-\beta v_{i,t}) = \mathbb{E}_{I_t\sim p_t}\exp\left\{\log\left(1-\frac{\beta g_{i,t}}{p_{i,t}}\right)\mathbb{I}_{I_t=i}\right\} = 1-\beta g_{i,t} \leq \exp(-\beta g_{i,t}).$$

This implies that the sequence $W_t = \exp\big(\beta G_{i,t} - \beta V_{i,t}\big)$, $t=1,\dots,n$, forms a supermartingale over the filtration generated by (g_t, I_t), $t=1,\dots,n$. Indeed, we have

$$\mathbb{E}\big(\exp(W_t)|(g_s,I_s), s=1,\dots,t-1\big) = \mathbb{E}_{g_t|(g_s,I_s),s=1,\dots,t-1}\mathbb{E}_{I_t\sim p_t}\exp(W_t) \leq \exp(W_{t-1}).$$

So, we have $\mathbb{E}\exp(W_n) \leq \mathbb{E}\exp(W_1) \leq 1$, which implies that with probability at least $1-\delta$, $V_{i,n} \geq G_{i,n} - \frac{\log(\delta^{-1})}{\beta}$ with probability at least $1-\delta$. So, for any fixed $k \in \{1,\dots,K\}$, we have

$$\max_{1\leq i\leq K} V_{i,n} \geq G_{k,n} - \frac{\log(\delta^{-1})}{\beta}. \tag{42}$$

Combining (41) and (42), we obtain that for any $\delta > 0$ and any fixed $k \in \{1,\dots,K\}$, with probability at least $1-\delta$, we have

$$\big(1-\gamma-(1+\gamma\zeta)\nu\beta K\big)G_{k,n} - (1+\gamma\zeta)\sum_{t=1}^{n} g_{I_t,t} \leq \frac{q}{q-1}\eta K^{\frac{1}{q}} + \frac{\log(\delta^{-1})}{\beta},$$

hence

$$G_{k,n} - \sum_{t=1}^{n} g_{I_t,t} \leq \big(\gamma(1+\zeta)+\nu\beta K\big)G_{k,n} + \frac{q}{q-1}\eta K^{\frac{1}{q}} + \frac{\log(\delta^{-1})}{\beta}. \tag{43}$$

Now, for $G_0 \geq 81K$, let us take

$$q=2, \quad \gamma = 3\sqrt{\frac{K}{G_0}}, \quad \beta = \frac{1}{\sqrt{2KG_0}}, \quad \text{and} \quad \eta = 2\sqrt{G_0}.$$

Then we have $c \approx 1.14$, $\nu \approx 0.522$, $\mu \leq 2.09$, $\zeta \leq 0.377$, and

$$G_{k,n} - \sum_{t=1}^{n} g_{I_t,t} \leq 4.5\sqrt{K\frac{G_{\max}^2}{G_0}} + 4\sqrt{KG_0} + \sqrt{2KG_0}\log(\delta^{-1}). \tag{44}$$

For deterministic adversaries, the arm achieving a cumulative reward equal to $G_{\max}$ is deterministic (it does not depend on the randomization of the forecaster). So, we may take k equal to this fixed arm, and obtain (18).

To prove the inequality for a fully oblivious adversary, let us take $k \in \operatorname{argmax}_{i\in\{1,\ldots,K\}} \mathbb{E}G_{i,n}$. From (44) which holds with probability at least $1-\delta$, it suffices to prove that with probability at least $1-\delta$, we have $G_{k,n} \geq G_{\max} - \sqrt{8\log(K\delta^{-1})G_{\max}}$. Let $\lambda > 0$. Since the reward vectors $g_1, \ldots, g_n$ are independent, and from the inequality $\exp(\lambda x) \leq 1 + [\exp(\lambda) - 1]x$ for any $x \in [0,1]$ (by convexity of the exponential function), for any $j \neq k$, from Lemma 35, we have

$$\mathbb{E}\exp(\lambda G_{j,n}) = \prod_{t=1}^{n} \mathbb{E}\exp(\lambda g_{j,t}) \leq \prod_{t=1}^{n} \exp\big[\big(\exp(\lambda) - 1\big)\mathbb{E}g_{j,t}\big] = \exp\big[\big(\exp(\lambda) - 1\big)\mathbb{E}G_{j,n}\big],$$

and

$$\begin{aligned}
\mathbb{E}\exp(-\lambda G_{k,n}) &= \prod_{t=1}^{n} \mathbb{E}\exp(-\lambda g_{k,t}) \\
&\leq \prod_{t=1}^{n} \mathbb{E}\left(1 - \lambda g_{k,t} + \frac{1}{2}\lambda^2 g_{k,t}^2\right) \\
&\leq \prod_{t=1}^{n} \left(1 - \lambda\Big(1 - \frac{\lambda}{2}\Big)\mathbb{E}g_{k,t}\right) \\
&\leq \prod_{t=1}^{n} \exp\left[-\lambda\Big(1 - \frac{\lambda}{2}\Big)\mathbb{E}g_{k,t}\right] = \exp\left[-\lambda\Big(1 - \frac{\lambda}{2}\Big)\mathbb{E}G_{k,n}\right].
\end{aligned}$$

This implies respectively that for any $j \neq k$, with probability at least $1-\delta$,

$$G_{j,n} \leq \mathbb{E}G_{j,n} + \lambda\Theta(\lambda)\mathbb{E}G_{j,n} + \frac{\log(\delta^{-1})}{\lambda},$$

and with probability at least $1-\delta$,

$$\mathbb{E}G_{k,n} \leq G_{k,n} + \frac{\lambda}{2}\mathbb{E}G_{k,n} + \frac{\log(\delta^{-1})}{\lambda}.$$

By optimizing the free parameter λ (using Lemma 32 below), and from a union bound, with probability at least $1-\delta$, we simultaneously have

$$G_{j,n} \leq \mathbb{E}G_{j,n} + \sqrt{2\mathbb{E}G_{j,n}\log(K\delta^{-1})} + \frac{\log(K\delta^{-1})}{3},$$

and

$$\mathbb{E}G_{k,n} \leq G_{k,n} + \sqrt{2\mathbb{E}G_{k,n}\log(K\delta^{-1})}.$$

Since we have $\mathbb{E}G_{k,n} \geq \mathbb{E}G_{j,n}$, we get consecutively $G_{k,n} \geq \mathbb{E}G_{j,n} - \sqrt{2\mathbb{E}G_{j,n}\log(K\delta^{-1})}$, and after computations, $G_{k,n} \geq G_{j,n} - \sqrt{8G_{j,n}\log(K\delta^{-1})}$ for any $j \neq k$. With probability at least $1-\delta$, we thus have $G_{k,n} \geq G_{\max} - \sqrt{8\log(K\delta^{-1})G_{\max}}$, which concludes the proof of (19).

To prove Corollary 19, first note that (20) holds for $9\sqrt{KG_{\max}} \geq G_{\max}$ since we trivially have $R_n \leq G_{\max}$. For $9\sqrt{KG_{\max}} < G_{\max}$, we may apply Theorem 18 with $G_0 = G_{\max}$ since $G_{\max} > 81K$, and obtain $R_n \leq 8.5\sqrt{KG_{\max}} + \sqrt{2KG_{\max}}\log(\delta^{-1})$. This implies (20) and (21).

Lemma 32 *Let* $\Theta(\lambda) = \frac{\exp(\lambda)-1-\lambda}{\lambda^2}$. *For any* $A > 0$, $\inf_{\lambda>0}\left\{\lambda\Theta(\lambda) + \frac{A^2}{2\lambda}\right\} \leq A + A^2/6$.

Proof Considering $\lambda = \log(1+A)$, we have $\inf_{\lambda>0}\left\{\lambda\Theta(\lambda) + \frac{A^2}{2\lambda}\right\} \leq \log(1+A)\Theta[\log(1+A)] + \frac{A^2}{2\log(1+A)} = A + \frac{A^2}{6} - \frac{1+A+\frac{A^2}{6}}{\log(1+A)}\Phi(A)$ where $\Phi(A) \triangleq \log(1+A) - \frac{A+\frac{A^2}{2}}{1+A+\frac{A^2}{6}}$. Since $\Phi(0) = 0$ and $\Phi'(A) = \frac{A^4}{36(1+A)(1+A+A^2/6)^2} \geq 0$, we get $\Phi(A) \geq 0$, hence the result. ■

To prove Theorem 20, we replace (42), which holds with probability at least $1-\delta$, by

$$\max_{1\leq i\leq K} V_{i,n} \geq G_{\max} - \frac{\log(\delta^{-1})}{\beta},$$

which, by a union bound, holds with probability at least $1-K\delta$. (It is this union bound that makes the $\log K$ factor appears in the bound.) This leads to the following modified version of (43): with probability $1-K\delta$,

$$G_{\max} - \sum_{t=1}^{n} g_{I_t,t} \leq \big(\gamma(1+\zeta) + \nu\beta K\big)G_{\max} + \frac{q}{q-1}\eta K^{\frac{1}{q}} + \frac{\log(\delta^{-1})}{\beta}.$$

Now, for $G_0 \geq 81K\log(3K)$, let us take $q = 2$,

$$\gamma = 3\sqrt{\frac{K\log(3K)}{G_0}}, \quad \beta = \sqrt{\frac{\log(3K)}{2KG_0}}, \quad \text{and} \quad \eta = 2\sqrt{\frac{G_0}{\log(3K)}}.$$

Then we have $c \approx 1.14$, $\nu \approx 0.522$, $\mu \leq 2.09$, $\zeta \leq 0.377$, and

$$G_{\max} - \sum_{t=1}^{n} g_{I_t,t} \leq 4.5\sqrt{\frac{G_{\max}^2}{G_0}K\log(3K)} + 4\sqrt{\frac{KG_0}{\log(3K)}} + \sqrt{\frac{2KG_0}{\log(3K)}}\log(\delta^{-1}).$$

This inequality holds with probability at least $1-K\delta$. This implies the result of Theorem 20.

C.10 Proof of Theorem 21 (page 2617)

Consider parameters $q > 1$, $0 < \gamma < 1$, $\eta > 0$ and $\beta > 0$ such that $\beta K < \gamma$. Introduce $c = -\frac{\gamma}{\beta K}\log\left(1 - \frac{\beta K}{\gamma}\right)$. We have $\max_{i,t} p_{i,t}v_{i,t} \leq c$. So (11) holds as soon as

$$\gamma \geq K\eta c\Theta\bigg(-\frac{\eta}{\beta}\log\Big(1 - \frac{\beta K}{\gamma}\Big)\bigg). \tag{45}$$

From (12) and as in the proof of Theorem 13, by using (40) with $\nu = \frac{\gamma}{\beta K} + \frac{1}{\log(1-\beta K/\gamma)}$, we obtain

$$(1-\gamma)\Big(\max_{1\leq i\leq K} V_{i,n}\Big) - \sum_{t=1}^{n} g_{I_t,t} - \nu\beta \sum_{i=1}^{K} V_{i,n} \leq (1-\gamma)\frac{\log K}{\eta},$$

hence

$$(1-\gamma-\nu\beta K)\Big(\max_{1\leq i\leq K} V_{i,t}\Big) - \sum_{t=1}^{n} g_{I_t,t} \leq (1-\gamma)\frac{\log K}{\eta}.$$

From the same argument as in the proof of Theorem 18, for any $i \in \{1,\dots,K\}$, we have $\mathbb{E}\exp\big(\beta G_{i,n} - \beta V_{i,n}\big) \leq 1$, and for any $\delta > 0$, $V_{i,n} \geq G_{i,n} - \frac{\log(\delta^{-1})}{\beta}$ holds with probability at least $1-\delta$. By a union bound, we get that with probability at least $1 - K\delta$,

$$\max_{1\leq i\leq K} V_{i,n} \geq G_{\max} - \frac{\log(\delta^{-1})}{\beta}.$$

With probability at least $1-\delta$, we thus have

$$G_{\max} - \sum_{t=1}^{n} g_{I_t,t} \leq (\gamma + \nu\beta K) G_{\max} + \frac{\log(K\delta^{-1})}{\beta} + \frac{\log K}{\eta}.$$

This inequality holds for any parameters $q > 1$, $0 < \gamma < 1$, $\eta > 0$ and $\beta > 0$ such that the inequality $\beta K < \gamma$ and (45) hold. We choose

$$\gamma = 2\sqrt{\frac{K\log(3K)}{G_0}}, \quad \beta = \sqrt{\frac{\log(3K)}{2KG_0}}, \quad \text{and} \quad \eta = 2\sqrt{\frac{\log(3K)}{KG_0}},$$

which gives $c \approx 1.23$, $\nu \approx 0.536$, and

$$G_{\max} - \sum_{t=1}^{n} g_{I_t,t} \leq \frac{5}{2}\sqrt{\frac{G_{\max}^2}{G_0} K\log(3K)} + \sqrt{\frac{2KG_0}{\log(3K)}}\log(K\delta^{-1}) + \frac{1}{2}\sqrt{KG_0\log(3K)}.$$

C.11 Proof of Theorem 22 (page 2617)

Consider parameters $0 < \gamma < 1$, $\eta > 0$ and $\beta > 0$. We have $\max_{i,t} p_{i,t} v_{i,t} \leq 1 + \frac{\beta K}{\gamma}$ and $\max_{i,t} v_{i,t} \leq (1+\beta)\frac{K}{\gamma}$. So (11) holds as soon as

$$\gamma \geq K\eta\Big(1 + \frac{\beta K}{\gamma}\Big)\Theta\Big(\frac{\eta(1+\beta)K}{\gamma}\Big). \tag{46}$$

Then, from (12), we have

$$(1-\gamma)\Big(\max_{1\leq i\leq K} V_{i,n}\Big) - \sum_{t=1}^{n} g_{I_t,t} - \beta n K \leq (1-\gamma)\frac{\log K}{\eta}.$$

Let $\xi_t = \max_{1\leq i\leq K} V_{i,t} - \min_{1\leq j\leq K} V_{j,t}$ and $\xi = \max_{1\leq t\leq n} \xi_t$. Consider a fixed switching strategy $(i_1,\dots,i_n) \in \{1,\dots,K\}^n$, and let $V_{(i_1,\dots,i_n)} = \sum_{t=1}^{n} v_{i_t,t}$. One can easily check that $\max_{1\leq i\leq K} V_{i,n} \geq V_{(i_1,\dots,i_n)} - \xi\mathcal{S}(i_1,\dots,i_n)$, and consequently

$$\max_{1\leq i\leq K} V_{i,n} \geq \max_{(i_1,\dots,i_n):\, \mathcal{S}(i_1,\dots,i_n)\leq S} V_{(i_1,\dots,i_n)} - \xi S.$$

Since $\exp(-x) \leq 1-x+x^2/2$ for $x \leq 0$, we have for any $t \in \{1,\dots,n\}$ and any $i \in \{1,\dots,K\}$

$$\begin{aligned}
\mathbb{E}_{I_t \sim p_t} \exp\left(-2\beta g_{i,t} \frac{\mathbb{I}_{I_t=i}}{p_{i,t}} \right) &\leq \mathbb{E}_{I_t \sim p_t}\left(1 - 2\beta g_{i,t} \frac{\mathbb{I}_{I_t=i}}{p_{i,t}} + 2\beta^2 g_{i,t}^2 \frac{\mathbb{I}_{I_t=i}}{p_{i,t}^2} \right) \\
&= 1 - 2\beta g_{i,t} + 2\beta^2 \frac{g_{i,t}^2}{p_{i,t}} \\
&\leq 1 - 2\beta\Big(g_{i,t} - \frac{\beta}{p_{i,t}} \Big) \\
&\leq \exp\left(-2\beta\Big(g_{i,t} - \frac{\beta}{p_{i,t}} \Big) \right),
\end{aligned}$$

hence

$$\mathbb{E}_{I_t \sim p_t} \exp\big(2\beta(g_{i,t} - v_{i,t}) \big) \leq 1.$$

For a fixed $(i_1,\dots,i_n)$, by using this inequality n times corresponding to the n time steps and their associated actions, this implies $\mathbb{E} \exp\big(2\beta(G_{(i_1,\dots,i_n)} - V_{(i_1,\dots,i_n)})\big) \leq 1$, hence with probability at least $1-\delta$,

$$G_{(i_1,\dots,i_n)} - V_{(i_1,\dots,i_n)} \leq \frac{\log(\delta^{-1})}{2\beta}.$$

Let $M = \sum_{j=0}^{S} \binom{n-1}{j} K(K-1)^j$ be the number of switching strategies of size not larger than S. By a union bound, we get that with probability at least $1-\delta$,

$$\max_{(i_1,\dots,i_n):\, \mathcal{S}(i_1,\dots,i_n) \leq S} V_{(i_1,\dots,i_n)} \geq \max_{(i_1,\dots,i_n):\, \mathcal{S}(i_1,\dots,i_n) \leq S} G_{(i_1,\dots,i_n)} - \frac{\log(M\delta^{-1})}{2\beta}.$$

By putting the previous inequalities together, we obtain that with probability at least $1-\delta$,

$$(1-\gamma) \max_{(i_1,\dots,i_n):\, \mathcal{S}(i_1,\dots,i_n) \leq S} G_{(i_1,\dots,i_n)} - \sum_{t=1}^{n} g_{I_t,t} \leq \beta n K + (1-\gamma)\frac{\log K}{\eta} + \xi S + \frac{\log(M\delta^{-1})}{2\beta},$$

hence

$$R_n^S = \max_{(i_1,\dots,i_n):\, \mathcal{S}(i_1,\dots,i_n) \leq S} G_{(i_1,\dots,i_n)} - \sum_{t=1}^{n} g_{I_t,t} \leq (\gamma + \beta K)n + \frac{\log K}{\eta} + \xi S + \frac{\log(M\delta^{-1})}{2\beta}.$$

We now upper bound M and ξ. We have

$$M = \sum_{j=0}^{S} \binom{n-1}{j} K(K-1)^j \leq K^{S+1} \sum_{j=0}^{S} \binom{n-1}{j} \leq K^{S+1}\left(\frac{en}{S}\right)^S = \frac{\exp(s)}{2},$$

where the second inequality comes from Sauer's lemma. Let

$$\tilde{\rho} = \exp\Big((1+\beta)\frac{K\eta}{\gamma} \Big) \frac{1+K\beta}{1-\gamma}.$$

By contradiction, we now prove

$$\xi \leq \tilde{\rho} - \frac{1}{\eta} \log\left(\frac{\beta}{\tilde{\rho}} - \frac{\gamma}{K}\right). \tag{47}$$

To this end, we start by bounding $C_t - C_{t-1}$. By the mean value theorem, with the notations of the third step of the proof of Theorem 27, there exists $W \in [V_{t-1}, V_t]$ such that

$$\begin{aligned}
C_t - C_{t-1} &= C(V_t) - C(V_{t-1}) \\
&= \sum_{i=1}^{K} \frac{\partial C}{\partial x_i}(W)(V_{i,t} - V_{i,t-1}) \\
&= \sum_{i=1}^{K} \frac{h_i(W)}{\sum_{j=1}^{K} h_j(W)} \frac{g_{i,t}\mathbb{1}_{I_t=i} + \beta}{f_i(V_{i,t-1})} \\
&= \frac{1}{\sum_{j=1}^{K} \eta(f_j(W) - \gamma/K)} \sum_{i=1}^{K} \eta h_i(W) \frac{g_{i,t}\mathbb{1}_{I_t=i} + \beta}{h_i(V_{i,t-1}) + \eta\gamma/K} \\
&\leq \frac{1}{1-\gamma} \sum_{i=1}^{K} h_i(W) \frac{\mathbb{1}_{I_t=i} + \beta}{h_i(V_{t-1})} \leq \frac{\rho}{1-\gamma} \sum_{i=1}^{K} (\mathbb{1}_{I_t=i} + \beta) = \rho \frac{1 + K\beta}{1-\gamma}.
\end{aligned}$$

From Lemma 28 (page 2624), we have $\rho \leq \exp\left((1+\beta)\frac{K\eta}{\gamma}\right)$, hence $C_t - C_{t-1} \leq \exp\Big((1+\beta)\frac{K\eta}{\gamma}\Big)\frac{1+K\beta}{1-\gamma} = \tilde{\rho}$. If (47) does not hold, then from Lemma 1, we have

$$\max_{1 \leq t \leq n} \left(C_t - \min_{1 \leq j \leq K} V_{j,t}\right) > \tilde{\rho} - \psi^{-1}(\beta/\tilde{\rho}).$$

Besides we have $C_0 - \min_{1 \leq j \leq K} V_{j,0} = -\psi^{-1}(1/K) \leq \tilde{\rho} - \psi^{-1}(\beta/\tilde{\rho})$, since $K\beta \leq \tilde{\rho}$. So there exist $T \in \{1, \ldots, n\}$ and $\ell \in \{1, \ldots, K\}$ such that $C_{T-1} - V_{\ell,T-1} \leq \tilde{\rho} - \psi^{-1}(\beta/\tilde{\rho})$ and $C_T - V_{\ell,T} > \tilde{\rho} - \psi^{-1}(\beta/\tilde{\rho})$. In particular, we have $\psi(V_{\ell,T} - C_T + \tilde{\rho}) < \frac{\beta}{\tilde{\rho}}$, hence

$$V_{\ell,T} - V_{\ell,T-1} \geq \frac{\beta}{p_{\ell,T}} = \frac{\beta}{\psi(V_{\ell,T-1} - C_{T-1})} \geq \frac{\beta}{\psi(V_{\ell,T} - C_T + \tilde{\rho})} \geq \tilde{\rho} \geq C_T - C_{T-1},$$

which contradicts the inequality $C_{T-1} - V_{\ell,T-1} < C_T - V_{\ell,T}$. This ends the proof of (47). We have thus proved that for any $0 < \gamma < 1$, $\eta > 0$ and $\beta > 0$ such that (46) holds, we have

$$R_n^S \leq (\gamma + \beta K)n + \frac{\log K}{\eta} + S\left\{\tilde{\rho} - \frac{1}{\eta}\log\left(\frac{\beta}{\tilde{\rho}} - \frac{\gamma}{K}\right)\right\} + \frac{\log(K\delta^{-1})}{2\beta} + \frac{S\log(Ken/S)}{2\beta},$$

with $\tilde{\rho} = \frac{1+K\beta}{1-\gamma}\exp\left((1+\beta)\frac{K\eta}{\gamma}\right)$. For the numerical application, we first notice that the bound trivially holds for $7\sqrt{Ks} \geq \sqrt{n}$. For $7\sqrt{Ks} < \sqrt{n}$, with $s = S\log\left(\frac{enK}{S}\right) + 2\log K$, we choose

$$\gamma = \sqrt{\frac{Ks}{2n}}, \quad \beta = 3\sqrt{\frac{s}{nK}}, \quad \text{and} \quad \eta = \frac{1}{5}\sqrt{\frac{s}{nK}}.$$

We then use $\gamma \leq \frac{1}{7\sqrt{2}}$, $\beta K \leq \frac{3}{7}$, $\beta \leq \frac{3}{14}$ to deduce $\tilde{\rho} \leq 2.25$, and $\tilde{\rho} S \leq 0.05\sqrt{nKs}$. We check (46) by the upper bound $\frac{K\eta}{\gamma}\big(1+\frac{\beta K}{\gamma}\big)\Theta\big(\frac{\eta(1+\beta)K}{\gamma}\big) \leq 0.84 < 1$. We also use $-\log\big(\frac{\beta}{\tilde{\rho}} - \frac{\gamma}{K}\big) \leq \frac{1}{2}\log\big(3nK/s\big) \leq \frac{1}{2}\log\big(3nK/S\big)$. We thus have

$$R_n^S \leq \left(3 + \frac{1}{\sqrt{2}} + 0.05 + \frac{5}{2} + \frac{1}{6}\right)\sqrt{nKs} + \frac{\log(\delta^{-1})}{2\beta} \leq 6.5\sqrt{nKs} + \frac{\log(\delta^{-1})}{6}\sqrt{\frac{nK}{s}}$$

The last inequality follows by integrating the deviations.

C.12 Proof of Theorem 24 (page 2619)

This proof requires some new arguments compared to the one for UCB1. First, we need to decouple the arm, while not being too loose. This is achieved by introducing appropriate stopping times. The decoupled upper bound on the pseudo-regret is (51). Secondly, we use peeling arguments to tightly control the terms in the right-hand side of (51).

We may assume $\mu_1 \geq \ldots \geq \mu_K$. Using the trivial equality $\sum_{i=1}^K \mathbb{E}T_i(n) = n$, we have

$$\begin{aligned}
\overline{R}_n &= \max_{1\leq i\leq K} \mathbb{E}\sum_{t=1}^n \big(g_{i,t} - g_{I_t,t}\big) \\
&= n\Big(\max_{1\leq i\leq K} \mathbb{E}g_{i,t}\Big) - \sum_{t=1}^n \mathbb{E}g_{I_t,t} \\
&= n\Big(\max_{1\leq i\leq K} \mu_i\Big) - \sum_{t=1}^n \mathbb{E}\mu_{I_t} \\
&= n\Big(\max_{1\leq i\leq K} \mu_i\Big) - \mathbb{E}\sum_{t=1}^n \mu_{I_t} \\
&= \left(\sum_{i=1}^K \mathbb{E}T_i(n)\right)\Big(\max_{1\leq i\leq K} \mu_i\Big) - \mathbb{E}\sum_{i=1}^K \mu_i T_i(n) = \sum_{i=1}^K \Delta_i \mathbb{E}T_i(n).
\end{aligned}$$

C.12.1 First Step: Decoupling the Arms

For an arm k_0, we trivially have $\sum_{k=1}^K \Delta_k T_k(n) \leq n\Delta_{k_0} + \sum_{k=k_0+1}^K \Delta_k T_k(n)$. Let $\Delta_{K+1} = +\infty$, $z_k = \mu_1 - \frac{\Delta_k}{2}$ for $k_0 < k \leq K+1$ and $z_{k_0} = +\infty$. Let $Z = \min_{1\leq s\leq n} B_{1,s}$ and $W_{j,k} = \mathbb{1}_{Z\in[z_{j+1},z_j)}(\Delta_k - \Delta_{k_0})T_k(n)$. By using $\mathbb{E}\sum_{k=1}^{k_0} T_k(n) = n - \mathbb{E}\sum_{k=k_0+1}^K T_k(n)$, we get

$$\overline{R}_n = \mathbb{E}\sum_{k=1}^K \Delta_k T_k(n) \leq n\Delta_{k_0} + \mathbb{E}\sum_{k=k_0+1}^K (\Delta_k - \Delta_{k_0})T_k(n).$$

We have

$$\sum_{k=k_0+1}^K (\Delta_k - \Delta_{k_0})T_k(n) = \sum_{k=k_0+1}^K \sum_{j=k_0}^K W_{j,k} = \sum_{j=k_0}^K \sum_{k=k_0+1}^j W_{j,k} + \sum_{j=k_0}^K \sum_{k=j+1}^K W_{j,k}. \qquad (48)$$

An Abel transformation takes care of the first sum of (48):

$$\sum_{j=k_0}^{K}\sum_{k=k_0+1}^{j} W_{j,k} \leq \sum_{j=k_0}^{K} \mathbb{1}_{Z\in[z_{j+1},z_j)} n(\Delta_j - \Delta_{k_0}) = n\sum_{j=k_0+1}^{K} \mathbb{1}_{Z<z_j}(\Delta_j - \Delta_{j-1}). \tag{49}$$

To bound the second sum of (48), we introduce the stopping times $\tau_k = \min\{t : B_{k,t} < z_k\}$ and remark that, by definition of MOSS, we have $\{Z \geq z_k\} \subset \{T_k(n) \leq \tau_k\}$, since once we have pulled τ_k times arm k its index will always be lower than the index of arm 1. This implies

$$\sum_{j=k_0}^{K}\sum_{k=j+1}^{K} W_{j,k} = \sum_{k=k_0+1}^{K}\sum_{j=k_0}^{k-1} W_{j,k} = \sum_{k=k_0+1}^{K} \mathbb{1}_{Z\geq z_k}\Delta_k T_k(n) \leq \sum_{k=k_0+1}^{K} \tau_k\Delta_k. \tag{50}$$

Combining (48), (49) and (50) and taking the expectation, we get

$$\overline{R}_n \leq n\Delta_{k_0} + \sum_{k=k_0+1}^{K} \Delta_k \mathbb{E}\tau_k + n\sum_{k=k_0+1}^{K} \mathbb{P}(Z < z_k)(\Delta_k - \Delta_{k-1}). \tag{51}$$

Let $\delta_0 = \sqrt{\frac{75K}{n}}$ and set k_0 such that $\Delta_{k_0} \leq \delta_0 < \Delta_{k_0+1}$. If $k_0 = K$, we trivially have $\overline{R}_n \leq n\delta_0 \leq \sqrt{75nK}$ so that (22) holds trivially. In the following, we thus consider $k_0 < K$.

Second step: Bounding $\mathbb{E}\tau_k$ for $k_0 + 1 \leq k \leq K$.

Let $\log_+(x) = \max(\log(x), 0)$. For $\ell_0 \in \mathbb{N}$, we have

$$\begin{aligned}
\mathbb{E}\tau_k - \ell_0 &= \sum_{\ell=0}^{+\infty} \mathbb{P}(\tau_k > \ell) - \ell_0 \\
&\leq \sum_{\ell=\ell_0}^{+\infty} \mathbb{P}(\tau_k > \ell) = \sum_{\ell=\ell_0}^{+\infty} \mathbb{P}(\forall t \leq \ell, B_{k,t} > z_k) \\
&\leq \sum_{\ell=\ell_0}^{+\infty} \mathbb{P}\left(\widehat{\mu}_{k,\ell} - \mu_k \geq \frac{\Delta_k}{2} - \sqrt{\frac{\log_+(n/K\ell)}{\ell}}\right).
\end{aligned} \tag{52}$$

Now let us take $\ell_0 = \lceil 7\log\left(\frac{n}{K}\Delta_k^2\right)/\Delta_k^2 \rceil$ with $\lceil x \rceil$ the smallest integer larger than x. For $\ell \geq \ell_0$, since $k > k_0$, we have

$$\log_+\left(\frac{n}{K\ell}\right) \leq \log_+\left(\frac{n}{K\ell_0}\right) \leq \log_+\left(\frac{n\Delta_k^2}{7K}\right) \leq \frac{\ell_0\Delta_k^2}{7} \leq \frac{\ell\Delta_k^2}{7},$$

hence $\frac{\Delta_k}{2} - \sqrt{\frac{\log_+(n/(K\ell))}{\ell}} \geq c\Delta_k$, with $c = \frac{1}{2} - \frac{1}{\sqrt{7}}$. Therefore, by using Hoeffding's inequality and (52), we get

$$\begin{aligned}
\mathbb{E}\tau_k - \ell_0 &\leq \sum_{\ell=\ell_0}^{+\infty} \mathbb{P}\left(\widehat{\mu}_{k,\ell} - \mu_k \geq c\Delta_k\right) \\
&\leq \sum_{\ell=\ell_0}^{+\infty} \exp\left(-2\ell(c\Delta_k)^2\right) = \frac{\exp\left(-2\ell_0(c\Delta_k)^2\right)}{1 - \exp\left(-2(c\Delta_k)^2\right)} \leq \frac{\exp(-14c^2\log(75))}{1 - \exp\left(-2c^2\Delta_k^2\right)},
\end{aligned}$$

where the last inequality uses $\ell_0 \Delta_k^2 \geq 7\log(75)$. Plugging the value of ℓ_0, we obtain

$$\begin{aligned}\Delta_k \mathbb{E}\tau_k &\leq \Delta_k \left(1 + \frac{7\log\left(\frac{n}{K}\Delta_k^2\right)}{\Delta_k^2}\right) + \frac{\Delta_k \exp(-14c^2\log(75))}{1-\exp\left(-2c^2\Delta_k^2\right)} \\ &\leq 1 + 7\frac{\log\left(\frac{n}{K}\Delta_k^2\right)}{\Delta_k} + \frac{\exp(-14c^2\log(75))}{2c^2(1-c^2)\Delta_k},\end{aligned}$$

where the last step uses that, since $1-\exp(-x) \geq x - x^2/2$ for any $x \geq 0$, we have

$$\frac{1}{1-\exp\left(-2c^2\Delta_k^2\right)} \leq \frac{1}{2c^2\Delta_k^2 - 2c^4\Delta_k^4} \leq \frac{1}{2c^2\Delta_k^2(1-c^2)}$$

Third step: Bounding $n\sum_{k=k_0+1}^{K} \mathbb{P}(Z < z_k)(\Delta_k - \Delta_{k-1})$.

Let X_t denote the reward obtained by arm 1 when it is drawn for the t-th time. The random variables $X_1, X_2, \ldots$ are i.i.d. so that we have the maximal inequality (Hoeffding, 1963, Inequality (2.17)): for any $x > 0$ and $m \geq 1$,

$$\mathbb{P}\left(\exists s \in \{1,\ldots,m\}, \sum_{t=1}^{s}(\mu_1 - X_t) > x\right) \leq \exp\left(-\frac{2x^2}{m}\right).$$

Since $z_k = \mu_1 - \Delta_k/2$ and since $u \mapsto \mathbb{P}\left(Z < \mu_1 - u/2\right)$ is a nonincreasing function, we have

$$\sum_{k=k_0+1}^{K} \mathbb{P}(Z < z_k)(\Delta_k - \Delta_{k-1}) \leq \Delta_{k_0+1}\mathbb{P}(Z < z_{k_0+1}) + \int_{\Delta_{k_0+1}}^{1} \mathbb{P}\left(Z < \mu_1 - \frac{u}{2}\right) du. \quad (53)$$

We will now concentrate on upper bounding $\mathbb{P}\left(Z < \mu_1 - \frac{u}{2}\right)$ for a fixed $u \in [\delta_0, 1]$. Let $f(u) = 8\log\left(\sqrt{\frac{n}{K}}u\right)/u^2$. We have

$$\begin{aligned}\mathbb{P}\left(Z < \mu_1 - \frac{1}{2}u\right) &= \mathbb{P}\left(\exists 1 \leq s \leq n : \sum_{t=1}^{s}(\mu_1 - X_t) > \sqrt{s\log_+\left(\frac{n}{Ks}\right)} + \frac{su}{2}\right) \\ &\leq \mathbb{P}\left(\exists 1 \leq s \leq f(u) : \sum_{t=1}^{s}(\mu_1 - X_t) > \sqrt{s\log_+\left(\frac{n}{Ks}\right)}\right) \\ &\quad + \mathbb{P}\left(\exists f(u) < s \leq n : \sum_{t=1}^{s}(\mu_1 - X_t) > \frac{su}{2}\right).\end{aligned}$$

For the first term, we use a peeling argument with a geometric grid of the form $\frac{1}{2^{\ell+1}}f(u) \leq s \leq \frac{1}{2^\ell}f(u)$. The numerical constant in δ_0 ensures that $f(u) \leq n/K$, which implies that for

any $s \leq f(u)$, $\log_+\left(\frac{n}{Ks}\right) = \log\left(\frac{n}{Ks}\right)$. We have

$$\begin{aligned}
&\mathbb{P}\left(\exists 1 \leq s \leq f(u) : \sum_{t=1}^{s}(\mu_1 - X_t) > \sqrt{s\log\left(\frac{n}{Ks}\right)}\right) \\
&\leq \sum_{\ell=0}^{+\infty}\mathbb{P}\left(\exists \frac{1}{2^{\ell+1}}f(u) \leq s \leq \frac{1}{2^\ell}f(u) : \sum_{t=1}^{s}(\mu_1 - X_t) > \sqrt{\frac{f(u)}{2^{\ell+1}}\log\left(\frac{n2^\ell}{Kf(u)}\right)}\right) \\
&\leq \sum_{\ell=0}^{+\infty}\exp\left(-2\frac{f(u)\frac{1}{2^{\ell+1}}\log\left(\frac{n2^\ell}{Kf(u)}\right)}{f(u)\frac{1}{2^\ell}}\right) = \sum_{\ell=0}^{+\infty}\frac{Kf(u)}{n}\frac{1}{2^\ell} = \frac{16K}{nu^2}\log\left(\sqrt{\frac{n}{K}}u\right).
\end{aligned}$$

For the second term we also use a peeling argument but with a geometric grid of the form $2^\ell f(u) \leq s \leq 2^{\ell+1}f(u)$:

$$\begin{aligned}
&\mathbb{P}\Big(\exists s \in \lceil f(u)\rceil, \ldots, n\} : \sum_{t=1}^{s}(\mu_1 - X_t) > \frac{su}{2}\Big) \\
&\qquad \leq \sum_{\ell=0}^{+\infty}\mathbb{P}\left(\exists 2^\ell f(u) \leq s \leq 2^{\ell+1}f(u) : \sum_{t=1}^{s}(\mu_1 - X_t) > 2^{\ell-1}f(u)u\right) \\
&\qquad \leq \sum_{\ell=0}^{+\infty}\exp\left(-2\frac{\left(2^{\ell-1}f(u)u\right)^2}{f(u)2^{\ell+1}}\right) \\
&\qquad = \sum_{\ell=0}^{+\infty}\exp\left(-2^\ell f(u)u^2/4\right) \\
&\qquad \leq \sum_{\ell=0}^{+\infty}\exp\left(-(\ell+1)f(u)u^2/4\right) = \frac{1}{\exp\left(f(u)u^2/4\right) - 1} = \frac{1}{nu^2/K - 1}.
\end{aligned}$$

Putting together the last three computations, we obtain

$$\mathbb{P}\left(Z < \mu_1 - \frac{1}{2}u\right) \leq \frac{16K}{nu^2}\log\left(\sqrt{\frac{n}{K}}u\right) + \frac{1}{nu^2/K - 1}.$$

Plugging this into (53) gives

$$\begin{aligned}
&\sum_{k=k_0+1}^{K} \mathbb{P}(Z < z_k)(\Delta_k - \Delta_{k-1}) \\
&\qquad \le \frac{16K}{n\Delta_{k_0+1}} \log\left(\sqrt{\frac{n}{K}}\Delta_{k_0+1}\right) + \frac{\Delta_{k_0+1}}{n\Delta_{k_0+1}^2/K - 1} \\
&\qquad\quad + \left[-\frac{16K}{nu}\log\left(e\sqrt{\frac{n}{K}}u\right) + \sqrt{\frac{K}{4n}}\log\left(\frac{\sqrt{\frac{n}{K}}u - 1}{\sqrt{\frac{n}{K}}u + 1}\right)\right]_{\Delta_{k_0+1}}^{1} \\
&\qquad \le \frac{16K}{n\Delta_{k_0+1}} \log\left(\frac{en\Delta_{k_0+1}^2}{K}\right) + \frac{\Delta_{k_0+1}}{n\Delta_{k_0+1}^2/K - 1} + \sqrt{\frac{K}{4n}}\log\left(\frac{\sqrt{\frac{n}{K}}\Delta_{k_0+1} + 1}{\sqrt{\frac{n}{K}}\Delta_{k_0+1} - 1}\right) \\
&\qquad \le \frac{16K}{n\Delta_{k_0+1}} \log\left(\frac{en\Delta_{k_0+1}^2}{K}\right) + \left(\frac{75}{74} + \frac{\sqrt{75}}{\sqrt{75}-1}\right)\frac{K}{n\Delta_{k_0+1}}
\end{aligned}$$

where the penultimate inequality uses $\Delta_{k_0+1} \ge \sqrt{\frac{75K}{n}}$ and $\log(1+x) \le x$ for any $x \ge 0$.

Gathering the results of the three steps, we get

$$\begin{aligned}
\overline{R}_n &\le n\Delta_{k_0} + \sum_{k=k_0+1}^{K}\left(1 + 7\frac{\log\left(\frac{n}{K}\Delta_k^2\right)}{\Delta_k} + \frac{\exp(-14c^2\log(75))}{2c^2(1-c^2)\Delta_k}\right) \\
&\qquad + \frac{16K}{\Delta_{k_0+1}}\log\left(\frac{en\Delta_{k_0+1}^2}{K}\right) + \left(\frac{75}{74} + \frac{\sqrt{75}}{\sqrt{75}-1}\right)\frac{K}{\Delta_{k_0+1}} \\
&\le n\Delta_{k_0} + K + (16+7)K\frac{\log\left(\frac{n}{K}\Delta_{k_0+1}^2\right)}{\Delta_{k_0+1}} + (16+16)\frac{K}{\Delta_{k_0+1}} \\
&\le n\delta_0 \mathbb{I}_{\Delta \le \delta_0} + 23K\frac{\log\left(\frac{n}{K}\Delta_{k_0+1}^2\right)}{\Delta_{k_0+1}} + \frac{33K}{\Delta_{k_0+1}} \\
&\le 23K\frac{\log\left(\frac{n}{K}\max(\Delta,\delta_0)^2\right)}{\max(\Delta,\delta_0)} + \frac{108K}{\max(\Delta,\delta_0)} \\
&\le 23K\frac{\log\left(\frac{110n}{K}\max(\Delta,\delta_0)^2\right)}{\max(\Delta,\delta_0)},
\end{aligned}$$

which implies (22) and also $\overline{R}_n \le 24\sqrt{nK}$. Since Proposition 34 implies $\mathbb{E}R_n - \overline{R}_n \le \sqrt{nK}$, we have proved (23). For (24), Proposition 36 implies

$$\mathbb{E}R_n - \overline{R}_n \le \min\left(\frac{K}{\Delta}, \frac{\sqrt{nK}}{2}\right) \le \frac{K\sqrt{75}}{2\max(\Delta,\delta_0)},$$

which implies

$$\mathbb{E}R_n \le 23K\frac{\log\left(\frac{133n}{K}\max(\Delta,\delta_0)^2\right)}{\max(\Delta,\delta_0)}.$$

Appendix D. Pseudo-regret vs Expected Regret

The first two propositions hold for the four prediction games considered in this work and defined in Figure 1.

Proposition 33 *For deterministic adversaries, we have* $\mathbb{E}R_n = \overline{R}_n$. *For oblivious adversaries, we have*

$$\mathbb{E}R_n \leq \sup_{\textit{deterministic adversaries}} \overline{R}_n.$$

In particular, this means that the worst oblivious adversary for a forecaster cannot lead to a larger regret than the worst deterministic adversary.

Proof The first assertion is trivial. For the second one, let $\mathbb{E}_{\text{adv}}$ be the expectation with respect to the eventual randomization of the adversary and $\mathbb{E}_{\text{for}}$ be the expectation with respect to the randomization of the forecaster. For oblivious adversaries, we have $\mathbb{E}R_n = \mathbb{E}_{\text{adv}}\mathbb{E}_{\text{for}}R_n$, hence

$$\mathbb{E}R_n \leq \sup_{\text{deterministic adversaries}} \mathbb{E}_{\text{for}}R_n = \sup_{\text{deterministic adversaries}} \overline{R}_n.$$

■

While the previous proposition is useful for upper bounding the regret of a forecaster against the worst oblivious adversary, it does not say anything about the difference between the expected regret and the pseudo-regret for a given adversary. The next proposition gives an upper bound on this difference for fully oblivious adversaries, which are (oblivious) adversaries generating independently the reward vectors.

Proposition 34 *For fully oblivious adversaries, we have*

$$\mathbb{E}R_n - \overline{R}_n \leq \sqrt{\frac{n \log K}{2}},$$

and

$$\mathbb{E}R_n - \overline{R}_n \leq \sqrt{2\log(K) \max_i \mathbb{E}\sum_{t=1}^n g_{i,t}} + \frac{\log K}{3}.$$

Proof The proof is similar to the one of the upper bound on the expected supremum of a finite number of subgaussian random variables. We use the following lemma

Lemma 35 *Let* $\lambda > 0$ *and* W *a random variable taking its values in* $[0,1]$. *We have*

$$\mathbb{E}\exp(\lambda W) \leq \exp\left[\big(\exp(\lambda) - 1\big)\mathbb{E}W\right].$$

Proof By convexity of the exponential function, we have $\exp(\lambda x) \leq 1 + \big(\exp(\lambda) - 1\big)x$ for any $x \in [0,1]$. So we have $\mathbb{E}\exp(\lambda W) \leq 1 + \big(\exp(\lambda) - 1\big)\mathbb{E}W$, hence Lemma 35. ■

Let $\lambda > 0$, then by Jensen's inequality and Lemma 35, we have

$$\begin{aligned}
\mathbb{E}\max_i \sum_{t=1}^n g_{i,t} &\leq \mathbb{E}\frac{1}{\lambda}\log\sum_{i=1}^K \exp\left(\lambda\sum_{t=1}^n g_{i,t}\right) \\
&\leq \frac{1}{\lambda}\log\sum_{i=1}^K \mathbb{E}\prod_{t=1}^n \exp(\lambda g_{i,t}) \\
&= \frac{1}{\lambda}\log\sum_{i=1}^K \prod_{t=1}^n \mathbb{E}\exp(\lambda g_{i,t}) \\
&\leq \frac{1}{\lambda}\log\sum_{i=1}^K \prod_{t=1}^n \exp\left([\exp(\lambda)-1]\mathbb{E}g_{i,t}\right) \\
&\leq \frac{\log K}{\lambda} + \frac{\exp(\lambda)-1}{\lambda}\max_i \mathbb{E}\sum_{t=1}^n g_{i,t}.
\end{aligned}$$

This implies

$$\mathbb{E}R_n - \overline{R}_n \leq \inf_{\lambda>0}\left(\frac{\log K}{\lambda} + \lambda\Theta(\lambda)\max_i \mathbb{E}\sum_{t=1}^n g_{i,t}\right),$$

where $\Theta(\lambda) = \frac{\exp(\lambda)-1-\lambda}{\lambda^2}$. By using Lemma 32, one obtains the second inequality of the theorem. Instead of using a variant of Bernstein's argument to control $\mathbb{E}\exp(\lambda g_{i,t})$, one can use Hoeffding's inequality. This leads to the first inequality by taking $\lambda = \sqrt{\frac{2\log K}{n}}$. ■

We can strengthen the previous result on the difference between the expected regret and the pseudo-regret when we consider the stochastic bandit game, in which the rewards coming from a given arm form an i.i.d. sequence. In particular, when there is a unique optimal arm, the following theorem states that the difference is exponentially small with n (instead of being of order $\sqrt{n}$).

Proposition 36 *For a given $\delta \geq 0$, let $I = \{i \in \{1,\dots,K\} : \Delta_i \leq \delta\}$ be the set of arms "δ-close" to the optimal ones, and $J = \{1,\dots,K\} \setminus I$ the remaining set of arms. In the stochastic bandit game, we have*

$$\mathbb{E}R_n - \overline{R}_n \leq \sqrt{\frac{n\log|I|}{2}} + \sum_{i\in J}\frac{1}{\Delta_i}\exp\left(-\frac{n\Delta_i^2}{2}\right),$$

and also

$$\mathbb{E}R_n - \overline{R}_n \leq \sqrt{\frac{n\log|I|}{2}} + \sum_{i\in J}\left\{\frac{8\sigma_i^2 + 4\Delta_i/3}{\Delta_i}\exp\left(-\frac{n\Delta_i^2}{8\sigma_i^2+4\Delta_i/3}\right) + \frac{8\sigma_{i^*}^2 + 4\Delta_i/3}{\Delta_i}\exp\left(-\frac{n\Delta_i^2}{8\sigma_{i^*}^2+4\Delta_i/3}\right)\right\},$$

where for any $j \in \{1, \ldots, K\}$, σ_j^2 denotes the variance of the reward distribution of arm j.

In particular when there exists a unique arm i^* such that $\Delta_{i^*} = 0$, we have

$$\mathbb{E}R_n - \overline{R}_n \leq \sum_{i \neq i^*} \frac{1}{\Delta_i} \exp\left(- \frac{n\Delta_i^2}{2} \right).$$

Note that the assumption on the uniqueness of the optimal arm in the last statement is necessary as we already discussed in Remark 25.

Proof Let $W_n^{(1)} = \max_{i \in I} \sum_{t=1}^n g_{i,t} - \sum_{t=1}^n g_{i^*,t}$ and $W_n^{(2)} = \max_{j \in \{1,\ldots,K\}} \sum_{t=1}^n g_{j,t} - \max_{i \in I} \sum_{t=1}^n g_{i,t}$. We have $\mathbb{E}R_n - \overline{R}_n = \mathbb{E}W_n^{(1)} + \mathbb{E}W_n^{(2)}$. From the same argument as in the proof of Proposition 34, we have

$$\mathbb{E}W_n^{(1)} \leq \sqrt{\frac{n \log |I|}{2}}.$$

Besides, we have

$$\begin{aligned}
\mathbb{E}W_n^{(2)} &= \int_0^{+\infty} \mathbb{P}(W_n^{(2)} > u) du \\
&\leq \sum_{j \in J} \int_0^{+\infty} \mathbb{P}\left(\sum_{t=1}^n g_{j,t} - \max_{i \in I} \sum_{t=1}^n g_{i,t} > u \right) du \\
&\leq \sum_{j \in J} \int_0^{+\infty} \mathbb{P}\left(G_{j,n} - G_{i^*,n} > u \right) du \\
&= \sum_{i \in J} \int_0^{+\infty} \mathbb{P}\left(G_{i,n} - \mathbb{E}G_{i,n} + \mathbb{E}G_{i^*,n} - G_{i^*,n} > u + n\Delta_i \right) du \\
&\leq \sum_{i \in J} \int_0^{+\infty} \left\{ \mathbb{P}\left(G_{i,n} - \mathbb{E}G_{i,n} > \frac{u + n\Delta_i}{2} \right) + \mathbb{P}\left(\mathbb{E}G_{i^*,n} - G_{i^*,n} > \frac{u + n\Delta_i}{2} \right) \right\} du.
\end{aligned}$$

This last integrand is upper bounded by $2\exp\left(-\frac{(u+n\Delta_i)^2}{2n}\right)$ from Hoeffding's inequality, and by $\exp\left(-\frac{(u+n\Delta_i)^2}{8n\sigma_i^2 + 4(u+n\Delta_i)/3}\right) + \exp\left(-\frac{(u+n\Delta_i)^2}{8n\sigma_{i^*}^2 + 4(u+n\Delta_i)/3}\right)$ from Bernstein's inequality. To control the two corresponding integrals, we note that for a nondecreasing convex function χ going to infinity at $+\infty$, we have

$$\int_x^{+\infty} \exp(-\chi(u)) du \leq \int_x^{+\infty} \frac{\chi'(u)}{\chi'(x)} \exp(-\chi(u)) du = \frac{\exp(-\chi(x))}{\chi'(x)}.$$

We apply this inequality to the functions $r \mapsto \frac{r^2}{2n}$ and $r \mapsto \frac{r^2}{8n\sigma_i^2 + 4r/3}$ to obtain respectively

$$\mathbb{E}W_n^{(2)} \leq 2 \sum_{i \in J} \int_{n\Delta_i}^{+\infty} \exp\left(-\frac{u^2}{2n} \right) du \leq \sum_{i \in J} \frac{1}{\Delta_i} \exp\left(- \frac{n\Delta_i^2}{2} \right),$$

and

$$\int_{n\Delta_i}^{+\infty} \exp\left(-\frac{u^2}{8n\sigma_i^2 + 4u/3}\right) du \leq \frac{(8\sigma_i^2 + 4\Delta_i/3)^2}{\Delta_i(16\sigma_i^2 + 4\Delta_i/3)} \exp\left(-\frac{n\Delta_i^2}{8\sigma_i^2 + 4\Delta_i/3}\right)$$
$$\leq \frac{8\sigma_i^2 + 4\Delta_i/3}{\Delta_i} \exp\left(-\frac{n\Delta_i^2}{8\sigma_i^2 + 4\Delta_i/3}\right),$$

hence

$$\mathbb{E}W_n^{(2)} \leq \sum_{i \in J} \left\{ \frac{8\sigma_i^2 + \frac{4\Delta_i}{3}}{\Delta_i} \exp\left(-\frac{n\Delta_i^2}{8\sigma_i^2 + 4\Delta_i/3}\right) + \frac{8\sigma_{i^*}^2 + \frac{4\Delta_i}{3}}{\Delta_i} \exp\left(-\frac{n\Delta_i^2}{8\sigma_{i^*}^2 + 4\Delta_i/3}\right) \right\}.$$

■

References

C. Allenberg, P. Auer, L. Györfi, and G. Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *ALT*, volume 4264 of *Lecture Notes in Computer Science*, pages 229–243. Springer, 2006.

J.-Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410:1876–1902, 2009.

P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. Gambling in a rigged casino: the adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, pages 322–331. IEEE Computer Society Press, 1995.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning Journal*, 47(2-3):235–256, 2002a.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19:357–367, 1967.

N. Cesa-Bianchi. Analysis of two gradient-based algorithms for on-line regression. *Journal of Computer and System Sciences*, 59(3):392–411, 1999.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

N. Cesa-Bianchi, Y. Freund, D.P. Helmbold, D. Haussler, R.E. Schapire, and M.K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Minimizing regret with label efficient prediction. *IEEE: Transactions on Information Theory*, 51:2152–2162, 2005.

D. A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3:100–118, 1975.

A. György and G. Ottucsák. Adaptive routing using expert advice. *Computer Journal-Oxford*, 49(2):180–189, 2006.

D. Helmbold and S. Panizza. Some label efficient learning results. In *Proceedings of the 10th annual conference on Computational learning theory*, pages 218–230. ACM New York, NY, USA, 1997.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

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