Constant regret for sequence prediction with limited advice

El Mehdi Saad, Gilles Blanchard

Constant regret for sequence prediction with limited advice

El Mehdi Saad[1] , Gilles Blanchard[1,2]
[1]Laboratoire de Mathématiques d'Orsay, CNRS, Université Paris-Saclay; [2]Inria

Abstract

We investigate the problem of cumulative regret minimization for individual sequence prediction with respect to the best expert in a finite family of size K under limited access to information. We assume that in each round, the learner can predict using a convex combination of at most p experts for prediction, then they can observe *a posteriori* the losses of at most m experts. We assume that the loss function is range-bounded and exp-concave. In the standard multi-armed bandits setting, when the learner is allowed to play only one expert per round and observe only its feedback, known optimal regret bounds are of the order $\mathcal{O}(\sqrt{KT})$. We show that allowing the learner to play one additional expert per round and observe one additional feedback improves substantially the guarantees on regret. We provide a strategy combining only $p = 2$ experts per round for prediction and observing $m \geq 2$ experts' losses. Its randomized regret (wrt. internal randomization of the learners' strategy) is of order $\mathcal{O}\big((K/m)\log(K\delta^{-1})\big)$ with probability $1-\delta$, i.e., is independent of the horizon T ("constant" or "fast rate" regret) if ($p \geq 2$ and $m \geq 3$). We prove that this rate is optimal up to a logarithmic factor in K. In the case $p = m = 2$, we provide an upper bound of order $\mathcal{O}(K^2 \log(K\delta^{-1}))$, with probability $1-\delta$. Our strategies do not require any prior knowledge of the horizon T nor of the confidence parameter δ. Finally, we show that if the learner is constrained to observe only one expert feedback per round, the worst-case regret is the "slow rate" $\Omega(\sqrt{KT})$, suggesting that synchronous observation of at least two experts per round is necessary to have a constant regret.

Keywords: Online Learning, Prediction with expert advice, Frugal Learning, Bandits feedback, Partial monitoring.

1 Introduction

We study the problem of online individual sequence prediction with expert advice, based on the setting presented by Cesa-Bianchi and Lugosi [2006, Chap. 2], under limited access to information. In this game, the learner's aim is to predict an unknown sequence $(y_1, y_2, \dots)$ of an outcome space $\mathcal{Y}$. The mismatch between the learner's predictions $(z_1, z_2, \dots)$, taking values in a closed convex subset $\mathcal{X}$ of a real vector space, and the target sequence is measured via a loss function $\ell(z, y)$. The learner's predictions may only depend on past observations. Following standard terminology used in prediction games, we will use the word "play" to mean the prediction output by the learner.

In each round $t \in [\![T]\!]$ (for a non-negative integer n, we denote $[\![n]\!] = \{1, \dots, n\}$), the learner has access to K experts predictions $(F_{1,t}, \dots, F_{K,t})$. The performance of the learner is

compared to that of the best single expert. More precisely, the objective is to have a cumulated regret as small as possible, where the regret is defined by

$$\mathcal{R}_T = \sum_{t=1}^{T} \ell(z_t, y_t) - \min_{i \in [\![K]\!]} \sum_{t=1}^{T} \ell(F_{i,t}, y_t).$$

Experts aggregation is a standard problem in machine learning, where the learner observes the predictions of all experts in each round and plays a convex combination of those. However, in many practical situations, querying the advice of every expert is unrealistic. Natural constraints arise, such as the financial cost of consultancy, time limitations in online systems, or computational budget constraints if each expert is actually the output of a complex prediction model. One might hope to make predictions in these scenarios while minimizing the underlying cost. Furthermore, we will distinguish between the constraint on the number of experts' advices used for prediction, and the number of feedbacks (losses of individual experts) observed *a posteriori*. This difference naturally arises in online settings where the advices are costly prior to the prediction task but just observing reported experts' losses after prediction can be cheaper. If the learner picks one single expert per round, plays the prediction of that expert and observes the resulting loss, the game is the standard multi-armed bandits problem. In this paper, we investigate intermediate settings, where the player has a constraint $p \leq K$ on the number of experts used for prediction (via convex combination) in each round and several feedbacks $m \leq K$ of actively chosen experts to see their losses. In the standard multi-armed bandit problem, the played arm is necessarily the observed arm, this restriction is known as the *coupling between exploitation and exploration*. In our protocol, we consider a generalization of that restriction through the *Inclusion Condition (IC)*: when $m \geq p$, if IC = True, we require that the set of played experts for prediction at round t, denoted S_t , is included in the set of observed experts, denoted C_t. More precisely, if IC = True, in each round t, the player first chooses p experts out of K and plays a convex combination of their prediction, then she observes the feedback (loss) of the individual selected experts, then picks $m - p$ additional experts to observe their losses. When IC = False, the choice of played and observed experts is decoupled; this means that the loss incurred by the p experts used for prediction is not necessarily observed.

A closely related question was considered by Seldin et al. [2014], obtaining $\mathcal{O}(\sqrt{T})$ regret bounds for a general loss function (see extended discussion in the next section.) Our emphasis here is on obtaining *constant bounds* guarantees on regret (i.e. independent of the time horizon T). Such "fast" rates, linked to assumptions related to strong convexity of the loss function ℓ, have been the subject of many works in learning (batch and online, in the stochastic setting) and optimization, but are comparatively under-explored in fixed sequence prediction.

In the literature on the prediction of fixed individual sequences, no assumptions are made about the distribution of the sequences. The attainability of fast rates (or constant regrets) is also possible under certain assumptions on the loss function ℓ: the full information setting was studied, mainly by Vovk [1990], Vovk [1998], Vovk [2001], where it was shown that fast rates are attainable under the *mixability* assumption on the loss function. The reader can find an extensive discussion of different assumptions considered in the literature for this problem in van Erven et al. [2015]. In the present paper, we make the following assumption on the loss function:

Protocol 1 The Game Protocol (p, m, IC).

Parameters:
p, the number of experts allowed for prediction.
m, the number of experts allowed for observation as feedback.
$\text{IC} \in \{\text{False}, \text{True}\}$, inclusion condition (if IC = True, we must have $p \leq m$).
for each round $t = 1, 2, \ldots, T$ **do**
Choose a subset $S_t \subseteq [\![K]\!]$ such that $|S_t| = p$, and convex combination weights $(\alpha_i)_{i \in S_t}$.
Play the convex combination $\sum_{i \in S_t} \alpha_{i,t} F_{i,t}$ and incur its loss.
if IC = True, **then**
Choose a subset $C_t \subseteq [\![K]\!]$ such that: $|C_t| = m$ and $S_t \subseteq C_t$.
else if IC = False, **then**
Choose a subset $C_t \subseteq [\![K]\!]$ such that: $|C_t| = m$.
end if
The environment reveals the losses $(\ell(F_{i,t}, y_t))_{i \in C_t}$.
end for

Assumption 1. *There exist $B, \eta > 0$, such that*

- ***Exp-concavity:*** *For all $y \in \mathcal{Y}$, $\ell(., y)$ is η-exp-concave over domain $\mathcal{X}$.*
- ***Range-boundedness:*** *For all $y \in \mathcal{Y}$:* $\sup_{x,x' \in \mathcal{X}} |\ell(x, y) - \ell(x', y)| \leq B$.

Remarks. *This assumption is satisfied in some usual settings of learning theory such as the least squares loss with bounded outputs: $\mathcal{X} = \mathcal{Y} = [x_{\min}, x_{\max}]$ and $\ell(x, x') = (x - x')^2$. Then ℓ satisfies Assumption 1, with $B = (x_{\max} - x_{\min})^2$ and $\eta = 1/(2B)$.*

Remarks. *The regret as well as all the algorithms to follow remain unchanged if we replace ℓ by $\tilde{\ell} : \mathcal{X} \to [0, B]$ defined by $\tilde{\ell}(x, y) := \ell(x, y) - \min_{x \in \mathcal{X}} \ell(x, y)$, so we can assume without loss of generality $\ell \in [0, B]$ instead of range-boundedness; the results obtained still hold in the latter more general case.*

Assumption 1 was considered in several previous works tracking fast rates both in batch and online learning (Koren and Levy, 2015, Mehta, 2017, Gonen and Shalev-Shwartz, 2016, Mahdavi et al., 2015, van Erven et al., 2015). We introduce a new characterization for the class of functions satisfying Assumption 1. Let $c > 0$, define $\mathcal{E}(c)$ as the class of functions $f : \mathcal{X} \to \mathbb{R}$, such that

$$\forall x, x' \in \mathcal{X} : \quad f\left(\frac{x + x'}{2}\right) \leq \frac{1}{2} f(x) + \frac{1}{2} f(x') - \frac{1}{2c} \big(f(x) - f(x')\big)^2. \tag{1}$$

We introduce this class to highlight the sufficient and minimal property of ℓ required for the proofs in this paper to work, namely we will only make use of (1) in the proofs of the results to come.

Lemma 1.1 below relates the class of functions $\mathcal{E}(.)$ to the set of functions satisfying Assumption 1 as well a sufficient condition (Lipschitz and Strongly Convex or LIST condition).

Lemma 1.1. *Let $y \in \mathcal{Y}$ be fixed.*

- *If $\ell(.,y)$ is B-range-bounded and η-exp-concave, then: $\ell(.,y) \in \mathcal{E}\left(\frac{\eta B^2}{4\log\left(1+\frac{\eta^2 B^2}{2}\right)}\right)$.*
- *If $\ell(.,y) \in \mathcal{E}(c)$ and is continuous, then: $\ell(.,y)$ is c-range-bounded and $(4/c)$-exp-concave.*
- *If $\ell(.,y)$ is L-Lipschitz and ρ-strongly convex, then $\ell(.,y) \in \mathcal{E}(4L^2/\rho)$.*

Figure 1 summarizes bounds on regret for bounded and exp-concave loss functions. We only consider fixed individual sequences, which corresponds to fully oblivious adversaries (see Audibert and Bubeck, 2010 for a definition of different types of adversaries).

	$p=1$		$p \geq 2$	
	Lower bound	Upper bound	Lower bound	Upper bound ($p=2$)
$m=1$	$\sqrt{KT}$ [1]	$\sqrt{KT}$ [2]	$\sqrt{KT}$ [Thm 5.2]	$\sqrt{KT}$ [2]
$m=2$	$\sqrt{KT}$ [3]	$\sqrt{KT}$ [2]	K [Thm 5.1]	**IC** = True : $K^2\log(K)$ **IC** = False : $K\log(K)$ [Thm 4.2 and 4.1]
$m \geq 3$	$\sqrt{\frac{K}{m}T}$ [3]	$\sqrt{\frac{K}{m}T\log(K)}$ [3]	$\frac{K}{m}$ [Thm 5.1]	$\frac{K}{m}\log(K)$ [Thm 4.1]

Figure 1: Existing bounds from the literature ([1] = Auer et al., 2002, [2]=Audibert and Bubeck, 2010, [3]=Seldin et al., 2014) and new bounds presented in this paper. All bounds hold up to numerical constant factors. Under Assumption 1, all new upper bounds hold with high probability if we replace the factor $\log(K)$ with $\log(K\delta^{-1})$, δ being the confidence parameter. Lower bounds are in expectation. When bounds are the same, we omit the distinction between the settings **IC** = True and **IC** = False (coupling between exploration and exploitation, see Protocol 1).

The remainder of this paper is organized as follows. Section 2 presents some results from the literature relevant to the studied problem. Section 3 introduces algorithms satisfying constant regrets in expectation in the case $p=2$ and $m \geq 3$; that section aims to present a preliminary view of the intuitions for attaining our objective. Next, we present in Section 4 our main results consisting of algorithms satisfying constant regrets with a high probability for $p, m \geq 2$. Finally, in Section 5, we present lower bounds for all the possible settings.

2 Discussion of related work

Games with limited feedback and $\mathcal{O}(\sqrt{T})$ regret: In the standard setting of multi-armed bandit problem, the learner has to repeatedly obtain rewards (or incur losses) by choosing from a fixed set of k actions and gets to see only the reward of the chosen action. Algorithms such as EXP3-IX [Neu, 2015] or EXP3.P [Auer et al., 2002] achieve the optimal regret of order $\mathcal{O}(\sqrt{KT})$ up to a logarithmic factor, with high probability. A more general setting closer to ours was introduced by Seldin et al. [2014]. Given a budget $m \in [\![K]\!]$, in each round t, the learner plays the prediction of one expert I_t, then gets to choose a subset of experts C_t such that $I_t \in C_t$ in order to see their prediction. A careful adaptation of the EXP3 algorithm to this setting leads to an expected regret of order $\mathcal{O}(\sqrt{(K/m)T})$, which is optimal up to logarithmic factor in K.

There are two significant differences between our framework and the setting presented by Seldin et al. [2014]. First, we allow the player to combine up to p experts out of K in each round for prediction. Second, we make an additional exp-concavity-type assumption (Assumption 1) on the loss function. These two differences allow us to achieve constant regrets bounds (independent of T).

Playing multiple arms per round was considered in the literature of multiple-play multi-armed bandits. This problem was investigated under a budget constraint C by Zhou and Tomlin [2018] and Xia et al. [2016]. In each round, the player picks m out of K arms, incurs the sum of their losses. In addition to observing the losses of the played arms, the learner learns a vector of costs which has to be covered by a pre-defined budget C. Once the budget is consumed, the game finishes. An extension of the EXP3 algorithm allows deriving a strategy in the adversarial setting with regret of order $\mathcal{O}(\sqrt{KC\log(K/m)})$. The cost of each arm is supposed to be in an interval $[c_{\min}, 1]$, for a positive constant $c_{\min}$. Hence the total number of rounds in this game T satisfies $T = \Theta(C/m)$. Another online problem aims at minimizing the cumulative regret in an adversarial setting with a small effective range of losses. Gerchinovitz and Lattimore [2016] have shown the impossibility of regret scaling with the effective range of losses in the bandit setting, while Thune and Seldin [2018] showed that it is possible to circumvent this impossibility result if the player is allowed one additional observation per round. However, it is impossible to achieve a regret dependence on T better than the rate of order $\mathcal{O}(\sqrt{T})$ in this setting.

Decoupling exploration and exploitation was considered by Avner et al. [2012]. In each round, the player plays one arm, then chooses one arm out of K to see its prediction (not necessarily the played arm as in the canonical multi-armed bandits problem). They devised algorithms for this setting and showed that the dependence on the number of arms K can be improved. However, it is impossible to achieve a regret dependence on T better than $\mathcal{O}(\sqrt{T})$.

Prediction with limited expert advice was also investigated by Helmbold and Panizza [1997],Cesa-Bianchi and Lugosi [2006, Chap. 6] and Cesa-Bianchi et al. [2005]. However, in these problems, known as label efficient prediction, the forecaster has full access to the experts advice but limited information about the past outcomes of the sequence to be predicted. More precisely, the outcome y_t is not necessarily revealed to the learner. In such a framework, the optimal regret is of order $\mathcal{O}(\sqrt{T})$.

Constant regrets in the full information setting: The setting where the learner plays a combination of all the experts and is allowed to see all their predictions in each round is known in the literature as experts aggregation problem. It is a well-established framework [Cesa-Bianchi and Lugosi, 2006] studied earlier by Freund and Schapire [1997], Kivinen and Warmuth [1999], Vovk [1998]. This setting was investigated under the assumption that the loss ℓ function is η-exp-concave (i.e., the function $\exp(-\eta\ell)$ is concave). The Weighted Average Algorithm algorithm [Kivinen and Warmuth, 1999] is known to achieve a constant regret of order $\mathcal{O}(\log(K)/\eta)$. While this result holds for any sequence of target variable and experts, it requires using a combination of all the experts in each round. In several situations, it is desirable to query and use the least number possible of experts advice for various reasons (such as cost or time restrictions). In this paper, we aim at achieving the same bounds (with high probability) under such constraints.

Fast rates in the batch setting: Another line of works investigated the problem of experts (or estimators) aggregation in the batch setting with stochastic and i.i.d samples (i.e., each expert's predictions are assumed to follow an independent and identical distribution, see Tsybakov, 2003). There are two distinct phases: a first step where the learner has access to training data points, then a prediction step where she outputs a combination of experts. The output in this setting is compared against the best expert. A non-exhaustive list of works considering this problem includes those of Audibert [2008], Lecué and Mendelson [2009], and Saad and Blanchard [2021], where the emphasis was put on obtaining $\mathcal{O}(1/T)$ "fast" rates for excess risk with high probability under some convexity assumptions on the loss function. However, these algorithms are not translatable to the adversarial setting since some of the previous strategies rely on the early elimination of sub-optimal experts. Saad and Blanchard [2021] presented a budgeted setting where the learner is constrained to see at most m experts forecasts per data point and can predict using p experts. This paper is an extension of their framework in the adversarial setting with a cumulative regret.

Online Convex Optimization with bandit feedback: A different objective is considered in the online convex optimization framework, where the losses are compared against the best convex combination of the experts. This problem was studied by Agarwal et al. [2010] and Shamir [2017] under limited feedback. More precisely, the learner can query the value of the loss function in two points from the convex envelope of the compact set over which the optimization is performed. In such a setting, it was shown that for Lipschitz and strongly-convex loss functions, it is possible to achieve an expected regret bounded by $\mathcal{O}(d^2 \log(T))$, where d is the dimension of the linear span of experts (which plays a similar role to K in our setting). Observe that online convex optimization algorithms (eg. as considered in the cited references) cannot be applied in our setting, where the player is not allowed to play (or observe) an arbitrary point in the convex envelope of the experts, but rather convex combinations with support on p (or m) experts. On the other hand, the goal aimed at is different as well, since we want to minimize the regret with respect to the best expert, not with respect to the best convex combination of experts (which would not be an attainable goal under the considered play restrictions).

Why aim at high probability bounds instead of expectation bounds? Consider an algorithm with internal randomization. From a practical point of view, bounds on its expected regret do not necessarily translate into a similar guarantee with high probability. In many applications, such as finance, controlling the fluctuations of risk is very important. From a mathematical point of view, the "phenomenon" of negative regrets occurs when the player has a chance of outperforming the benchmark (such as the best-fixed expert in hindsight) for some rounds. In this case, an algorithm may have optimal expected regret but sub-optimal deviations. A manifestation of this problem is for the EXP3 algorithm in multi-armed bandit setting ($p = m = 1$ in Protocol 1), which has a worst case regret of $\sqrt{KT}$ in expectation, but the random regret can be linear $\Omega(T)$ with constant probability (see the exercises of Chapter 11 of Lattimore and Szepesvári, 2020).

3 Main results: Algorithm with upper bounds in expectation

In this section, we introduce a new algorithm with constant bounds on the expected regret, for the setting: $p = 2$ and $m \geq 3$. The aim of this section is to present some central intuitions, which are complemented in the next section to achieve stronger guarantees. To ease notation, we denote for each $i \in [\![K]\!]$ and $t \in [\![T]\!]$: $\ell_{i,t} := \ell(F_{i,t}, y_t)$.

The high-level idea of Algorithm 2 is common in the literature. It consists in constructing unbiased estimates of unseen losses, which are fed to the classical exponential weighting (EW) scheme over the experts. The first novelty introduced here is that the estimates are centered in a "data-dependent" way, whose goal is to reduce variance. This variance control is essential in our analysis (see sketch of the proof below) in order to have constant regrets.

Let us denote $\hat{p}_t$ the probability distribution derived by the EW principle using estimated cumulated losses $\hat{L}_{i,t}$ over the set of experts at round t. The second novelty consists in sampling just two experts I_t and J_t, independently at random following $\hat{p}_t$, and $m-2$ additional experts uniformly at random for exploration. Then, we play the mid-point of the predictions of I_t and J_t (i.e., predict we predict $\frac{1}{2}F_{I_t,t} + \frac{1}{2}F_{J_t,t}$).

The main idea for getting a constant regret bound is to compensate the variance term introduced by the estimates $(\hat{\ell}_{i,t})$ by the negative second order term in inequality (1) satisfied by the loss. The following theorem presents a constant bound on the expected regret, with a sketch of the proof.

Define the following constant

$$\bar{\lambda} := \min\left\{ \frac{4\log\left(1 + \frac{\eta^2 B^2}{2}\right)}{\eta B^2},\ \frac{1}{B} \right\}. \tag{2}$$

Theorem 3.1. *Suppose Assumption 1 holds. For any input parameter: $\lambda \in \left(0, \frac{m-2}{4K}\bar{\lambda}\right)$, where $\bar{\lambda}$ is defined in* (2)*, the expected regret of Algorithm 2 satisfies:*

$$\mathbb{E}[\mathcal{R}_T] \leq \frac{\log(K)}{\lambda},$$

where the expectation is with respect to the learner's own randomization.

Algorithm 2 Prediction with limited advice ($p = 2, m \geq 3$)

Input Parameters: λ, m.
Initialize: $\hat{L}_{i,0} = 0$ for all $i \in [\![K]\!]$.
for each round $t = 1, 2, \dots$ **do**
Let

$$\hat{p}_{i,t} = \frac{\exp\Big(-\lambda \hat{L}_{i,t-1}\Big)}{\sum_j \exp\Big(-\lambda \hat{L}_{j,t-1}\Big)}.$$

Draw I_t and J_t according to $\hat{p}_t$ independently.
Play: $\frac{1}{2}F_{I_t,t} + \frac{1}{2}F_{J_t,t}$, and incur its loss.
Sample $m-2$ experts uniformly at random without replacement from $[\![K]\!]$. Denote $\mathcal{U}_t$ this set of experts.
Query $C_t = \mathcal{U}_t \cup \{I_t, J_t\}$.
for $i \in [\![K]\!]$ **do**
Let

$$\hat{\ell}_{i,t} = \frac{K}{m-2}\mathbb{1}(i \in \mathcal{U}_t)\ \ell_{i,t} + \left(1 - \frac{K}{m-2}\mathbb{1}(i \in \mathcal{U}_t)\right)\ \ell_{I_t,t}.$$

Update $\hat{L}_{i,t} = \hat{L}_{i,t-1} + \hat{\ell}_{i,t}$.
end for
end for

Remarks. *Comparing this result with the guarantees of the classical exponential weights averaging (EWA) algorithm, one can notice that in the full information feedback setting ($m = K$), our guarantee is of the same order, up to a numerical constant, as the constant regret bound for EWA for exp-concave losses. The advantage of our procedure is that it necessitates sampling only two experts from the EW distribution instead of full averaging. In the partial feedback case ($m < K$), Algorithm 2 guarantees a regret of order $\mathcal{O}(K \log(K)/m)$, as one would expect, the factor K/m reflects the proportion of the information available to the learner. The last bound is tight, up to a logarithmic factor in K (see Theorem 5.1).*

Sketch of the proof. Let $(\mathcal{F}_t)$ denote the natural filtration associated to the process of available information, $(S_t, C_t, (\ell_{i,t})_{t \in C_t})$, and denote $\mathbb{P}_{t-1}$ resp. $\mathbb{E}_{t-1}$ the conditional probability resp. expectation with respect to $\mathcal{F}_{t-1}$ ("past observations"). The loss functions ℓ_t satisfy Assumption 1. Therefore, using Lemma 1.1, the expected cumulative loss of Algorithm 2 is given by

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}\left[\ell_t\left(\frac{F_{I_t,t} + F_{J_t,t}}{2}\right)\right] &\leq \sum_{t=1}^{T} \mathbb{E}\left[\frac{1}{2}\ell_{I_t,t} + \frac{1}{2}\ell_{J_t,t} - \frac{\bar{\lambda}}{2}(\ell_{I_t,t} - \ell_{J_t,t})^2\right] \\ &= \underbrace{\sum_{t=1}^{T}\sum_{i=1}^{K} \mathbb{E}[\hat{p}_{i,t}\ \ell_{i,t}]}_{\text{Term 1}} - \underbrace{\frac{\bar{\lambda}}{2}\sum_{t=1}^{T}\sum_{i,j=1}^{K} \mathbb{E}\Big[\hat{p}_{i,t}\hat{p}_{j,t}(\ell_{i,t} - \ell_{j,t})^2\Big]}_{\text{Term 2}}. \qquad (3)\end{aligned}$$

Observe that by construction of Algorithm 2, the elements in $\mathcal{U}_t$ were sampled uniformly at random without replacement from $[\![K]\!]$. Moreover, $\mathcal{U}_t$ is independent of I_t. Therefore, $\hat{\ell}_{i,t}$ is an unbiased estimator of $\ell_{i,t}$ conditionally to the available information: $\mathbb{E}_{t-1}[\hat{\ell}_{i,t}] = \ell_{i,t}$.

Using the tower rule, Term 1 therefore writes $\sum_t \sum_i \mathbb{E}[\hat{p}_{i,t}\hat{\ell}_{i,t}]$. Next, we use Lemma E.1 in the Appendix (by cancellation of consecutive logarithmic terms) with $\mu_t = \sum_{i=1}^K \hat{p}_{i,t}\ell_{i,t}$ for each $t \in [\![T]\!]$. We have the following upper bound for Term 1 in (3):

$$\sum_{t=1}^{T}\sum_{i=1}^{K}\mathbb{E}\Big[\hat{p}_{i,t}\ \hat{\ell}_{i,t}\Big] \leq \min_{i\in[\![K]\!]}\sum_{t=1}^{T}\mathbb{E}\Big[\hat{\ell}_{i,t}\Big] + \frac{\log(K)}{\lambda} + \lambda\sum_{t=1}^{T}\sum_{i=1}^{K}\mathbb{E}\Big[\hat{p}_{i,t}\Big(\hat{\ell}_{i,t}-\mu_t\Big)^2\Big]. \tag{4}$$

We use the definition of $\hat{\ell}_{i,t}$ and the tower rule to upper bound the last term in (3):

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{K}\hat{p}_{i,t}\Big(\hat{\ell}_{i,t}-\mu_t\Big)^2\right] &\leq \frac{2K}{m-2}\mathbb{E}\left[\sum_{i=1}^{K}\hat{p}_{i,t}(\ell_{i,t}-\mu_t)^2\right] + \frac{2K}{m-2}\mathbb{E}\Big[(\ell_{I_t,t}-\mu_t)^2\Big]\\ &= \frac{4K}{m-2}\mathbb{E}\left[\sum_{i=1}^{K}\hat{p}_{i,t}(\ell_{i,t}-\mu_t)^2\right].\end{aligned}$$

Finally, we combine (3), (4) and the bound above to obtain

$$\mathbb{E}[\mathcal{R}_T] \leq \frac{\log(K)}{\lambda} + \lambda\frac{4K}{m-2}\mathbb{E}\left[\sum_{i=1}^{K}\hat{p}_{i,t}(\ell_{i,t}-\mu_t)^2\right] - \bar{\lambda}\sum_{t=1}^{T}\sum_{i,j=1}^{K}\mathbb{E}\Big[\hat{p}_{i,t}\hat{p}_{j,t}(\ell_{i,t}-\ell_{j,t})^2\Big].$$

Recall that if X and Y are two independent and identically distributed variables, we have $\mathbb{E}[(X-Y)^2] = 2\,\mathrm{Var}(X)$. Applying this identity to Term 2 in (3), we have

$$\mathbb{E}[\mathcal{R}_T] \leq \frac{\log(K)}{\lambda} + \left(\lambda\frac{4K}{m-2} - \frac{1}{B}\right)\mathbb{E}\left[\sum_{i=1}^{K}\hat{p}_{i,t}(\ell_{i,t}-\mu_t)^2\right].$$

We conclude using $\lambda < \frac{m-2}{4K}\bar{\lambda}$. □

4 Main results: Algorithms with high probability upper bounds

In this section, we present new algorithms with guarantees that hold with high probability with respect to the player's own randomization. As discussed in Section 2, high probability guarantees are important to assess any algorithm's goodness due to potential exposure to negative regrets phenomena and thus the possibility of deviations having larger order than the expectation.

We introduce sampling strategies for three different settings: $p = 2$ and $m \geq 3$, ($p = 2, m = 2, \mathrm{IC} = \mathrm{False}$) and ($p = 2, m = 2, \mathrm{IC} = \mathrm{True}$), presented in Algorithms 3 and 4; Algorithm 3 is common to the first two settings. To ease notations, we denote for each $i \in [\![K]\!]$ and $t \in [\![T]\!]$: $\ell_{i,t} := \ell(F_{i,t}, y_t)$.

In Algorithms 3 and 4, we build on the idea presented in Algorithm 2 and construct estimates of unseen losses, which are fed into an EW scheme from which experts are sampled. Let $\hat{p}_t$ denotes the resulting estimated EW distribution. The main differences between the algorithms below and Algorithm 2 are (a) the constructed loss estimates and (b) the sampling strategy when $m = 2$ and $\mathrm{IC} = \mathrm{True}$.

Modified loss estimates: We start with the same unbiased loss estimates, with data-dependent centering, from Algorithm 2, but additionally introduce a *negative* (or "optimistic") bias on the estimated losses, which takes into account an estimated variance. This can be conceptually compared to the uniform confidence bound (UCB) algorithm in the standard stochastic bandit setting, which will select "optimistically" arms which have the highest potential reward given past information (here, loss is a negative reward). In this sense, this term tends to encourage diversity in expert sampling (i.e. encourage sampling experts with a possibly higher estimated loss but also larger variance than the best estimated experts so far). This is used in both Algorithms 3 and 4.

In the case $m \geq 3$ or $(m = 2, \mathrm{IC} = \mathrm{False})$, there is still at least one free observation left for exploration decoupled from exploitation. In these settings, Algorithm 3 uses the same sampling scheme as Algorithm 2, namely sampling independently at random two experts following $\hat{p}_t$ and playing the central point of the sampled predictions. The remaining "pure exploration" observations are sampled uniformly at random, with replacement.

Modified sampling scheme: the case $(m = 2, \mathrm{IC} = \mathrm{True})$ is more difficult since there is no "free exploration" observation possible. This is the counterpart of the exploration/exploitation tradeoff of the standard bandit setting, in the framework where we aim at constant regrets (so that playing combinations of at least two arms is necessary, see next section). Taking inspiration from the standard bandit setting literature ($p = m = 1$), introducing a small uniform exploration component appears necessary for the sampling strategy for algorithms achieving optimal high probability guarantees (Audibert and Bubeck, 2010, Auer et al., 2002, Beygelzimer et al., 2011, Bubeck and Cesa-Bianchi, 2012). For example, EXP3.P mixes the EW sampling rule with a uniform distribution over the arms. On the other hand, EXP-IX [Neu, 2015] incorporates the exploration component implicitly through a biased estimate of the losses. However, this uniform exploration costs $\mathcal{O}(\sqrt{KT})$ on the cumulative regret. Hence, aiming at constant regret necessitates a more subtle sampling rule.

We introduce a two-step sampling strategy. The first expert, denoted A_t, is sampled following $\hat{p}_t$. The second expert, denoted B_t, is sampled uniformly at random (possibly B_t and A_t are identical). The predictions of (A_t, B_t) are observed after making a prediction. For the playing strategy, we sample two experts independently (conditionally to A_t and B_t) at random, following the restriction of the law $\hat{p}_t$ on $\{A_t, B_t\}$, and we play the central point of the two sampled experts. Therefore, depending on the outcome of the second step, the algorithm's prediction can be either one of the two pre-selected experts or the central point of the two experts. This strategy ensures the necessary uniform exploring component needed in the adversarial problems.

The possibility of having constant regrets guarantees is due to Property (1), satisfied for the loss functions ℓ under Assumption 1: Lemma 1.1 suggests that when predicting the central point of two experts, the learner benefits from the distance between the played predictions. This remark is exploited in constructing of the distribution $\hat{p}_t$.

To summarize, the playing strategy relies on three essential ideas: the (conditional for $m = 2$) independence of the played experts, the centering scheme for the losses estimates, and the second order term to diversify the played arms.

Remarks.

- *The proposed algorithm can be implemented in an efficient way, so that after a one-time computational cost of $\mathcal{O}(K)$ for initialization, the computational cost of each*

Algorithm 3 ($p = 2$, $m \geq 3$) or ($p = 2$, $m = 2$, IC = False)

Input Parameters: λ, m.
Initialize: $\hat{L}_{i,0} = 0, \hat{V}_{i,0} = 0$ for all $i \in [\![K]\!]$.
Let $\tilde{m} = \max\{m - 2, 1\}$.
for each round $t = 1, 2, \ldots$ **do**
Let

$$\hat{p}_{i,t} = \frac{\exp\Big(-\lambda \hat{L}_{i,t-1} + \lambda^2 \hat{V}_{i,t-1}\Big)}{\sum_{j=1}^{K} \exp\Big(-\lambda \hat{L}_{j,t-1} + \lambda^2 \hat{V}_{j,t-1}\Big)}. \tag{5}$$

Sample I_t and J_t according to $\hat{p}_t$ from $[\![K]\!]$ independently.
Play: $\frac{1}{2}F_{I_t,t} + \frac{1}{2}F_{J_t,t}$, and incur its loss.
Sample $\tilde{m}$ experts without replacement, independently and uniformly at random from $[\![K]\!]$. Denote $\mathcal{U}_t$ this set of experts.
if $m \geq 3$ **then**
Let $C_t = \{I_t, J_t\} \cup \mathcal{U}_t$.
else if $m = 2$ **then**
Let $C_t = \{I_t\} \cup \mathcal{U}_t$.
end if
Observe: $\ell_{i,t}$ for $i \in C_t$.
for $i \in [\![K]\!]$ **do**
Let

$$\hat{\ell}_{i,t} = \frac{K}{\tilde{m}} \mathbb{1}(i \in \mathcal{U}_t)\ \ell_{i,t} + \left(1 - \frac{K}{\tilde{m}} \mathbb{1}(i \in \mathcal{U}_t)\right)\ \ell_{I_t,t} \tag{6}$$

$$\hat{v}_{i,t} = \Big(\hat{\ell}_{i,t} - \ell_{I_t,t}\Big)^2 \tag{7}$$

Update $\hat{L}_{i,t} = \hat{L}_{i,t-1} + \hat{\ell}_{i,t}$ and $\hat{V}_{i,t} = \hat{V}_{i,t-1} + \hat{v}_{i,t}$.
end for
end for

round, including suitably keeping track of the distribution $\hat{p}_t$ and sampling from it, is $\mathcal{O}(m \log K)$ (see Appendix K for details). Therefore, the computational complexity also depends mildly on the number of experts K.

- *Since our analysis suggests that we can restrict possible plays to mid-points of just two experts, one could argue that the coupled setting ($p = m = 2$, IC=True) looks quite similar to learning with expert advice with bandit feedback, where the possible arms would be the K^2 "bi-experts" that are mid-points of original experts (i, j). One could therefore think of a more direct approach: simply applying a bandit-type strategy, say EXP3.P or EXP3-IX (Auer et al., 2002 and Neu, 2015, respectively) to these K^2 "arms". However, existing generic results only guarantee a "slow" $\mathcal{O}(\sqrt{T})$ regret with respect to the best "bi-expert", and this cannot be compensated in general by exp-concavity, as the best "bi-expert" may not be much better than the best expert (if the experts are "correlated": see proof of lower bounds in Theorem 5.1 and 5.2). Furthermore, in the playing strategy of EXP3.P and EXP3-IX, each pair of experts is played $\Omega(\sqrt{K^2 T})$ times, due the uniform exploration component of their sampling schemes. This will lead regrets scaling with $\sqrt{T}$.*

Algorithm 4 ($p = 2$, $m = 2$, IC = True)

Input Parameters: λ.
Initialize: $\hat{L}_{i,0} = 0$ for all $i \in [\![K]\!]$.
for each round $t = 1, 2, \ldots$ **do**
Let

$$\hat{p}_{i,t} = \frac{\exp\Big(-\lambda \hat{L}_{i,t-1} + \lambda^2 \hat{V}_{i,t-1}\Big)}{\sum_{j=1}^{K} \exp\Big(-\lambda \hat{L}_{j,t-1} + \lambda^2 \hat{V}_{j,t-1}\Big)}.$$

Sample one expert from $[\![K]\!]$, denoted A_t, according to $\hat{p}_t$, and one expert from $[\![K]\!]$, denoted B_t, independently and uniformly at random. Let $C_t = \{A_t, B_t\}$.
for $i \in C_t$ **do**
Let

$$\hat{q}_{i,t} = \frac{\exp\Big(-\lambda \hat{L}_{i,t-1} + \lambda^2 \hat{V}_{i,t-1}\Big)}{\sum_{j \in C_t} \exp\Big(-\lambda \hat{L}_{j,t-1} + \lambda^2 \hat{V}_{j,t-1}\Big)}.$$

Draw I_t from C_t according to $\hat{q}_t$.
Draw J_t from C_t according to $\hat{q}_t$ independently from I_t.
Play: $\frac{1}{2}F_{I_t,t} + \frac{1}{2}F_{J_t,t}$, and incur its loss.
Observe: $\ell_{i,t}$ for $i \in C_t$.
end for
for $i \in [\![K]\!]$ **do**
Let

$$\hat{\ell}_{i,t} = K\ \mathbb{1}(B_t = i)\ \ell_{i,t} + (1 - K\ \mathbb{1}(B_t = i))\ \ell_{A_t,t}$$
$$\hat{v}_{i,t} = \Big(\hat{\ell}_{i,t} - \ell_{A_t,t}\Big)^2$$

Update: $\hat{L}_{i,t} = \hat{L}_{i,t-1} + \hat{\ell}_{i,t}$ and $\hat{V}_{i,t} = \hat{V}_{i,t-1} + \hat{v}_{i,t}$.
end for
end for

Theorem 4.1. *Suppose Assumption 1 holds.*

Consider the case ($m \geq 3$ *and* $p = 2$) *or* ($m = 2$ *and* $p = 2$ *and IC = False*). *For any input parameter* $\lambda \in \Big(0, \frac{m-1}{128K}\bar{\lambda}\Big)$, *where* $\bar{\lambda}$ *is defined in* (2), *the regret of Algorithm 3 satisfies with probability at least* $1 - 8\delta$, *with respect to the player's own randomization*

$$\mathcal{R}_T \leq c\ \frac{1}{\lambda} \log\left(\frac{\bar{\lambda} K}{\lambda \delta}\right),$$

where c is a numerical constant.

Theorem 4.2. *Suppose Assumption 1 holds.*

Consider the case $p = m = 2$ *and IC = True. For any input parameter* $\lambda \in \Big(0, \frac{\bar{\lambda}}{352K^2}\Big)$, *where* $\bar{\lambda}$ *is defined in* (2), *the regret of Algorithm 4 satisfies with probability at least* $1 - 8\delta$,

with respect to the player's own randomization

$$\mathcal{R}_T \leq c \left(\frac{1}{\lambda} + \frac{K}{\bar{\lambda}}\right) \log\left(\frac{\bar{\lambda} K}{\lambda \delta}\right),$$

where c is a numerical constant.

Discussion Notice that prior knowledge on the confidence level δ is not required by Algorithms 3 and 4. The presented bounds in theorems above are valid for any $\delta \in (0,1)$. Observe that taking λ close to $m/(128K)$ $\bar{\lambda}$ leads to a bound of the order $\mathcal{O}(K \log(K\delta^{-1})/m)$ in Theorem 4.1, which is minimax optimal up to a $\log(K)$ factor (Theorem 5.1). Taking λ close to $1/(352K^2)$ $\bar{\lambda}$, leads to a bound of the order $\mathcal{O}(K^2 \log(K\delta^{-1}))$ in the special setting $p = m = 2$ with IC = True. This bound presents a gap of factor K with the lower bound presented in Theorem 5.1. We emphasize that in the last setting, the player chooses two experts to combine their predictions and observes only the feedback of these two experts. Hence, unlike the setting considered in Theorem 4.1, the player is deprived of additional 'freely chosen' experts to explore their losses. This constraint necessitates a more careful playing strategy, presented in Algorithm 4.

5 Lower bounds

In this section, we provide lower bounds matching the upper bounds in Theorem 4.1, up to a logarithmic factor in K (except for the case $p = m = 2$, where we have a gap of factor K). The techniques of the proof are similar to the ones presented by Auer et al. [1995]. The main difference comes from the construction of the experts' distributions.

Theorem 5.1. *Let ℓ be the squared loss: $\ell(x,y) = (x-y)^2$ on $\mathcal{X} = \mathcal{Y} = [0,1]$. Consider the game protocol presented in Algorithm 1 with $m \geq 2$ and $p \geq 2$ and IC $\in \{$False, True$\}$. The expected regret satisfies:*

$$\inf \sup \mathbb{E}[\mathcal{R}_T] \geq c \, \frac{K}{m},$$

where c is a numerical constant, the infinimum is over all playing strategies and the supremum is over all individual sequences.

Remarks. *The lower bound presented in Theorem 5.1 is valid for any $p \leq K$. Algorithms 3 and 4 match it (up to a log factor in K) using only $p = 2$, suggesting that no significant improvements can be obtained if we are allowed to predict using more than two experts.*

Theorem below is of theoretical interest, it shows that if only one feedback is received per round, then constant regrets are not achievable.

Theorem 5.2. *Let ℓ be the squared loss: $\ell(x,y) = (x-y)^2$ on $\mathcal{X} = \mathcal{Y} = [0,1]$. Consider the game protocol presented in Algorithm 1 with $m = 1$ and $p \in [\![K]\!]$ and IC $\in \{$False, True$\}$, we have*

$$\inf \sup \mathbb{E}[\mathcal{R}_T] \geq c \, \sqrt{KT},$$

where c is a numerical constant, the infinimum is over all playing strategies and the supremum is over all individual sequences.

For the sake of completeness, we state the following lower bound from Seldin et al. [2014].

Theorem 5.3 (Direct consequence of Seldin et al., 2014)**.** *Let ℓ be the squared loss: $\ell(x,y) = (x-y)^2$ on $\mathcal{X} = \mathcal{Y} = [0,1]$. Consider the game protocol presented in Algorithm 1 with $p = 1$ and $m \in [\![K]\!]$ and IC $\in \{$False, True$\}$, we have*

$$\inf \sup \mathbb{E}[\mathcal{R}_T] \geq c \sqrt{\frac{K}{m} T},$$

where c is a numerical constant, the infinimum is over all playing strategies and the supremum is over all individual sequences.

6 Discussion and open questions

- In the setting $p = m = 2$ with coupled exploration-exploitation (IC = True), Algorithm 4 presents a strategy with a bound of order $\mathcal{O}(K^2 \log(K\delta^{-1}))$, while the lower bound presented in Theorem 5.1 is of order $\mathcal{O}(K)$. It would be of interest to close this gap.
- Previous works on achieving constant regret under a full observation model only assumed exp-concavity of the loss (see e.g. Cesa-Bianchi and Lugosi, 2006, Chap. 3). In the limited observation setting, we additionally assume that the loss function is bounded by a constant B known to the player. It would be of interest to determine if this condition is necessary. We note, however that loss boundedness is an important ingredient in applying Bernstein-type inequalities for bounds in high probability.
- In the stochastic (i.i.d. experts and target variables) setting, a variation of the expert elimination strategy proposed by Saad and Blanchard [2021] (suitably adapted to tackle cumulative regret) can be shown to have fast rates for regret in an instance-free setting, as well as suitable instance-dependent performance bounds (i.e., the bound depends on the average performance of experts and their correlation, eliminating clearly sub-optimal experts earlier). This a fairly different strategy from the exponential weighting variations proposed here. In the bandit setting, Seldin and Slivkins [2014] have proposed a strategy that reaches almost optimal bounds both in the stochastic and the adversarial settings. It would be interesting to investigate whether such an omnibus strategy exists.
- We have shown that $p = 2$ is sufficient to get constant regret with respect to the best expert, using a strong convexity-type assumption on the loss. For $p = K$, for an exp-concave loss there exist strategies having constant regret with respect to the best convex combination of experts (e.g. Cesa-Bianchi and Lugosi, 2006, Theorem. 3.3), albeit with a $O(K)$ scaling of the regret. It would be interesting to study if "intermediate" situations exist, for example if it is possible to have constant regret with respect to k-combinations of experts using only $p = \mathcal{O}(k)$ expert predictions.

References

Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT*, pages 28–40, 2010.

Jean-Yves Audibert. Progressive mixture rules are deviation suboptimal. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20, 2008.

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *The Journal of Machine Learning Research*, 11:2785–2836, 2010.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331. IEEE, 1995.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Orly Avner, Shie Mannor, and Ohad Shamir. Decoupling exploration and exploitation in multi-armed bandits. *arXiv preprint arXiv:1205.2874*, 2012.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26. JMLR Workshop and Conference Proceedings, 2011.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Nicolo Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162, 2005.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Xiequan Fan, Ion Grama, and Quansheng Liu. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20:1 – 22, January 2015. doi: 10.1214/EJP.v20-3496. URL `https://hal.inria.fr/hal-01108032`.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Pierre Gaillard, Gilles Stoltz, and Tim Van Erven. A second-order bound with excess losses. In *Conference on Learning Theory*, pages 176–196. PMLR, 2014.

Sébastien Gerchinovitz and Tor Lattimore. Refined lower bounds for adversarial bandits. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1198–1206, 2016.

Alon Gonen and Shai Shalev-Shwartz. Tightening the sample complexity of empirical risk minimization via preconditioned stability. *arXiv preprint arXiv:1601.04011*, 2016.

David Helmbold and Sandra Panizza. Some label efficient learning results. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 218–230, 1997.

Sham M Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. In *NIPS*, pages 801–808, 2008.

Jyrki Kivinen and Manfred K Warmuth. Averaging expert predictions. In *European Conference on Computational Learning Theory*, pages 153–167. Springer, 1999.

Tomer Koren and Kfir Levy. Fast rates for exp-concave empirical risk minimization. *Advances in Neural Information Processing Systems*, 28, 2015.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Guillaume Lecué and Shahar Mendelson. Aggregation via empirical risk minimization. *Probability theory and related fields*, 145(3-4):591–613, 2009.

Mehrdad Mahdavi, Lijun Zhang, and Rong Jin. Lower and upper bounds on the generalization of stochastic exponentially concave optimization. In *Conference on Learning Theory*, pages 1305–1320. PMLR, 2015.

Nishant Mehta. Fast rates with high probability in exp-concave statistical learning. In *Artificial Intelligence and Statistics*, pages 1085–1093. PMLR, 2017.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 3168–3176, 2015.

Constantin Niculescu and Lars-Erik Persson. *Convex functions and their applications*, volume 23. Springer, 2006.

El Mehdi Saad and Gilles Blanchard. Fast rates for prediction with limited expert advice. *Advances in Neural Information Processing Systems*, 34, 2021.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, pages 1287–1295. PMLR, 2014.

Yevgeny Seldin, Peter Bartlett, Koby Crammer, and Yasin Abbasi-Yadkori. Prediction with limited advice and multiarmed bandits with paid observations. In *International Conference on Machine Learning*, pages 280–287. PMLR, 2014.

Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *The Journal of Machine Learning Research*, 18(1):1703–1713, 2017.

Tobias Sommer Thune and Yevgeny Seldin. Adaptation to easy data in prediction with limited advice. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 2914–2923, 2018.

Alexander Tsybakov. Optimal rates of aggregation. In *Learning theory and kernel machines*, pages 303–313. Springer, 2003.

Tim van Erven, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16(54):1793–1861, 2015.

Vladimir Vovk. Aggregating strategies. *Proc. of Computational Learning Theory, 1990*, 1990.

Vladimir Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2):153–173, 1998.

Vladimir Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

Yingce Xia, Tao Qin, Weidong Ma, Nenghai Yu, and Tie-Yan Liu. Budgeted multi-armed bandits with multiple plays. In *IJCAI*, pages 2210–2216, 2016.

Datong Zhou and Claire Tomlin. Budget-constrained multi-armed bandits with multiple plays. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Appendix: detailed proofs

A Notation

The following notation pertains to all the considered algorithms, where t is a given training round and T is the game horizon:

- For any $x > 0$, let $\log_2^+(x) = \max\{0, \log_2(x)\}$.
- Let $\mathcal{R}_T$ denote the cumulative random regret of the player over T rounds.
- Let S_t denote the set of combined experts to make a prediction at round t.
- Let C_t denote the set of observed experts after making the prediction at round t.
- For each $i \in S_t$, let $\alpha_{i,t}$ denote the weight of expert i in the convex combination played in round t.
- Let $(\mathcal{F}_t)_t$ denote the natural filtration associated with the process $(S_t, C_t, (\ell_{i,t})_{i \in C_t})_t$.
- Denote the conditional expectation with respect to $\mathcal{F}_t$ by $\mathbb{E}_t[.] = \mathbb{E}[.|\mathcal{F}_t]$.
- For each expert $i \in [\![K]\!]$, let N_i denote the number of times the prediction of expert i was observed during the game (over T rounds).
- For each expert $i \in [\![K]\!]$, let M_i denote the number of times the prediction of expert i was used for prediction during the game (over T rounds): $M_i := |\{t \in [\![T]\!] : \quad i \in S_t\}|$.
- For each expert $i \in [\![K]\!]$, we define $\ell_{i,t} = \ell(F_{i,t}, y_t)$.
- Denote by $\ell_t : \mathcal{X} \to \mathbb{R}$ such that $\forall x \in [\![X]\!] : \ell_t(x) = \ell(x, y_t)$.

Notation associated to Algorithms 3 and 4

- Let I_t and J_t denote the experts used for prediction in round t.
- Let $\mathcal{U}_t$ the set of experts queried for exploration (sampled uniformly without replacement from $[\![K]\!]$). In Algorithm 4 let $\mathcal{U}_t = \{B_t\}$.
- Let $\tilde{m} = \max\{1, m-2\}$.

B Some preliminary technical results

The following device is standard (it is used for instance for proving Bennett's inequality).

Lemma B.1. *Let X be a random variable with finite variance, such that $X \leq b$ almost surely for some $b > 0$. For any $\lambda > 0$:*

$$\log\Big(\mathbb{E}e^{\lambda X}\Big) \leq \lambda \mathbb{E}[X] + \frac{\phi(\lambda b)}{b^2}\mathbb{E}[X^2].$$

Where $\phi(x) = \exp(x) - 1 - x$.

Proof. The function $x \mapsto x^{-2}\phi(x)$ is non-decreasing on $\mathbb{R}$. As a consequence, if $X \leq b$ a.s., for any $\lambda > 0$ it holds $\exp(\lambda X) \leq \frac{\phi(\lambda b)}{b^2} X^2 + 1 + \lambda X$, a.s. Taking the expectation, then applying the inequality $\log(1+t) \leq t$ yields the result. □

Corollary B.2. *Let X be a random variable with finite variance, such that $X \geq -b$ almost surely for $b > 0$. For any $\lambda \in \left(0, \frac{1}{b}\right)$:*

$$\log\left(\mathbb{E}e^{-\lambda X}\right) \leq -\lambda \mathbb{E}[X] + \lambda^2 \mathbb{E}[X^2].$$

Proof. This corollary is a direct consequence of applying Lemma B.1 to the variable $-X \leq b$, then using the fact that $\forall x \leq 1 : \phi(x) \leq x^2$. □

We now introduce some technical lemmas used in the proofs. Let us start by reminding the following standard result (see Theorem 1.1.4 Niculescu and Persson, 2006).

Lemma B.3. *A continuous function $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is a convex set, is convex if and only if: for any $x, x' \in \mathcal{X}$:*

$$f\left(\frac{x+x'}{2}\right) \leq \frac{1}{2}f(x) + \frac{1}{2}f(x').$$

Lemmas below give some bounds for some functions.

Lemma B.4.

- *We have for any $x \in \mathbb{R}$*

$$1 + \frac{x^2}{2} \leq \cosh(x) \leq \exp(x^2/2).$$

- *Let $c > 0$. We have for any $x \in [0, c]$*

$$\log(1+x) \geq \frac{\log(1+c)}{c} x.$$

Proof. The first and third result is a direct consequence of Taylor's expansion. The second result follows simply by concavity of $x \to \log(1+x)$. □

Lemma B.5. *We have for any $x, y > 0$*

$$\log_2^+(x) - \frac{x}{y} \leq \log_2^+(y).$$

Proof. Let $x, y > 0$, we have

$$\begin{aligned} \log_2(y) &= \log_2(x) - \log_2\left(\frac{x}{y}\right) \\ &\geq \log_2(x) - \frac{x}{y}, \end{aligned}$$

where we used the fact that $\log_2(t) \leq t$ for any $t > 0$. To conclude we use the inequality

$$(a)_+ - b \leq (a-b)_+,$$

valid for any $a \in \mathbb{R}$ and $b > 0$. □

C Proof of Lemma 1.1

Let $y \in \mathcal{Y}$. In this proof, we will denote $\ell(.)$ instead of $\ell(., y)$ so as to ease notation.

C.1 First claim

By exp-concavity of ℓ, we have for any $x, x' \in \mathcal{X}$

$$\frac{1}{2}\exp\{-\eta\ell(x)\} + \frac{1}{2}\exp\{-\eta\ell(x')\} \leq \exp\Big\{-\eta\ell\Big(\frac{x+x'}{2}\Big)\Big\}.$$

Multiplying both sides by $\exp\Big\{\frac{1}{2}\eta\ell(x) + \frac{1}{2}\eta\ell(x')\Big\}$, we have

$$1 + \frac{\eta^2(\ell(x) - \ell(x'))^2}{2} \leq \exp\Big\{\frac{\eta}{2}\ell(x) + \frac{\eta}{2}\ell(x') - \eta\ell\Big(\frac{x+x'}{2}\Big)\Big\},$$

where we used the first result of Lemma B.4 to lower bound the left hand side.

Introducing the logarithm and using the second result of Lemma B.4, we obtain

$$\frac{2\log\Big(1 + \frac{\eta^2 B^2}{2}\Big)}{\eta^2 B^2}\eta^2(\ell(x) - \ell(x'))^2 \leq \frac{\eta}{2}\ell(x) + \frac{\eta}{2}\ell(x') - \eta\ell\Big(\frac{x+x'}{2}\Big).$$

We conclude that

$$\ell\Big(\frac{x+x'}{2}\Big) \leq \frac{1}{2}\ell(x) + \frac{1}{2}\ell(x') - \frac{1}{2c}(\ell(x) - \ell(x'))^2,$$

where

$$c = \frac{\eta B^2}{4\log\Big(1 + \frac{\eta^2 B^2}{2}\Big)}.$$

C.2 Second claim

Let $c > 0$, we denote $\mathcal{E}(c)$ the set of functions $f : \mathcal{X} \to \mathbb{R}$, such that for any $x, x' \in \mathcal{X}$:

$$f\Big(\frac{x+x'}{2}\Big) \leq \frac{1}{2}f(x) + \frac{1}{2}f(x') - \frac{1}{2c}(f(x) - f(x'))^2. \tag{8}$$

Lemma C.1. *For any $c > 0$, we have for any $f \in \mathcal{E}(c)$*

$$\sup_{x,x' \in \mathcal{X}} |f(x) - f(x')| \leq c.$$

Proof. Put $\Delta_{xx'} = f(x') - f(x)$, and $\Delta^* = \sup_{x,x' \in \mathcal{X}} \Delta_{xx'}$. We first prove that $\Delta^* \leq 3c$. Assume this is not the case and let $x, x' \in \mathcal{X}$ be such that $\Delta_{xx'} > 3c$. Let $z := \frac{1}{2}(x + x')$. Using $f \in \mathcal{E}(c)$, we obtain

$$\Delta_{xz} = f(z) - f(x) \leq \frac{1}{2}(f(x') - f(x)) - \frac{1}{2c}(f(x') - f(x))^2 = \frac{1}{2}\Delta_{xx'} - \frac{1}{2c}\Delta_{xx'}^2 \leq -\Delta_{xx'},$$

where the last inequality holds because $\Delta_{xx'} > 3c$. Hence $\Delta_{zx} > 3c$ and in turn, if $x_1 := \frac{1}{2}(x+z)$, reiterating the above argument we get $\Delta_{x_1 z} > 3c$ and in particular $f(x_1) < f(z)$. Also, we have $\Delta_{zx'} = \Delta_{zx} + \Delta_{xx'} > 3c$, therefore putting $x_1' := \frac{1}{2}(x'+z)$, again by the same token we get $f(x_1') < f(z)$. This is a contradiction, since $z = \frac{1}{2}(x_1 + x_1')$, thus Assumption 1 implies that $f(z) \leq \max(f(x_1), f(x_1'))$.

Since Δ^* is finite, $m := \inf_{x \in X} f(x)$ is finite. For any $\varepsilon > 0$, let x_ε be such that $f(x_\varepsilon) \leq m + \varepsilon$. For any $x' \in X$, putting again $z := \frac{1}{2}(x + x')$, it must be the case that $\Delta_{x_\varepsilon z} \geq -\varepsilon$, and using again the above display it must hold $-\varepsilon \leq \Delta_{x_\varepsilon z} \leq \frac{1}{2}\Delta_{x_\varepsilon x'} - \frac{1}{2c}\Delta^2_{x_\varepsilon x'}$. This implies $\Delta_{x_\varepsilon x'} \leq c + G(\varepsilon)$ for any $x' \in \mathcal{X}$, with $G(\varepsilon) = O(\varepsilon)$. Since $\Delta^* \leq \varepsilon + \sup_{x' \in \mathcal{X}} \Delta_{x_\varepsilon x'}$, we conclude to $\Delta^* \leq c$ by letting $\varepsilon \to 0$. □

Lemma C.2. *For any $c > 0$, we have for any continuous function $f \in \mathcal{E}(c)$: f is $(4/c)$-exp-concave.*

Proof. Fix $c > 0$ and $f \in \mathcal{E}(c)$. Let $x, x' \in \mathcal{X}$. Let us prove that

$$\frac{1}{2}\exp\left\{-\frac{4}{c}f(x)\right\} + \frac{1}{2}\exp\left\{-\frac{4}{c}f(x')\right\} \leq \exp\left\{-\frac{4}{c}f\left(\frac{x+x'}{2}\right)\right\}. \tag{9}$$

Recall that since $f \in \mathcal{E}(c)$, inequality (8) gives

$$\frac{2}{c^2}\big(f(x) - f(x')\big)^2 \leq \frac{2}{c}f(x) + \frac{2}{c}f(x') - \frac{4}{c}f\left(\frac{x+x'}{2}\right).$$

We introduce the exp function on both sides of the inequality and use the first result of Lemma B.4 to lower bound the left hand side. We have

$$\frac{1}{2}\exp\left\{\frac{2}{c}\big(f(x) - f(x')\big)\right\} + \frac{1}{2}\exp\left\{\frac{2}{c}\big(f(x') - f(x)\big)\right\} \leq \exp\left\{\frac{2}{c}f(x) + \frac{2}{c}f(x')\right\} \exp\left\{-\frac{4}{c}f\left(\frac{x+x'}{2}\right)\right\},$$

which proves (9). We conclude using the characterization provided by Lemma B.3. □

C.3 Third claim

Lemma C.3. *Let $f : \mathcal{X} \to \mathbb{R}$ be a L-Lipschitz and ρ-strongly convex function, then $f \in \mathcal{E}(4L^2/\rho)$.*

Proof. By strong convexity of f, we have for any $x, x' \in \mathcal{X}$

$$f\left(\frac{x+x'}{2}\right) \leq \frac{1}{2}f(x) + \frac{1}{2}f(x') - \frac{\rho}{8}\|x - x'\|^2.$$

Moreover, $f(.)$ is L-Lipschitz, hence: $|f(x) - f(x')| \leq L\|x - x'\|$. Therefore

$$f\left(\frac{x+x'}{2}\right) \leq \frac{1}{2}f(x) + \frac{1}{2}f(x') - \frac{\rho}{8L^2}\big(f(x) - f(x')\big)^2.$$

□

D Concentration inequality for martingales

We recall Bennett's inequality:

Theorem D.1. *Let $Z, Z_1, \dots, Z_n$ be i.i.d random variables with values in $[-B, B]$ and let $\delta > 0$. Then with probability at least $1 - \delta$ in $(Z_1, \dots, Z_n)$ we have*

$$\left| \mathbb{E}[Z] - \frac{1}{n} \sum_{i=1}^{n} Z_i \right| \leq \sqrt{\frac{2 \operatorname{Var}[Z] \log(2/\delta)}{n}} + \frac{2B \log(2/\delta)}{3n}.$$

We recall Freedman's inequality (the exposition here is lifted from Fan et al., 2015). Let $(\xi_i, \mathcal{F}_i)_{i \geq 1}$ be a (super)martingale difference sequence. Define $S_n := \sum_{i=1}^{n} \xi_i$ (then $(S_n, \mathcal{F}_n)$ is a (super)martingale), and $\langle S \rangle_n := \sum_{i=1}^{n} \mathbb{E}[\xi_i^2 | \mathcal{F}_{i-1}]$ the quadratic characteristic of S.

Theorem D.2 (Freedman's inequality)**.** *Assume $\xi_i \leq B$ for all $i \geq 1$, where B is a constant. Then for all $t, v > 0$:*

$$\mathbb{P}\Big[S_k \geq t \text{ and } \langle S \rangle_k \leq v^2 \text{ for some } k \geq 1 \Big] \leq \exp\left(-\frac{t^2}{2(v^2 + Bt)} \right). \tag{10}$$

The following direct consequence also appears in [Kakade and Tewari, 2008, Lemma 3] for fixed k. Here we give a version that holds uniformly in k. See also [Gaillard et al., 2014, Theorem 12] for a related result.

Corollary D.3. *Assume $\xi_i \leq B$ for all $i \geq 1$, where B is a constant. Then for all $\delta \in (0, 1/3)$, with probability at least $1 - 3\delta$ it holds*

$$\forall k \geq 1 : S_k \leq 2\sqrt{\langle S \rangle_k \varepsilon(\delta, k)} + 4B\varepsilon(\delta, k),$$

where $\varepsilon(\delta, k) := \log \delta^{-1} + 2 \log(1 + \log_2^+(\langle S \rangle_k / B^2))$.

If $|\xi_i| \leq B$ for all $i \geq 1$, observe that $\varepsilon(\delta, k) \leq \log \delta^{-1} + O(\log \log k)$.

Proof. By standard calculations, it holds that if $t \geq v\sqrt{2 \log \delta^{-1}} + 2B \log \delta^{-1}$, then $\frac{t^2}{2(v^2 + Bt)} \geq \log \delta^{-1}$. Therefore (10) implies that for any $v > 0$ and $\delta \in (0, 1)$, it holds

$$\mathbb{P}\left[\exists k \geq 1 : S_k \geq \sqrt{2v^2 \log \delta^{-1}} + 2B \log \delta^{-1} \text{ and } \langle S \rangle_k \leq v^2 \right] \leq \delta. \tag{11}$$

Denote $v_j^2 := 2^j B^2$, $\delta_j := (j \vee 1)^{-2} \delta$, $j \geq 0$, and define the non-decreasing sequence of stopping times $\tau_{-1} = 1$ and $\tau_j := \min\{k \geq 1 : \langle S \rangle_k > v_j^2\}$ for $j \geq 0$. Define the events for $j \geq 0$:

$$A_j := \left\{ \exists k \geq 1 : S_k \geq \sqrt{2v_j^2 \log \delta_j^{-1}} + 2B \log \delta_j^{-1} \text{ and } \langle S \rangle_k \leq v_j^2 \right\},$$
$$A_j' := \left\{ \exists k \text{ with } \tau_{j-1} \leq k < \tau_j : S_k \geq 2\sqrt{\langle S \rangle_k \varepsilon(\delta, k)} + 4B\varepsilon(\delta, k) \right\}.$$

From the definition of v_j^2, δ_j, we have $j = \log_2(v_j^2 / B^2)$ for $j \geq 1$. For $j \geq 1$, $\tau_{j-1} \leq k < \tau_j$ implies $v_{j-1}^2 = v_j^2/2 < \langle S \rangle_k \leq v_j^2$, and further

$$\log \delta_j^{-1} = \log \delta^{-1} + 2 \log \log_2(v_j^2 / B^2) \leq \varepsilon(\delta, k).$$

Therefore it holds $A'_j \subseteq A_j$. Furthermore, for $j = 0$, we have $v_0^2 = B^2, \delta_0 = \delta$. Further, if $k < \tau_0$ it implies $\langle S \rangle_k < B^2$ and therefore $\varepsilon(\delta, k) = \log \delta^{-1}$. Thus, provided $\log \delta^{-1} \geq 1$ i.e. $\delta \leq 1/e$, it holds

$$A'_0 \subseteq \Big\{ \exists k \text{ with } k < \tau_0 : S_k \geq 4B \log \delta_0^{-1} \Big\}$$
$$\subseteq \Big\{ \exists k \geq 1 : S_k \geq \sqrt{2v_0^2 \log \delta_0^{-1}} + 2B \log \delta_0^{-1} \text{ and } \langle S \rangle_k \leq v_0^2 \Big\} = A_0.$$

Therefore, since by (11) it holds $\mathbb{P}[A_j] \leq \delta_j$ for all $j \geq 0$:

$$\mathbb{P}\Big[\exists k \leq n : S_k \geq 2\sqrt{\langle S \rangle_k \varepsilon(\delta, k)} + 4B\varepsilon(\delta, k) \Big] = \mathbb{P}\Big[\bigcup_{j \geq 0} A'_j \Big] \leq \mathbb{P}\Big[\bigcup_{j \geq 0} A_j \Big] \leq \delta \sum_{j \geq 0} (j \vee 1)^{-2} \leq 3\delta.$$

□

Corollary D.4. *Assume $\xi_i \leq b$ for all $i \geq 1$, where b is a constant. Let $(\nu_t)_t$ denote an $\mathcal{F}_t$-measurable sequence, such that for any $k \geq 1$: $\langle S \rangle_k \leq \sum_{i=1}^k \nu_i$. Then for all $c > 0$ and $\delta \in (0, 1/3)$, with probability at least $1 - 3\delta$ it holds*

$$\forall k \geq 1 : S_k - \frac{c}{b} \sum_{i=1}^k \nu_k \leq \Big(\frac{8}{c} + 4 \Big) \Big(\log(\delta^{-1}) + 2 \log_2^+ \Big(\frac{32 + 16c}{c^2} \Big) \Big) b.$$

Proof. Let $c > 0$ and fix $\delta \in (0, 1/3)$, we have using Corollary D.3: with probability at least $1 - 3\delta$, it holds for any $k \geq 1$

$$\begin{aligned}
S_k - \frac{c}{b} \sum_{i=1}^k \nu_i &\leq 2\sqrt{\langle S \rangle_k \epsilon(\delta, k)} + 4b\epsilon(\delta, k) - \frac{c}{b} \sum_{i=1}^k \nu_i \\
&\leq 2\sqrt{\langle S \rangle_k \epsilon(\delta, k)} + 4b\epsilon(\delta, k) - \frac{c}{b} \langle S \rangle_k \\
&\leq 2\Big(\frac{c}{4b} \langle S \rangle_k + \frac{4b}{c} \epsilon(\delta, k) \Big) + 4b\epsilon(\delta, k) - \frac{c}{b} \langle S \rangle_k \\
&\leq \Big(\frac{8}{c} + 4 \Big) b\epsilon(\delta, k) - \frac{c}{2b} \langle S \rangle_k \\
&= \Big(\frac{8}{c} + 4 \Big) b \Big(\log \delta^{-1} + 2 \log \Big(1 + \log_2^+ (\langle S \rangle_k / b^2) \Big) \Big) - \frac{c}{2b} \langle S \rangle_k \\
&\leq \Big(\frac{8}{c} + 4 \Big) b \Big(\log \delta^{-1} + 2 \log_2^+ (\langle S \rangle_k / b^2) \Big) - \frac{c}{2b} \langle S \rangle_k
\end{aligned}$$

The result follows by upper-bounding the function $x \to \log_2^+(x) - x/y$, for $x, y > 0$ using Lemma B.5. □

E Additional technical results

The following lemma is a consequence of Corollary B.2, the chaining rule (i.e cancellation in the sum of logarithmic terms) and Fubini's theorem. Let $(\hat{h}_{i,t})_{t \in [\![T]\!], i \in [\![K]\!]}$ be a $\mathcal{F}_t$-adapted process.

For each $i \in [\![K]\!]$ and $t \in [\![T]\!]$ we define: $\hat{H}_{i,t} := \sum_{i=1}^{t} \hat{h}_{i,s}$, we use the convention that $\hat{H}_{i,0} = 0$. Let $t \in [\![T]\!]$ and $\lambda > 0$, we define the sequence $(\hat{p}_{i,t})_{i \in [\![K]\!]}$:

$$\hat{p}_{i,t} := \frac{\exp\Big\{-\lambda \hat{H}_{i,t-1}\Big\}}{\sum_{j=1}^{K} \exp\Big\{-\lambda \hat{H}_{j,t-1}\Big\}}. \tag{12}$$

For each $t \in [\![T]\!]$, define:

$$\hat{Z}_t := \sum_{i=1}^{K} \exp\{-\lambda \hat{H}_{i,t}\} \tag{13}$$

$$M_t := \log\Big(\hat{Z}_t\Big) - \mathbb{E}_{t-1}\Big[\log(\hat{Z}_t)\Big]. \tag{14}$$

Lemma E.1. *Let $b > 0$ and $(\hat{h}_{i,t})_{t \in [\![T]\!], i \in [\![K]\!]}$ be a sequence of numbers taking values in an interval of length b. For each $i \in [\![K]\!]$ and $t \in [\![T]\!]$, let $\mathbb{E}_{t-1}[\hat{h}_{i,t}] = h_{i,t}$. Let $(\alpha_t)_{t \in [\![T]\!]}$ be a sequence such that α_t is $\mathcal{F}_{t-1}$-measurable and:*

$$\forall i \in [\![K]\!], t \in [\![T]\!], \ \Big|\hat{h}_{i,t} - \alpha_t\Big| \leq b.$$

Then for any $\lambda \in (0, 1/b)$, for all $t \in [\![T]\!]$ we have:

$$\sum_{t=1}^{T} \sum_{i=1}^{K} \hat{p}_{i,t} \ h_{i,t} \leq \min_{i \in [\![K]\!]} \sum_{t=1}^{T} \hat{h}_{i,t} + \frac{\log(K)}{\lambda} + \frac{1}{\lambda} \sum_{t=1}^{T-1} M_t + \lambda \sum_{t=1}^{T} \sum_{i=1}^{K} \hat{p}_{i,t} \ \mathbb{E}_{t-1}\Big[\Big(\hat{h}_{i,t} - \alpha_t\Big)^2\Big],$$

where the sequence $(\hat{p}_{i,t})_{t \in [\![T]\!], i \in [\![K]\!]}$ is defined by (12) *and (M_t) is defined by* (14)*.*

Proof. Let $t \in [\![T]\!]$, we denote by $\hat{p}_t$ the probability distribution on $[\![K]\!]$ defined by the weights $(\hat{p}_{i,t})_{i \in [\![K]\!]}$. We apply Corollary B.2 to the random variable $X_t := \hat{h}_{I,t} - \alpha_t$, where I is drawn from $[\![K]\!]$ following $\hat{p}_t$: for any $\lambda \in (0, 1/b)$,

$$\log\left(\sum_{i=1}^{K} \hat{p}_{i,t} \exp\Big\{-\lambda\Big(\hat{h}_{i,t} - \alpha_t\Big)\Big\}\right) \leq -\lambda \sum_{i=1}^{K} \hat{p}_{i,t}\Big(\hat{h}_{i,t} - \alpha_t\Big) + \lambda^2 \sum_{i=1}^{K} \hat{p}_{i,t}\Big(\hat{h}_{i,t} - \alpha_t\Big)^2.$$

Rearranging terms we obtain:

$$\begin{aligned}
\sum_{i=1}^{K} \hat{p}_{i,t} \ \hat{h}_{i,t} &\leq \alpha_t - \frac{1}{\lambda} \log\left(\left(\sum_{i=1}^{K} \hat{p}_{i,t} \exp\{-\lambda \hat{h}_{i,t}\}\right) \exp\{\lambda \alpha_t\}\right) + \lambda \sum_{i=1}^{K} \hat{p}_{i,t}\Big(\hat{h}_{i,t} - \alpha_t\Big)^2 \\
&= -\frac{1}{\lambda} \log\left(\sum_{i=1}^{K} \hat{p}_{i,t} \exp\{-\lambda \hat{h}_{i,t}\}\right) + \lambda \sum_{i=1}^{K} \hat{p}_{i,t}\Big(\hat{h}_{i,t} - \alpha_t\Big)^2 \\
&= -\frac{1}{\lambda}\Big(\log\Big(\hat{Z}_t\Big) - \log\Big(\hat{Z}_{t-1}\Big)\Big) + \lambda \sum_{i=1}^{K} \hat{p}_{i,t}\Big(\hat{h}_{i,t} - \alpha_t\Big)^2,
\end{aligned}$$

where $\hat{Z}_t$ is defined by (13). Taking the conditional expectation with respect to $\mathcal{F}_{t-1}$ gives

$$\sum_{i=1}^{K} \hat{p}_{i,t} h_{i,t} \leq -\frac{1}{\lambda}\Big(\mathbb{E}_{t-1}\Big[\log\Big(\hat{Z}_t\Big)\Big] - \log\Big(\hat{Z}_{t-1}\Big)\Big) + \lambda \sum_{i=1}^{K} \hat{p}_{i,t} \mathbb{E}_{t-1}\Big[\Big(\hat{h}_{i,t} - \alpha_t\Big)^2\Big].$$

Summing over $t \in [\![T]\!]$ we obtain:

$$\sum_{t=1}^{T}\sum_{i=1}^{K} \hat{p}_{i,t}\ h_{i,t} \leq \frac{\log(Z_0)}{\lambda} - \frac{\log\Big(\hat{Z}_T\Big)}{\lambda} + \frac{1}{\lambda}\sum_{t=1}^{T-1} M_t + \lambda \sum_{t=1}^{T}\sum_{i=1}^{K} \hat{p}_{i,t}\ \mathbb{E}_{t-1}\Big[\Big(\hat{h}_{i,t} - \alpha_t\Big)^2\Big].$$

Finally observe that $Z_0 = K$ and that:

$$\begin{aligned} -\frac{1}{\lambda}\log\Big(\hat{Z}_T\Big) &= -\frac{1}{\lambda}\log\left(\sum_i \exp\{-\lambda \hat{H}_{i,t}\}\right) \\ &\leq \min_{i \in [\![K]\!]} \hat{H}_{i,t}. \end{aligned}$$

□

F A preliminary result for the proof of Theorem 4.1 and 4.2

In this section we present two key results for the proof of Theorem 4.1 and 4.2. Lemma F.5 provides a bound for the cases $(p = 2, m \geq 3)$ and $(p = 2, m = 2, \mathrm{IC} = \mathrm{False})$. Lemma F.6 presents a similar bound for the particular case $(p = 2, m = 2, \mathrm{IC} = \mathrm{True})$. We decided to separate these two settings because each one requires a different condition on λ.

We consider the notation of Algorithms 3 and 4. In Algorithm 3 ($m \geq 3$), we take $A_t = I_t$. Recall that $\tilde{m} = \max\{1, m-2\}$ (as defined in Section A).

Lemma F.1. *For any $k \geq 1$,*

$$\mathbb{E}_{t-1}\Big[\Big(\hat{\ell}_{i,t} - \ell_{A_t,t}\Big)^k\Big] = \left(\frac{K}{\tilde{m}}\right)^{k-1} \mathbb{E}_{t-1}\Big[(\ell_{i,t} - \ell_{A_t,t})^k\Big],$$

where $\tilde{m} = \max\{1, m-2\}$.

Proof. Suppose that $m \geq 3$. Consider the notation of Algorithm 3. Let $k \geq 1$, we have

$$\begin{aligned} \mathbb{E}_{t-1}\Big[\Big(\hat{\ell}_{i,t} - \ell_{A_t,t}\Big)^k\Big] &= \mathbb{E}_{t-1}\left[\left(\frac{K}{m-2}\mathbb{1}(i \in \mathcal{U}_t)\ell_{i,t} + \left(1 - \frac{K}{m-2}\mathbb{1}(i \in \mathcal{U}_t)\right)\ell_{A_t,t} - \ell_{A_t,t}\right)^k\right] \\ &= \mathbb{E}_{t-1}\left[\left(\frac{K}{m-2}\mathbb{1}(i \in \mathcal{U}_t)\ell_{i,t} - \frac{K}{m-2}\mathbb{1}(i \in \mathcal{U}_t)\ell_{A_t,t}\right)^k\right] \\ &= \left(\frac{K}{m-2}\right)^k \mathbb{E}_{t-1}[\mathbb{1}(i \in \mathcal{U}_t)](\ell_{i,t} - \ell_{A_t,t})^k \\ &= \left(\frac{K}{m-2}\right)^{k-1} \mathbb{E}_{t-1}\Big[(\ell_{i,t} - \ell_{A_t,t})^k\Big], \end{aligned}$$

where we used the fact that U_t and A_t are independent conditionally to $\mathcal{F}_{t-1}$.

Suppose that $m = 2$. Consider the notation of Algorithm 4. Let $k \geq 1$, we have

$$\begin{aligned}
\mathbb{E}_{t-1}\Big[\Big(\hat{\ell}_{i,t} - \ell_{A_t,t}\Big)^k\Big] &= \mathbb{E}_{t-1}\Big[\Big(\hat{\ell}_{i,t} - \ell_{A_t,t}\Big)^k\Big] \\
&= \mathbb{E}_{t-1}\Big[(K\mathbb{1}(B_t = i)\ell_{i,t} + (1 - K\mathbb{1}(B_t = i))\ell_{A_t,t} - \ell_{A_t,t})^k\Big] \\
&= K^k\mathbb{E}_{t-1}\Big[\mathbb{1}(B_t = i)(\ell_{i,t} - \ell_{A_t,t})^k\Big] \\
&= K^{k-1}\mathbb{E}_{t-1}\Big[(\ell_{i,t} - \ell_{A_t,t})^k\Big].
\end{aligned}$$

□

Introduce the notation

$$\hat{\mu}_t := \sum_{i \in [\![K]\!]} \hat{p}_{i,t}\ell_{i,t}, \tag{15}$$

$$\hat{\xi}_t := \frac{1}{2}\sum_{i,j \in [\![K]\!]} \hat{p}_{i,t}\hat{p}_{j,t}\ (\ell_{i,t} - \ell_{j,t})^2, \tag{16}$$

where $(\hat{p}_{i,t})$ is defined in (12). For each $t \in [\![T]\!]$, let

$$\begin{aligned}
\hat{Z}_t &= \sum_{i=1}^K \exp\Big\{-\lambda\hat{L}_{i,t} + \lambda^2\hat{V}_{i,t}\Big\} \\
M_t &= \log\Big(\hat{Z}_t\Big) - \mathbb{E}_{t-1}\Big[\hat{Z}_t\Big],
\end{aligned} \tag{17}$$

where $\hat{L}_{i,t} = \sum_{s=1}^t \hat{\ell}_{i,t}$ and $\hat{V}_{i,t} = \sum_{s=1}^t \hat{v}_{i,t}$, in agreement with the notation used in Algorithms 3 and 4, and in Section E.

Lemma F.2. *Let $\lambda \in \Big(0, \frac{2\tilde{m}}{K}\bar{\lambda}\Big)$, where $\bar{\lambda}$ is defined in (2) and $\tilde{m} = \max\{m-2, 1\}$. For each $i \in [\![K]\!]$, $t \in [\![T]\!]$, let $\hat{h}_{i,t} = \hat{\ell}_{i,t} - \lambda\hat{v}_{i,t}$. We have*

$$\sum_{t=1}^T \hat{\mu}_t \leq \min_{i \in [\![K]\!]} \sum_{t=1}^T \hat{h}_{i,t} + \frac{1}{\lambda}\sum_{t=1}^T M_t + \frac{\log(K)}{\lambda} + \frac{11\lambda K}{\tilde{m}}\sum_{t=1}^T \hat{\xi}_t,$$

where $\hat{\mu}_t$ is defined in (15), $\hat{\xi}_t$ is defined in (16) and M_t is defined in (17).

Proof. Let $h_{i,t} := \mathbb{E}_{t-1}[\hat{h}_{i,t}] = \ell_{i,t} - \lambda\mathbb{E}_{t-1}[\hat{v}_{i,t}]$, we apply Lemma E.1 to the sequence $(\hat{h}_{i,t})_{i,t}$. We take $\alpha_t = \hat{\mu}_t$, which is an $\mathcal{F}_{t-1}$-measurable process. For each $i \in [\![K]\!]$ and $t \geq 0$, we have

$$\sum_{t=1}^T\sum_{i=1}^K \hat{p}_{i,t}h_{i,t} \leq \min_{i \in [\![K]\!]} \sum_{t=1}^T \hat{h}_{i,t} + \frac{\log(K)}{\lambda} + \frac{1}{\lambda}\sum_{t=1}^T M_t + \lambda\sum_{t=1}^T\sum_{i=1}^K \hat{p}_{i,t}\mathbb{E}_{t-1}\Big[\Big(\hat{h}_{i,t} - \hat{\mu}_t\Big)^2\Big]. \tag{18}$$

Now, let us develop a lower bound on the left hand side of the inequality above. Recall that in Algorithm 3, we take $A_t = I_t$, then $A_t \sim \hat{p}_t$. In Algorithm 4, Lemma G.1 shows that $A_t \sim \hat{p}_t$.

Fix $t \in [\![T]\!]$, we have:

$$\begin{aligned}
\sum_{i=1}^{K} \hat{p}_{i,t} h_{i,t} &= \sum_{i=1}^{K} \hat{p}_{i,t}(\ell_{i,t} - \lambda \mathbb{E}_{t-1}[\hat{v}_{i,t}]) \\
&= \sum_{i=1}^{K} \hat{p}_{i,t}\ell_{i,t} - \lambda \sum_{i=1}^{K} \hat{p}_{i,t} \mathbb{E}_{t-1}\Big[\Big(\hat{\ell}_{i,t} - \ell_{A_t,t}\Big)^2\Big] \\
&= \sum_{i=1}^{K} \hat{p}_{i,t}\ell_{i,t} - \lambda \frac{K}{\tilde{m}} \left(\sum_{i=1}^{K} \hat{p}_{i,t}(\ell_{i,t} - \hat{\mu}_t)^2 \right) - \lambda \frac{K}{\tilde{m}} \mathbb{E}_{t-1}\Big[(\ell_{A_t,t} - \hat{\mu}_t)^2\Big] \\
&= \hat{\mu}_t - 2\lambda \frac{K}{\tilde{m}} \hat{\xi}_t, \qquad (19)
\end{aligned}$$

where we used in the second line the definition $\hat{v}_{i,t} = \Big(\hat{\ell}_{i,t} - \ell_{A_t,t}\Big)^2$, Lemma F.1 with $k = 2$ in the third line, and the fact that A_t is distributed following $\hat{p}$ in the third and fourth line.

Next, we develop an upper bound on the last term of the right hand side of (18). We have

$$\sum_{t=1}^{T} \sum_{i=1}^{K} \hat{p}_{i,t}\ \mathbb{E}_{t-1}\Big[\Big(\hat{h}_{i,t} - \hat{\mu}_t\Big)^2\Big] \leq 2 \sum_{t=1}^{T} \sum_{i=1}^{K} \hat{p}_{i,t} \Big\{ \mathbb{E}_{t-1}\Big[\Big(\hat{\ell}_{i,t} - \hat{\mu}_t\Big)^2\Big] + \lambda^2 \mathbb{E}_{t-1}\Big[\hat{v}_{i,t}^2\Big] \Big\}. \qquad (20)$$

Fix $t \in [\![T]\!]$. Let us bound each of the terms in the right hand side of the inequality above

$$\begin{aligned}
\sum_{i=1}^{K} \hat{p}_{i,t} \mathbb{E}_{t-1}\Big[\Big(\hat{\ell}_{i,t} - \hat{\mu}_t\Big)^2\Big] &\leq \sum_{i=1}^{K} 2\hat{p}_{i,t} \Big(\mathbb{E}_{t-1}\Big[\Big(\hat{\ell}_{i,t} - \ell_{A_t,t}\Big)^2\Big] + \mathbb{E}_{t-1}\Big[(\ell_{A_t,t} - \hat{\mu}_t)^2\Big] \Big) \\
&= 2\mathbb{E}_{t-1}\Big[(\ell_{A_t,t} - \hat{\mu}_t)^2\Big] + 2\frac{K}{\tilde{m}} \sum_{i=1}^{K} \hat{p}_{i,t} \mathbb{E}_{t-1}\Big[(\ell_{i,t} - \ell_{A_t,t})^2\Big] \\
&= 2\hat{\xi}_t + 2\frac{K}{\tilde{m}} \sum_{i=1}^{K} \hat{p}_{i,t} \Big\{ (\ell_{i,t} - \hat{\mu}_t)^2 + \mathbb{E}_{t-1}\Big[(\ell_{A_t,t} - \hat{\mu}_t)^2\Big] \Big\} \\
&\leq \frac{6K}{\tilde{m}} \hat{\xi}_t, \qquad (21)
\end{aligned}$$

where we used Lemma F.1 for the second line. Moreover, using the same Lemma F.1 with $k = 4$, we have

$$\begin{aligned}
\sum_{i=1}^{K} \hat{p}_{i,t} \mathbb{E}_{t-1}\Big[\hat{v}_{i,t}^2\Big] &= \sum_{i=1}^{K} \hat{p}_{i,t} \left(\frac{K}{\tilde{m}}\right)^3 \mathbb{E}_{t-1}\Big[(\ell_{i,t} - \ell_{A_t,t})^4\Big] \\
&\leq \left(\frac{K}{\tilde{m}}\right)^3 B^2 \sum_{i=1}^{K} \hat{p}_{i,t} \mathbb{E}_{t-1}\Big[(\ell_{i,t} - \ell_{A_t,t})^2\Big] \\
&= 2\left(\frac{K}{\tilde{m}}\right)^3 B^2 \hat{\xi}_t. \qquad (22)
\end{aligned}$$

We plug the bounds obtained from (21) and (22) into inequality (19), and obtain

$$\sum_{t=1}^{T} \sum_{i=1}^{K} \hat{p}_{i,t}\ \mathbb{E}_{t-1}\Big[\Big(\hat{h}_{i,t} - \hat{\mu}_t\Big)^2\Big] \leq 2\left(\frac{6K}{\tilde{m}} + 2\lambda^2 \frac{K^3}{(\tilde{m})^3} B^2 \right) \sum_{t=1}^{T} \hat{\xi}_t. \qquad (23)$$

Recall that by definition (2), $\bar{\lambda} \leq \frac{1}{B}$. Hence, $\lambda < \frac{2\tilde{m}}{K}\bar{\lambda}$ gives

$$\lambda^2 \frac{K^2}{\tilde{m}^2} B^2 \leq 4,$$

we plug this bound into (23) and obtain

$$\sum_{t=1}^{T}\sum_{i=1}^{K} \hat{p}_{i,t}\, \mathbb{E}_{t-1}\Big[\Big(\hat{h}_{i,t} - \hat{\mu}_t\Big)^2\Big] \leq 20\frac{K}{\tilde{m}}\sum_{t=1}^{T}\hat{\xi}_t. \tag{24}$$

Next, we plug the bounds obtained in (19) and (24) into (18) to obtain

$$\sum_{t=1}^{T}\hat{\mu}_t \leq \min_{i\in[\![K]\!]}\sum_{t=1}^{T}\hat{h}_{i,t} + \frac{1}{\lambda}\sum_{t=1}^{T}M_t + \frac{\log(K)}{\lambda} + \frac{22\lambda K}{\tilde{m}}\sum_{t=1}^{T}\hat{\xi}_t.$$

□

Lemma F.3. *Let $\lambda \in \Big(0, \frac{2\tilde{m}}{K}\bar{\lambda}\Big)$, where $\bar{\lambda}$ is defined in (2) and $\tilde{m} = \max\{1, m-2\}$. Consider the martingale difference sequence $(M_t)_{t\in[\![T]\!]}$ defined in (17). We have*

- $\forall t \in [\![T]\!] : |M_t| \leq 3\lambda\frac{K}{\tilde{m}}B.$
- $\sum_{t=1}^{T}\mathbb{E}[M_t^2] \leq 5\frac{K}{\tilde{m}}\lambda^2\sum_{t=1}^{T}\hat{\xi}_t.$

Proof. Observe that the sequence $(M_t, \mathcal{F}_t)_{t\in[\![T]\!]}$ is a martingale difference. For any $t \in [\![T]\!]$, we have

$$\begin{aligned}
M_t &= \mathbb{E}\Big[\log\Big(\hat{Z}_{t+1}\Big)|\mathcal{F}_t\Big] - \log\Big(\hat{Z}_t\Big)\\
&= \log\left(\frac{\hat{Z}_t}{\hat{Z}_{t-1}}\right) - \mathbb{E}_{t-1}\left[\log\left(\frac{\hat{Z}_t}{\hat{Z}_{t-1}}\right)\right]\\
&= \log\left(\sum_{i=1}^{K}\hat{p}_{i,t}\exp\{-\lambda\hat{\ell}_{i,t} + \lambda^2\hat{v}_{i,t}\}\right) - \mathbb{E}_{t-1}\left[\log\left(\sum_{i=1}^{K}\hat{p}_{i,t}\exp\{-\lambda\hat{\ell}_{i,t} + \lambda^2\hat{v}_{i,t}\}\right)\right],
\end{aligned}$$

where we used the fact that $\hat{Z}_{t-1}$ is $\mathcal{F}_{t-1}$-measurable in the second line.

The loss function $\ell(.,y)$ is B-range-bounded for any y. Let $c_{\min}$ and c_{max} denote the lower and upper bounds, respectively, for the values of ℓ ($c_{\max} - c_{\min} \leq B$). Therefore, for any $i \in [\![K]\!]$, $\hat{\ell}_{i,t} \in \Big[c_{\min} - \frac{K}{\tilde{m}}B, c_{\max} + \frac{K}{\tilde{m}}B\Big]$ and $\hat{v}_{i,t} \in [0, (\frac{K}{\tilde{m}})^2B^2]$. Therefore

$$\exp\Big(\lambda c_{\max} - \frac{K}{\tilde{m}}\lambda B\Big) \leq \exp\Big(-\lambda\hat{\ell}_{i,t} + \lambda^2\hat{v}_{i,t}\Big) \leq \exp\left(-\lambda c_{\min} + \lambda\frac{K}{\tilde{m}}B + 2\lambda^2\frac{K^2}{\tilde{m}^2}B^2\right).$$

Hence

$$\lambda c_{\max} - \lambda\frac{KB}{\tilde{m}} \leq \log\left(\sum_{i=1}^{K}\hat{p}_{i,t}\exp\{-\lambda\hat{\ell}_{i,t} + \lambda^2\hat{v}_{i,t}\}\right) \leq -\lambda c_{\min} + \lambda\frac{KB}{\tilde{m}} + 2\lambda^2\frac{K^2B^2}{\tilde{m}^2}$$

Recall that M_t is a centered variable and $\lambda < \frac{\tilde{m}}{128KB}$. Therefore

$$|M_t| \leq 4\lambda \frac{K}{\tilde{m}} B. \tag{25}$$

Now, let us bound the quadratic characteristic of $(M_t)_t$. We have

$$\begin{aligned}\mathbb{E}_{t-1}\Big[M_t^2\Big] &= \mathrm{Var}_{t-1}\Big(\log\Big(\hat{Z}_t\Big)\Big) \\ &= \mathrm{Var}_{t-1}\Big(\log\Big(\hat{Z}_t\Big) - \log\Big(\hat{Z}_{t-1}\Big)\Big),\end{aligned} \tag{26}$$

where we used the fact that $\hat{Z}_{t-1}$ is $\mathcal{F}_{t-1}$-measurable.

Furthermore we have

$$\begin{aligned}\hat{Z}_t &= \sum_{i=1}^K \exp\Big(-\lambda \hat{L}_{i,t} + \lambda^2 \hat{V}_{i,t}\Big) \\ &= \sum_{i=1}^K \exp\Big(-\lambda \hat{L}_{i,t-1} + \lambda^2 \hat{V}_{i,t}\Big) \exp\Big(-\lambda \hat{\ell}_{i,t} + \lambda^2 \hat{v}_{i,t}\Big) \\ &= \sum_{i=1}^K \hat{p}_{i,t} \hat{Z}_{t-1} \exp\Big(-\lambda \hat{\ell}_{i,t} + \lambda^2 \hat{v}_{i,t}\Big).\end{aligned}$$

Hence

$$\begin{aligned}\frac{\hat{Z}_t}{\hat{Z}_{t-1}} &= \sum_{i=1}^K \hat{p}_{i,t} \exp\Big(-\lambda \hat{\ell}_{i,t} + \lambda^2 \hat{v}_{i,t}\Big) \\ &= \sum_{i=1}^K \hat{p}_{i,t} \exp\left(-\lambda\left(\ell_{A_t,t} + \frac{K}{\tilde{m}} \mathbb{1}(i \in \mathcal{U}_t)(\ell_{i,t} - \ell_{A_t,t})\right) + \lambda^2 \frac{K^2}{\tilde{m}^2} \mathbb{1}(i \in \mathcal{U}_t)(\ell_{i,t} - \ell_{A_t,t})^2\right) \\ &= \exp(-\lambda \ell_{A_t,t}) \sum_{i=1}^K \hat{p}_{i,t} \exp\left(-\lambda \frac{K}{\tilde{m}} \mathbb{1}(i \in \mathcal{U}_t)(\ell_{i,t} - \ell_{A_t,t}) + \lambda^2 \frac{K^2}{\tilde{m}^2} \mathbb{1}(i \in \mathcal{U}_t)(\ell_{i,t} - \ell_{A_t,t})^2\right) \\ &= \exp(-\lambda \ell_{A_t,t}) \mathbb{E}_{A'_t}\left[\exp\left(-\lambda \frac{K}{\tilde{m}} \mathbb{1}(A'_t \in \mathcal{U}_t)(\ell_{A'_t,t} - \ell_{A_t,t}) + \lambda^2 \frac{K^2}{\tilde{m}^2} \mathbb{1}(A'_t \in \mathcal{U}_t)(\ell_{A'_t,t} - \ell_{A_t,t})^2\right)\right],\end{aligned} \tag{27}$$

where A'_t is a random variable, independent of A_t, such that for each $i \in [\![K]\!]$, $\mathbb{P}(A'_t = i) = \hat{p}_{i,t}$, and $\mathbb{E}_{A'_t}$ is the expectation with respect to the random variable A'_t. So as to ease notation, denote

$$D_t := \frac{K}{\tilde{m}} \mathbb{1}(A'_t \in \mathcal{U}_t)(\ell_{A'_t,t} - \ell_{A_t,t}) - \lambda \frac{K^2}{\tilde{m}^2} \mathbb{1}(A'_t \in \mathcal{U}_t)(\ell_{A'_t,t} - \ell_{A_t,t})^2.$$

We take the logarithm of both sides of inequality (27), we have

$$\log\Big(\hat{Z}_t\Big) - \log\Big(\hat{Z}_{t-1}\Big) = -\lambda \ell_{A_t,t} + \log\Big(\mathbb{E}_{A'_t}[\exp(-\lambda D_t)]\Big).$$

We inject the equality above in (26). We obtain

$$\begin{aligned}\mathbb{E}_{t-1}\Big[M_t^2\Big] &= \mathrm{Var}_{t-1}\Big(-\lambda \ell_{A_t,t} + \log\Big(\mathbb{E}_{A'_t}[\exp(-\lambda D_t)]\Big)\Big) \\ &\leq 2\,\mathrm{Var}_{t-1}(\lambda \ell_{A_t,t}) + 2\,\mathrm{Var}_{t-1}\Big(\log\Big(\mathbb{E}_{A'_t}[\exp(-\lambda D_t)]\Big)\Big) \\ &\leq 2\,\mathrm{Var}_{t-1}(\lambda \ell_{A_t,t}) + 2\mathbb{E}_{t-1}\Big[\log^2\Big(\mathbb{E}_{A'_t}[\exp(-\lambda D_t)]\Big)\Big].\end{aligned} \tag{28}$$

Observe that

$$|\lambda D_t| = \left|\lambda\frac{K}{\tilde{m}}\mathbb{1}(A'_t \in \mathcal{U}_t)\Big(\ell_{A'_t,t} - \ell_{A_t,t}\Big) - \lambda^2\frac{K^2}{\tilde{m}^2}\mathbb{1}(A'_t \in \mathcal{U}_t)(\ell_{A'_t,t} - \ell_{A_t,t})^2\right| \leq \frac{1}{5}.$$

where we used $\lambda \in \Big(0, \frac{\tilde{m}}{128KB}\Big)$.

The function $x \mapsto \log^2(x)$ is convex on $[e^{-1}, e]$. Hence, using Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}_{t-1}\Big[\log^2\Big(\mathbb{E}_{A'_t}[\exp(-\lambda D_t)]\Big)\Big] &\leq \mathbb{E}_{t-1}\mathbb{E}_{A'_t}\Big[\log^2(\exp(-\lambda D_t))\Big] \\ &= \mathbb{E}_{t-1}\mathbb{E}_{A'_t}\Big[\lambda^2 D_t^2\Big] \qquad (29)\end{aligned}$$

From (28) and (29), we conclude that

$$\begin{aligned}\mathbb{E}_{t-1}\Big[M_t^2\Big] &\leq 2\lambda^2\operatorname{Var}_{t-1}(\ell_{A_t,t}) + 2\mathbb{E}_{t-1}\mathbb{E}_{A'_t}\Big[\lambda^2 D_t^2\Big] \\ &\leq 2\lambda^2\hat{\xi}_t + 2\mathbb{E}_{t-1}\mathbb{E}_{A'_t}\Big[\lambda^2 D_t^2\Big]. \qquad (30)\end{aligned}$$

where we used $\operatorname{Var}_{t-1}(\ell_{A_t,t}) = \hat{\xi}_t$. Furthermore:

$$\begin{aligned}\mathbb{E}_{t-1}\mathbb{E}_{A'_t}\Big[\lambda^2 D_t^2\Big] &\leq 2\mathbb{E}_{t-1}\mathbb{E}_{A'_t}\left[\frac{\lambda^2K^2}{\tilde{m}^2}\mathbb{1}(A'_t \in \mathcal{U}_t)\Big(\ell_{A'_t,t} - \ell_{A_t,t}\Big)^2 + \frac{K^4\lambda^4}{\tilde{m}^4}\mathbb{1}(A'_t \in \mathcal{U}_t)\Big(\ell_{A'_t,t} - \ell_{A_t,t}\Big)^4\right] \\ &\leq 2\left(\frac{\lambda^2K^2}{\tilde{m}^2} + \frac{\lambda^4K^4}{\tilde{m}^4}B^2\right)\mathbb{E}_{t-1}\mathbb{E}_{A'_t}\Big[\mathbb{1}(A'_t \in \mathcal{U}_t)(\ell_{A'_t,t} - \ell_{A_t,t})^2\Big] \\ &\leq 3\frac{\lambda^2K^2}{\tilde{m}^2}\mathbb{E}_{t-1}\mathbb{E}_{A'_t}\Big[\mathbb{1}(A'_t \in \mathcal{U}_t)(\ell_{A'_t,t} - \ell_{A_t,t})^2\Big] \\ &\leq 3\frac{\lambda^2K^2}{\tilde{m}^2}\mathbb{E}_{A'_t}\Big[\mathbb{E}_{t-1}[\mathbb{1}(A'_t \in \mathcal{U}_t)]\mathbb{E}_{t-1}\Big[(\ell_{A'_t,t} - \ell_{A_t,t})^2\Big]\Big] \\ &= 3\frac{\lambda^2K^2}{\tilde{m}^2}\frac{\tilde{m}}{K}\sum_{i,j=1}^{K}\hat{p}_{i,t}\hat{p}_{j,t}(\ell_{i,t} - \ell_{j,t})^2 \\ &= 3\frac{K}{\tilde{m}}\lambda^2\hat{\xi}_t, \qquad (31)\end{aligned}$$

where we used the independence of U_t and A_t conditionally to $\mathcal{F}_{t-1}$.

We plug (31) into (30). Therefore, it holds

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}_{t-1}\Big[M_t^2\Big] &\leq \sum_{t=1}^{T}\left(2\lambda^2\hat{\xi}_t + 3\frac{K}{\tilde{m}}\lambda^2\hat{\xi}_t\right) \\ &\leq 5\frac{K}{\tilde{m}}\lambda^2\sum_{t=1}^{T}\hat{\xi}_t.\end{aligned}$$

□

The following lemma provides a bound with high probability on the quantity $\hat{L}_{i,T} - \lambda\hat{V}_{i,T}$, for each $i \in [\![K]\!]$.

Lemma F.4. *For any $i \in [\![K]\!]$ and $\lambda \in (0, \frac{\tilde{m}\bar{\lambda}}{128K})$, with $\bar{\lambda}$ defined in (2) and $\tilde{m} = \max\{1, m-2\}$. We have for any $\delta \in (0, 1/3)$, with probability at least $1 - 6\delta$:*

$$\hat{L}_{i,T} - \lambda \hat{V}_{i,T} \leq L_{i,T} + \frac{721}{\lambda} \log\Big(\frac{\tilde{m}}{KB\lambda\delta}\Big).$$

Proof. Let $i \in [\![K]\!]$. Recall that we have for any $t \in [\![T]\!]$

$$\hat{\ell}_{i,t} - \ell_{i,t} = \Big(\frac{K}{\tilde{m}} \mathbb{1}(i \in \mathcal{U}_t) - 1\Big)(\ell_{i,t} - \ell_{A_t,t})$$
$$\hat{\ell}_{i,t} - \ell_{A_t,t} = \frac{K}{\tilde{m}} \mathbb{1}(i \in \mathcal{U}_t)(\ell_{i,t} - \ell_{A_t,t}).$$

We introduce the following notation

$$\nu_{i,t} := \mathbb{E}_{t-1}\Big[(\ell_{i,t} - \ell_{A_t,t})^2\Big].$$

We have

$$\begin{aligned}
\hat{L}_{i,T} - \lambda \hat{V}_{i,T} &= L_{i,T} + \sum_{t=1}^{T} \Big(\hat{\ell}_{i,t} - \ell_{i,t}\Big) - \lambda \sum_{t=1}^{T} \Big(\frac{K}{\tilde{m}}\Big)^2 \mathbb{1}(i \in \mathcal{U}_t)(\ell_{i,t} - \ell_{A_t,t})^2 \\
&= L_{i,T} + \underbrace{\sum_{t=1}^{T} \Big(\hat{\ell}_{i,t} - \ell_{i,t}\Big) - \lambda \frac{K}{2\tilde{m}} \sum_{t=1}^{T} \nu_{i,t}}_{\text{Term 21}} \\
&\quad + \underbrace{\lambda \frac{K}{2\tilde{m}} \sum_{t=1}^{T} \nu_{i,t} - \lambda \sum_{t=1}^{T} \Big(\frac{K}{\tilde{m}}\Big)^2 \mathbb{1}(i \in \mathcal{U}_t)(\ell_{i,t} - \ell_{A_t,t})^2}_{\text{Term 22}}. \qquad (32)
\end{aligned}$$

Bounding Term 21: Observe that $(\hat{\ell}_{i,t} - \ell_{i,t})_t$ is a martingale difference with respect to the filtration $\mathcal{F}$, bounded in absolute value by $\frac{K}{\tilde{m}}B$. Let us bound its quadratic characteristic. Recall that A_t and $\mathcal{U}_t$ are independent conditionally to $\mathcal{F}_{t-1}$. We have

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}_{t-1}\Big[(\hat{\ell}_{i,t} - \ell_{i,t})^2\Big] &= \sum_{t=1}^{T} \mathbb{E}_{t-1}\left[\left(1 - \frac{K}{\tilde{m}} \mathbb{1}(i \in \mathcal{U}_t)\right)^2 (\ell_{i,t} - \ell_{A_t,t})^2\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{t-1}\left[\left(1 - \frac{K}{\tilde{m}} \mathbb{1}(i \in \mathcal{U}_t)\right)^2\right] \mathbb{E}_{t-1}\Big[(\ell_{i,t} - \ell_{A_t,t})^2\Big] \\
&\leq \frac{K}{\tilde{m}} \sum_{t=1}^{T} \mathbb{E}_{t-1}\Big[(\ell_{i,t} - \ell_{A_t,t})^2\Big] \\
&= \frac{K}{\tilde{m}} \sum_{t=1}^{T} \nu_{i,t}.
\end{aligned}$$

Next, we apply Corollary D.4 to the sequence $(\hat{\ell}_{i,t} - \ell_{i,t})_{t \in [\![T]\!]}$: We take $c = \lambda KB/(4\tilde{m}) \leq 1$, with probability at least $1 - 3\delta$, it holds

$$\sum_{t=1}^{T} \Big(\hat{\ell}_{i,t} - \ell_{i,t}\Big) - \lambda \frac{K}{2\tilde{m}} \sum_{t=1}^{T} \nu_{i,t} \leq \frac{720}{\lambda} \log\Big(\frac{\tilde{m}}{KB\lambda\delta}\Big). \qquad (33)$$

Bounding Term 22: Define the sequence $(Q_t)_{t\in[\![T]\!]}$ as follows:

$$Q_t := -\lambda\frac{K^2}{\tilde{m}^2}\mathbb{1}(i\in\mathcal{U}_t)(\ell_{i,t}-\ell_{A_t,t})^2 + \lambda\frac{K}{\tilde{m}}\nu_{i,t}.$$

Notice that (Q_t) is a martingale difference sequence with respect to the filtration $\mathcal{F}$, and bounded in absolute value by $2\lambda\frac{K^2B^2}{\tilde{m}^2}$. Let us bound its quadratic characteristic. We have

$$\begin{aligned}\sum_{t=1}^T \mathbb{E}_{t-1}\Big[Q_t^2\Big] &\le \lambda^2\sum_{t=1}^T \mathbb{E}_{t-1}\left[\frac{K^4}{\tilde{m}^4}\mathbb{1}(i\in\mathcal{U}_t)(\ell_{i,t}-\ell_{A_t,t})^4\right]\\ &\le \lambda^2\frac{K^4B^2}{\tilde{m}^4}\sum_{t=1}^T \mathbb{E}_{t-1}[\mathbb{1}(i\in\mathcal{U}_t)]\mathbb{E}_{t-1}\Big[(\ell_{i,t}-\ell_{A_t,t})^2\Big]\\ &= \frac{K^3\lambda^2B^2}{\tilde{m}^3}\sum_{t=1}^T\nu_{i,t}.\end{aligned}$$

Next, we apply Corollary D.4 to this sequence. We take $c=1$, we have with probability at least $1-3\delta$:

$$\begin{aligned}\sum_{t=1}^T Q_t - \lambda\frac{K}{2\tilde{m}}\sum_{t=1}^T\nu_{i,t} &\le 36\lambda\frac{K^2}{\tilde{m}^2}B^2\log\Big(\delta^{-1}\Big)\\ &\le \frac{9}{32}B\log(\delta^{-1}).\end{aligned} \tag{34}$$

Conclusion: To conclude, we inject bounds obtain in (33) and (34) into (32).

□

We provide a key lemma that will be used in the proof of Theorem 4.1 and 4.2.

Lemma F.5. *Let $\lambda\in\Big(0,\frac{\tilde{m}}{128K}\bar{\lambda}\Big)$, where $\bar{\lambda}$ is defined in (2). Consider Algorithm 3 with inputs (λ,m). We have with probability at least $1-9\delta$*

$$\sum_{t=1}^T\hat{\mu}_t - \frac{7\bar{\lambda}}{32}\sum_{t=1}^T\hat{\xi}_t \le \min_{i\in[\![K]\!]} L_{i,T} + c\,\frac{1}{\lambda}\log\Big(\frac{\tilde{m}}{B\lambda\delta}\Big)$$

where $\tilde{m}=\max\{1,m-1\}$ and c is a numerical constant.

Proof. For each $i\in[\![K]\!]$ and $t\in[\![T]\!]$, let $\hat{h}_{i,t} := \hat{\ell}_{i,t}-\lambda\hat{v}_{i,t}$ and $h_{i,t} := \mathbb{E}_{t-1}\Big[\hat{h}_{i,t}\Big]$. Using Lemma F.2, we have

$$\begin{aligned}\sum_{t=1}^T\hat{\mu}_t - \frac{7\bar{\lambda}}{32}\sum_{t=1}^T\hat{\xi}_t &\le \min_{i\in[\![K]\!]}\sum_{t=1}^T\hat{h}_{i,t} + \frac{1}{\lambda}\sum_{t=1}^T M_t + \frac{\log(K)}{\lambda} + \Big(\frac{11\lambda K}{\tilde{m}} - \frac{7}{32}\bar{\lambda}\Big)\sum_{t=1}^T\hat{\xi}_t\\ &\le \min_{i\in[\![K]\!]}\sum_{t=1}^T\hat{h}_{i,t} + \frac{1}{\lambda}\sum_{t=1}^T M_t - \frac{\bar{\lambda}}{8}\sum_{t=1}^T\hat{\xi}_t + \frac{\log(K)}{\lambda},\end{aligned} \tag{35}$$

where we used the fact that $\lambda \in \left(0, \frac{\tilde{m}}{128K}\bar{\lambda}\right)$.

In order to conclude, we only need bounds on the terms $\min_{i\in[\![K]\!]} \sum_{t=1}^T \hat{h}_{i,t}$ and $\frac{1}{\lambda}\sum_{t=1}^T M_t$. Recall that Lemma F.3 shows that (M_t) is a martingale difference sequence and provides a bound on its conditional variance. Hence, applying Corollary D.4 to this sequence with $c = 3B\bar{\lambda}/40$, with probability at least $1-3\delta$, it holds

$$\frac{1}{\lambda}\sum_{t=1}^T M_t - \frac{\tilde{m}\bar{\lambda}}{40\bar{\lambda}^2 K}\sum_{t=1}^T 5\frac{K}{\tilde{m}}\lambda^2\hat{\xi}_t \le \frac{324K}{\tilde{m}\bar{\lambda}}\left(\log\delta^{-1} + 2\log_2^+\left(\frac{7024}{B^2\bar{\lambda}^2}\right)\right).$$

We conclude that

$$\frac{1}{\lambda}\sum_{t=1}^T M_t - \frac{\bar{\lambda}}{8}\sum_{t=1}^T \hat{\xi}_t \le 8428\frac{K}{\tilde{m}\bar{\lambda}}\log\left(\frac{1}{B\bar{\lambda}\delta}\right). \tag{36}$$

Next, to bound the term $\min_{i\in[\![K]\!]} \sum_{t=1}^T \hat{h}_{i,t}$ we use Lemma F.4. We have with probability at least $1-6\delta$

$$\begin{aligned}\min_{i\in[\![K]\!]}\sum_{t=1}^T \hat{h}_{i,t} &= \min_{i\in[\![K]\!]} \hat{L}_{i,T} - \lambda\hat{V}_{i,T} \\ &\le \min_{i\in[\![K]\!]} L_{i,T} + \frac{721}{\lambda}\log\left(\frac{\tilde{m}}{B\lambda\delta}\right).\end{aligned} \tag{37}$$

Finally, we inject (36) and (37) into (35) and use $\lambda \in \left(0, \frac{\tilde{m}}{128K}\bar{\lambda}\right)$. We obtain that with probability at least $1-9\delta$

$$\sum_{t=1}^T \hat{\mu}_t - \frac{7\bar{\lambda}}{32}\sum_{t=1}^T \hat{\xi}_t \le \min_{i\in[\![K]\!]} L_{i,T} + c\,\frac{1}{\lambda}\log\left(\frac{\tilde{m}}{B\lambda\delta}\right),$$

where c is a numerical constant. □

The following Lemma is specific to the case $m = p = 2$ and IC = True in Algorithm 4.

Lemma F.6. *Let $\lambda \in \left(0, \frac{\bar{\lambda}}{352K^2}\right)$, where $\bar{\lambda}$ is defined in (2). Consider Algorithm 4 with input λ. We have with probability at least $1-9\delta$*

$$\sum_{t=1}^T \hat{\mu}_t - \frac{3\bar{\lambda}}{32K}\sum_{t=1}^T \hat{\xi}_t \le \min_{i\in[\![K]\!]} L_{i,T} + c\,\frac{1}{\lambda}\log\left(\frac{1}{B\lambda\delta}\right),$$

where c is a numerical constant.

Proof. For each $i \in [\![K]\!]$ and $t \in [\![T]\!]$, let $\hat{h}_{i,t} := \hat{\ell}_{i,t} - \lambda\hat{v}_{i,t}$ and $h_{i,t} := \mathbb{E}_{t-1}\left[\hat{h}_{i,t}\right]$. Using Lemma F.2, we have

$$\begin{aligned}\sum_{t=1}^T \hat{\mu}_t - \frac{3\bar{\lambda}}{32K}\sum_{t=1}^T \hat{\xi}_t &\le \min_{i\in[\![K]\!]}\sum_{t=1}^T \hat{h}_{i,t} + \frac{1}{\lambda}\sum_{t=1}^T M_t + \frac{\log(K)}{\lambda} + \left(11\lambda K - \frac{3\bar{\lambda}}{32K}\right)\sum_{t=1}^T \hat{\xi}_t \\ &\le \min_{i\in[\![K]\!]}\sum_{t=1}^T \hat{h}_{i,t} + \frac{1}{\lambda}\sum_{t=1}^T M_t - \frac{\bar{\lambda}}{16K}\sum_{t=1}^T \hat{\xi}_t + \frac{\log(K)}{\lambda},\end{aligned} \tag{38}$$

where we used the fact that $\lambda \in \left(0, \frac{\bar{\lambda}}{352K^2}\right)$.

The remainder of the proof is similar to the proof of Lemma F.5.

Lemma F.3 provides the following bound with probability at least $1 - 3\delta$

$$\frac{1}{\lambda}\sum_{t=1}^{T} M_t - \frac{\bar{\lambda}}{16K}\sum_{t=1}^{T}\hat{\xi}_t \leq \frac{3520}{\bar{\lambda}}\log\left(\frac{1}{B\bar{\lambda}\delta}\right). \tag{39}$$

Moreover, Lemma F.4 provides the following bound with probability at least $1 - 6\delta$

$$\min_{i\in[\![K]\!]}\sum_{t=1}^{T}\hat{h}_{i,t} = \min_{i\in[\![K]\!]} L_{i,T} + \frac{721}{\lambda}\log\left(\frac{1}{B\lambda\delta}\right). \tag{40}$$

Finally, we inject (39) and (40) into (38). We obtain that with probability at least $1 - 9\delta$

$$\sum_{t=1}^{T}\hat{\mu}_t - \frac{3\bar{\lambda}}{32K}\sum_{t=1}^{T}\hat{\xi}_t \leq \min_{i\in[\![K]\!]} L_{i,T} + c\,\frac{1}{\lambda}\log\left(\frac{1}{B\lambda\delta}\right),$$

where c is a numerical constant. □

G On the sampling strategy in the case $m = p = 2, \mathbf{IC} = \mathbf{True}$

Let $\boldsymbol{p}$ denote a distribution over $[\![K]\!]$. Let $\mathcal{E} = \{A, B\}$ denote a random set of elements in $[\![K]\!]$, such that A is sampled from $[\![K]\!]$ following $\boldsymbol{p}$ and B is sampled independently and uniformly at random from $[\![K]\!]$ (possibly $A = B$ and $\mathcal{E}$ is a singleton). Therefore, we have for each $u, v \in [\![K]\!]$, such that $u \neq v$:

$$\mathbb{P}(\mathcal{E} = \{u, v\}) = \frac{\boldsymbol{p}_u + \boldsymbol{p}_v}{K},$$

and

$$\mathbb{P}(\mathcal{E} = \{u\}) = \frac{\boldsymbol{p}_u}{K}.$$

Finally, let $\boldsymbol{p}_{\mathcal{E}}$ denote the restriction of the distribution $\boldsymbol{p}$ on $\mathcal{E}$, conditional to $\mathcal{E}$. Let X denote a random variable following $\boldsymbol{p}_{\mathcal{E}}$

$$\forall i \in \mathcal{E} : \boldsymbol{p}_{\mathcal{E}}(X = i) = \boldsymbol{p}(X = i|\mathcal{E}) = \frac{\boldsymbol{p}_i}{\sum_{j\in\mathcal{E}}\boldsymbol{p}_j}.$$

Let I and J denote two random variables on $[\![K]\!]$ sampled conditionally to $\mathcal{E}$, independently following $\boldsymbol{p}_{\mathcal{E}}$ (with replacement).

In this section, we prove two results: the marginal distribution of I on $[\![K]\!]$ is identical to $\boldsymbol{p}$, and a bound on the probabilities of the joint unconditional distribution of (I, J).

Lemma G.1. *For each $i \in [\![K]\!]$,*

$$\mathbb{P}(I = i) = \boldsymbol{p}_i.$$

Proof. Fix $i \in [\![K]\!]$. Let $\mathcal{K}$ denote the set of subsets of $[\![K]\!]$, constituted of at most two elements.

For any subset $\boldsymbol{a} \in \mathcal{K}$, define

$$\boldsymbol{p}_{\boldsymbol{a}} := \sum_{i \in \boldsymbol{a}} \boldsymbol{p}_i.$$

We have

$$\begin{aligned}
\mathbb{P}(I = i) &= \sum_{\boldsymbol{a} \in \mathcal{K}} \mathbb{P}(I = i, \mathcal{E} = \boldsymbol{a}) \\
&= \mathbb{P}(I = i | \mathcal{E} = \{i\})\ \mathbb{P}(\mathcal{E} = \{i\}) + \sum_{u \in [\![K]\!] \setminus \{i\}} \mathbb{P}(I = i | \mathcal{E} = \{u, i\})\ \mathbb{P}(\mathcal{E} = \{u, i\}) \\
&= \frac{\boldsymbol{p}_i}{K} + \sum_{u \in [\![K]\!] \setminus \{i\}} \frac{\boldsymbol{p}_i}{\boldsymbol{p}_u + \boldsymbol{p}_i}\ \frac{\boldsymbol{p}_u + \boldsymbol{p}_i}{K} \\
&= \frac{\boldsymbol{p}_i}{K} + \frac{\boldsymbol{p}_i}{K}(K - 1) \\
&= \boldsymbol{p}_i.
\end{aligned}$$

□

Lemma G.2. *For each $i, j \in [\![K]\!]$,*

$$\mathbb{P}(I = i, J = j) \geq \frac{1}{K} \boldsymbol{p}_i \boldsymbol{p}_j.$$

Proof. Fix $i, j \in [\![K]\!]$. Let $\mathcal{K}$ denote the set of subsets of $[\![K]\!]$, constituted of at most two elements.

Suppose that $i = j$. We have

$$\begin{aligned}
\mathbb{P}(I = i, J = i) &= \sum_{\boldsymbol{a} \in \mathcal{K}} \mathbb{P}(I = i, J = i, \mathcal{E} = \boldsymbol{a}) \\
&= \sum_{\boldsymbol{a} \in \mathcal{K}} \mathbb{P}(I = i, J = i | \mathcal{E} = \boldsymbol{a}) \mathbb{P}(\mathcal{E} = \boldsymbol{a}) \\
&= \sum_{\boldsymbol{a} \in \mathcal{K}} \mathbb{P}(I = i | \mathcal{E} = \boldsymbol{a})^2 \mathbb{P}(\mathcal{E} = \boldsymbol{a}),
\end{aligned}$$

where we used the fact that I and J are independent conditionally to $\mathcal{E}$ and that I and J follow the same distribution. We use Jensen's inequality:

$$\begin{aligned}
\mathbb{P}(I = i, J = i) &\geq \left(\sum_{a \in \mathcal{K}} \mathbb{P}(I = i | \mathcal{E} = \boldsymbol{a}) \mathbb{P}(\mathcal{E} = \boldsymbol{a}) \right)^2 \\
&= \boldsymbol{p}_i^2.
\end{aligned}$$

Now suppose that $i \neq j$. We have

$$\begin{aligned}\mathbb{P}(I=i,J=j) &= \mathbb{P}(I=i,J=j,\mathcal{E}=\{i,j\})\\ &= \mathbb{P}(I=i|\mathcal{E}=\{i,j\})\mathbb{P}(J=j|\mathcal{E}=\{i,j\})\mathbb{P}(\mathcal{E}=\{i,j\})\\ &= \frac{\boldsymbol{p}_i}{\boldsymbol{p}_i+\boldsymbol{p}_j}\frac{\boldsymbol{p}_j}{\boldsymbol{p}_i+\boldsymbol{p}_j}\frac{\boldsymbol{p}_i+\boldsymbol{p}_j}{K}\\ &= \frac{\boldsymbol{p}_i\boldsymbol{p}_j}{K}.\end{aligned}$$

□

H Proof of Theorems 4.1 and 4.2

We consider the notation of Algorithms 3 and 4. Let $\hat{\pi}_{ij,t} = \mathbb{P}(I_t = i, J_t = j|\mathcal{F}_{t-1})$. Introduce ($\hat{\mu}_t$ and $\hat{\xi}_t$ are the same quantities as in the previous section):

$$\begin{aligned}\hat{\mu}_t &:= \sum_{i\in[\![K]\!]} \hat{p}_{i,t}\ell_{i,t},\\ \hat{\nu}_t &:= \frac{1}{2}\sum_{i,j\in[\![K]\!]} \hat{\pi}_{ij,t}\ (\ell_{i,t}-\ell_{j,t})^2\\ \hat{\xi}_t &:= \frac{1}{2}\sum_{i,j\in[\![K]\!]} \hat{p}_{i,t}\hat{p}_{j,t}\ (\ell_{i,t}-\ell_{j,t})^2\end{aligned}$$

We have, using (8) with $c = 1/\bar{\lambda}$ (implied by Assumption 1, see Lemma 1.1):

$$\begin{aligned}\sum_{t=1}^T \ell_t\Big(\frac{F_{I_t}+F_{J_t}}{2}\Big) &\leq \sum_{t=1}^T\Big(\frac{1}{2}\ell_{I_t,t}+\frac{1}{2}\ell_{J_t,t}-\frac{\bar{\lambda}}{2}(\ell_{I_t,t}-\ell_{J_t,t})^2\Big)\\ &= \underbrace{\frac{1}{2}\sum_{t=1}^T \mathbb{U}_t+\frac{1}{2}\sum_{t=1}^T \mathbb{U}'_t-\frac{\tilde{m}\bar{\lambda}}{32K}\sum_{t=1}^T\hat{\xi}_t-\frac{\bar{\lambda}}{2}\sum_{t=1}^T\mathbb{W}_t-\frac{\bar{\lambda}}{4}\sum_{t=1}^T\hat{\nu}_t}_{\text{Term 1}}\\ &\quad+\underbrace{\sum_{t=1}^T\hat{\mu}_t+\frac{\tilde{m}\bar{\lambda}}{32K}\sum_{t=1}^T\hat{\xi}_t-\frac{\bar{\lambda}}{4}\sum_{t=1}^T\hat{\nu}_t}_{\text{Term 2}},\end{aligned}$$

where

$$\mathbb{U}_t := \ell_{I_t,t}-\hat{\mu}_t; \qquad \mathbb{U}'_t := \ell_{J_t,t}-\hat{\mu}_t; \qquad \mathbb{W}_t := (\ell_{I_t,t}-\ell_{J_t,t})^2-\hat{\nu}_t.$$

Section H.1 below is common to Theorem 4.1 and 4.2. In Section H.2, we distinguish between the case where ($p = m = 2, \mathrm{IC} = \mathrm{True}$) and ($p = 2, m \geq 3$) or ($p = 2, m = 2, \mathrm{IC} = \mathrm{False}$).

H.1 Bounding Term 1

Recall that in Algorithm 3 we have by definition of I_t, conditionally to $\mathcal{F}_{t-1}$: $I_t \sim \hat{p}_t$. Furthermore, in Algorithm 4, using Lemma G.1, conditionally to $\mathcal{F}_{t-1}$, we have: $I_t \sim \hat{p}_t$.

Hence, $(\mathbb{U}_t)_{t\in[\![T]\!]}$ is a martingale difference sequence bounded in absolute value by B. Moreover, we have for all $t \in [\![T]\!]$

$$\mathbb{E}\Big[\mathbb{U}_t^2|\mathcal{F}_{t-1}\Big] = \hat{\xi}_t.$$

Next we apply the high probability bound provided by Corollary D.4 to the sequence $(\mathbb{U}_t)_{t\in[\![T]\!]}$, with $c = \tilde{m}B\bar{\lambda}/(32K)$. We have with probability at least $1-3\delta$

$$\sum_{t=1}^{T}\mathbb{U}_t - \frac{\tilde{m}}{32K}\bar{\lambda}\sum_{t=1}^{T}\hat{\xi}_t \leq 7700\frac{K}{\tilde{m}\bar{\lambda}}\log\Big(\frac{K}{\tilde{m}B\bar{\lambda}\delta}\Big). \tag{41}$$

Recall that in Algorithm 3 and 4, I_t and J_t have the same marginal distribution. Therefore, with probability at least $1-3\delta$, (41) holds with $\mathbb{U}_t$ replaced by $\mathbb{U}'_t$.

Similarly, the sequence $((-\bar{\lambda}/2)\mathbb{W}_t)_{t\in[\![T]\!]}$ is a martingale difference bounded in absolute value by $\bar{\lambda}B^2$. For any $t \in [\![T]\!]$,

$$\frac{\bar{\lambda}^2}{4}\mathbb{E}\Big[\mathbb{W}_t^2|\mathcal{F}_t\Big] \leq \frac{\bar{\lambda}^2}{4}\mathbb{E}\Big[(\ell_{I_t,t}-\ell_{J_t,t})^4|\mathcal{F}_{t-1}\Big] \leq \frac{\bar{\lambda}^2B^2}{4}\hat{\nu}_t.$$

Next, we apply Corollary D.4 to the sequence $((-\bar{\lambda}/2)\mathbb{W}_t)_{t\in[\![T]\!]}$: We take $c=1$, we have with probability $1-3\delta$:

$$\begin{aligned}-\frac{\bar{\lambda}}{2}\sum_{t=1}^{T}\mathbb{W}_t - \frac{\bar{\lambda}}{4}\sum_{t=1}^{T}\hat{\nu}_t &\leq 72\bar{\lambda}B^2\log(\delta^{-1})\\ &\leq 72B\log(\delta^{-1}).\end{aligned} \tag{42}$$

Using (41) and (42), we conclude that with probability $1-9\delta$

$$\text{Term 1} \leq 7772\frac{K}{\tilde{m}\bar{\lambda}}\log\Big(\frac{K}{\tilde{m}B\bar{\lambda}\delta}\Big). \tag{43}$$

H.2 Bounding Term 2

We divide this part of the proof into two section (depending on the expression of the joint distribution $\hat{\pi}_t$).

H.2.1 Case ($p=2$ and $m\geq 3$) or ($p=2$, $m=2$ and IC = False)

Recall that conditionally to $\mathcal{F}_{t-1}$, the played experts I_t and J_t are sampled independently according to $\hat{p}_t$ from $[\![K]\!]$. Therefore for any $i,j\in[\![K]\!]$, $\hat{\pi}_{ij,t} = \hat{p}_{i,t}\hat{p}_{j,t}$ and $\hat{\nu}_t = \hat{\xi}_t$.

Hence, Term 2 satisfies the following bound

$$\text{Term 2} \leq \sum_{t=1}^{T}\hat{\mu}_t - \frac{7\bar{\lambda}}{32}\sum_{t=1}^{T}\hat{\xi}_t.$$

Using the first claim of Lemma F.5, we have if $\lambda \in \left(0, \frac{\tilde{m}}{128K}\bar{\lambda}\right)$

$$\text{Term } 2 \leq \min_{i \in [\![K]\!]} L_{i,T} + c\,\frac{1}{\lambda}\log\left(\frac{\tilde{m}}{B\lambda\delta}\right), \tag{44}$$

where c is a numerical constant. The conclusion of the theorem follows by combining the upper bounds obtained in (43) and (44).

H.2.2 Case $m = p = 2$ and IC = True:

Using Lemma G.1 we have $I_t \sim \hat{p}_t$. Furthermore, using Lemma G.2 we have that for any $i, j \in [\![K]\!]$, any $t \in [\![T]\!]$:

$$\hat{\pi}_{ij,t} \geq \frac{1}{K}\hat{p}_{i,t}\hat{p}_{j,t}.$$

Therefore $\hat{\nu}_t \geq \frac{1}{K}\hat{\xi}_t$, and we have the following bound on Term 2:

$$\text{Term } 2 \leq \sum_{t=1}^{T}\hat{\mu}_t - \frac{3\bar{\lambda}}{32K}\sum_{t=1}^{T}\hat{\xi}_t.$$

Using the second claim of Lemma F.6, we have if $\lambda \in \left(0, \frac{\bar{\lambda}}{352K^2}\right)$

$$\sum_{t=1}^{T}\hat{\mu}_t - \frac{7}{32B}\sum_{t=1}^{T}\hat{\xi}_t \leq \min_{i \in [\![K]\!]} L_{i,T} + c\,\frac{1}{\lambda}\log\left(\frac{1}{\lambda B\delta}\right). \tag{45}$$

The conclusion of the theorem follows by combining the upper bounds obtained in (43) and (45).

I Proofs of lower bounds, Theorem 5.1 and Theorem 5.2

The proofs of Theorem 5.1 and Theorem 5.2 are presented in four steps. The only difference between the proofs is in the last step. Thus the first three steps are common to both proofs.

We adapt the main steps of Auer et al. [1995] to our setting. The gist of the proof is the following. We construct a distribution with very correlated experts. In this situation, going from a weighted average of experts to a single expert with the largest weight does not change the prediction risk much. Then, we use some classical arguments in deriving lower bounds for the expected regret using information theory results.

Let $T > 0$ be fixed, we consider that the loss function is the squared loss and we focus on the particular setting where the target variables (Y_t) are identically 0.

First step: Specifying the distributions. We start by considering a deterministic forecaster. We denote by $\mathbb{P}_i$ the joint distribution of expert predictions, where all experts are identical and distributed as one and the same Bernoulli variable with parameter $1/2$, except the optimal expert i who has distribution $\mathcal{B}\left(\frac{1}{2} - \epsilon\right)$ but is still strongly correlated to the others.

More precisely, let $(U_t)_{t\in[\![T]\!]}$ be a sequence of independent random variables distributed according the uniform distribution on $[0,1]$. We consider that in each round the expert predictions have the following joint distribution $\mathbb{P}_i$:

- For $j \neq i$: $F_{j,t} = \mathbb{1}\Big(U_t \leq \frac{1}{2}\Big)$.
- $F_{i,t} = \mathbb{1}\Big(U_t \leq \frac{1}{2} - \epsilon\Big)$.

Recall that in this setting we have for any $k, j \in [\![K]\!] \setminus \{i\}$

$$\mathbb{E}_i[F_{j,t}F_{k,t}] = \frac{1}{2}$$
$$\mathbb{E}_i[F_{i,t}F_{j,t}] = \frac{1}{2} - \epsilon.$$

Finally, we denote by $\mathbb{P}_0$ the joint distribution where all experts are equal to the same Bernoulli(1/2) variables, i.e., experts predictions are defined by $F_{i,t} = \mathbb{1}(U_t \leq 1/2)$, $i \in [\![K]\!]$.

Second step: Strategy Reduction. Suppose that the player follows a deterministic strategy $\mathcal{A}$. In each round t, given $\mathcal{F}_{t-1}$, this strategy selects a subsets S_t of $[\![K]\!]$ of size m and a sequence of non-negative weights $(\alpha_{i,t})_{i\in S_t}$, such that $\sum_i \alpha_{i,t} = 1$, and plays the convex combination $\sum_{i\in S_t} \alpha_{i,t}F_{i,t}$.

For such a strategy $\mathcal{A}$, we associate a strategy $\hat{\mathcal{A}}$, such that in each round, we run the strategy $\mathcal{A}$ except that we play only the expert with the largest weight $\hat{i}_t \in \operatorname{Arg\,Max}_{i\in S_t} \alpha_{i,t}$.

Let us analyse the difference of the cumulative loss between the strategies $\mathcal{A}$ and $\hat{\mathcal{A}}$. Let $l_t(\mathcal{A})$ denote the loss of the strategy $\mathcal{A}$ at round t. We have

$$\mathbb{E}_i\Big[l_t(\mathcal{A}) - l_t(\hat{\mathcal{A}})\Big] = \mathbb{E}_i\left[\Big(\sum_{j\in S_t} \alpha_{j,t}F_{j,t}\Big)^2\right] - \mathbb{E}_i\left[\Big(\sum_{j\in S_t} \mathbb{1}\Big(\hat{i}_t = j\Big)F_{j,t}\Big)^2\right].$$

If $i \notin S_t$ then we have $\mathbb{E}_i[l_t(\mathcal{A}) - l_t(\hat{\mathcal{A}})] = 0$.

If $i \in S_t$ and $\hat{i}_t = i$, we have (let $j \in [\![K]\!]$ such that $j \neq i$)

$$\begin{aligned}
\mathbb{E}_i\Big[l_t(\mathcal{A}) - l_t(\hat{\mathcal{A}})\Big] &= \mathbb{E}_i\Big[\big((1-\alpha_{i,t})F_{j,t} + \alpha_{i,t}F_{i,t}\big)^2\Big] - \mathbb{E}_i[F_{i,t}] \\
&= (1-\alpha_{i,t})^2\frac{1}{2} + \alpha_{i,t}^2\Big(\frac{1}{2} - \epsilon\Big) + 2\alpha_{i,t}(1-\alpha_{i,t})\Big(\frac{1}{2} - \epsilon\Big) - \frac{1}{2} + \epsilon \\
&= \epsilon(1-\alpha_{i,t})^2 \\
&\geq 0.
\end{aligned}$$

If $i \in S_t$ and $\hat{i}_t \neq i$, we have (let $j \in [\![K]\!]$ such that $j \neq i$)

$$\begin{aligned}
\mathbb{E}_i\Big[l_t(\mathcal{A}) - l_t(\hat{\mathcal{A}})\Big] &= \mathbb{E}_i\Big[((1-\alpha_{i,t})F_{j,t} + \alpha_{i,t}F_{i,t})^2\Big] - \mathbb{E}_i[F_{j,t}] \\
&= (1-\alpha_{i,t})^2\frac{1}{2} + \alpha_{i,t}^2\left(\frac{1}{2}-\epsilon\right) + 2\alpha_{i,t}(1-\alpha_{i,t})\left(\frac{1}{2}-\epsilon\right) - \frac{1}{2} \\
&= \epsilon\alpha_{i,t}^2 - 2\epsilon\alpha_{i,t} \\
&\geq -\frac{3}{4}\epsilon,
\end{aligned}$$

where we used the fact that $\alpha_{i,t} \in [0, 1/2]$, since $\hat{i}_t \neq i$.

To summarize, in the worst case, the excess loss between $\mathcal{A}$ and $\hat{\mathcal{A}}$ is $-\frac{3}{4}\epsilon$. Hence, we have the following lower bound on the expected regret between the two strategies:

$$\mathcal{R}_T(\mathcal{A}) - \mathcal{R}_T(\hat{\mathcal{A}}) \geq -\frac{3}{4}T\epsilon. \tag{46}$$

Third step: Information theoretic tools. Let us introduce the following notation: assume the player follows a deterministic strategy $\mathcal{A}$, and let $Z_t = (C_t, \mathbb{1}_t(F_{i,t})_{i\in C_t})$ denote the information disclosed to the player at time t. Denote $\boldsymbol{Z}^t = (Z_1, \ldots, Z_t)$ the entire information available to the player since the start. The quantities $Z_t, \boldsymbol{Z}^t$ are considered as random variables, whose distribution is determined by the underlying experts distribution, and the player strategy $\mathcal{A}$.

Lemma I.1. *Let $F(\boldsymbol{Z}^T)$ be any fixed function of the player observations, taking values in $[0, B]$. Then for any $i \in [\![K]\!]$ and any player strategy $\mathcal{A}$,*

$$\mathbb{E}_i\Big[F\Big(\boldsymbol{Z}^T\Big)\Big] \leq \mathbb{E}_0\Big[F\Big(\boldsymbol{Z}^T\Big)\Big] + \frac{B}{2}\sqrt{\mathbb{E}_0[N_i]\log(1-2\epsilon)^{-1}},$$

where $N_i = \sum_{i=1}^T \mathbf{1}\{i \in C_t\}$.

In the case where $|C_t| = 1$ for all t, the following sharper bound holds:

$$\mathbb{E}_i\Big[F\Big(\boldsymbol{Z}^T\Big)\Big] \leq \mathbb{E}_0\Big[F\Big(\boldsymbol{Z}^T\Big)\Big] + \frac{B}{2}\sqrt{\mathbb{E}_0[N_i]\log(1-4\epsilon^2)^{-1}},$$

Proof. Fix $i \in [\![K]\!]$. Denote $\mathbb{Q}_i$ the distribution of $\boldsymbol{Z}^T$ induced by expert distribution $\mathbb{P}_i$ and a fixed player strategy $\mathcal{A}$ (omitted from the notation for simplicity). For any function G bounded by R, it is well-known that it holds $|\mathbb{E}_{X\sim\mathbb{P}}[G(X)] - \mathbb{E}_{X\sim\mathbb{Q}}[G(X)]| \leq 2R\|\mathbb{P} - \mathbb{Q}\|_{TV}$, where $\|\cdot\|_{TV}$ denotes the total variation distance. Hence, by shifting F by $-B/2$, we get

$$\mathbb{E}_i\Big[F(\boldsymbol{Z}^T)\Big] - \mathbb{E}_0\Big[F(\boldsymbol{Z}^T)\Big] \leq B\|\mathbb{Q}_i - \mathbb{Q}_0\|_{TV} \leq B\sqrt{\frac{1}{2}\mathrm{KL}(\mathbb{Q}_0\|\mathbb{Q}_i)},$$

by Pinsker's inequality, where KL(.) denotes the Kullback-Leibler divergence.

Next, we will compute the quantity $\mathrm{KL}(\mathbb{Q}_0\|\mathbb{Q}_i)$. The chain rule for relative entropy (Theorem 2.5.3 in Cover, 1999) gives:

$$\mathrm{KL}(\mathbb{Q}_0\|\mathbb{Q}_i) = \sum_{t=1}^T \mathrm{KL}\Big(\mathbb{Q}_0\Big\{Z_t|\boldsymbol{Z}^{t-1}\Big\}\|\mathbb{Q}_i\Big\{Z_t|\boldsymbol{Z}^{t-1}\Big\}\Big), \tag{47}$$

where

$$\mathrm{KL}\Big(\mathbb{Q}_0\Big\{Z_t|\boldsymbol{Z}^{t-1}\Big\}\|\mathbb{Q}_i\{Z_t|\boldsymbol{Z}^{t-1}\Big\}\Big) := \sum_{\boldsymbol{z}^t} \mathbb{Q}_0\Big\{\boldsymbol{z}^{t-1}\Big\}\mathbb{Q}_0\Big\{z_t|\boldsymbol{z}^{t-1}\Big\}\log\Bigg(\frac{\mathbb{Q}_0\{z_t|\boldsymbol{z}^{t-1}\}}{\mathbb{Q}_i\{z_t|\boldsymbol{z}^{t-1}\}}\Bigg)$$
$$= \sum_{\substack{\boldsymbol{z}^t \\ \text{s.t. } i\in C_t}} \mathbb{Q}_0\Big\{\boldsymbol{z}^{t-1}, C_t\Big\}\mathbb{Q}_0\{z_t|C_t\}\log\Big(\frac{\mathbb{Q}_0\{z_t|C_t\}}{\mathbb{Q}_i\{z_t|C_t\}}\Big).$$

The last line holds because $\mathbb{Q}_\bullet\{z_t|\boldsymbol{z}^{t-1}\} = \mathbb{Q}_\bullet\{z_t|\boldsymbol{z}^{t-1}, C_t\}\mathbb{Q}_\bullet\{C_t|\boldsymbol{z}^{t-1}\}$, and it holds $\mathbb{Q}_0\{C_t|\boldsymbol{z}^{t-1}\} = \mathbb{Q}_i\{C_t|\boldsymbol{z}^{t-1}\}$ since the strategy's play only depends on past observations; also $\mathbb{Q}_\bullet\{z_t|\boldsymbol{z}^{t-1}, C_t\} = \mathbb{Q}_\bullet\{z_t|C_t\}$ since the observed experts' losses at round t are independent of the past given the choice of C_t. Furthermore, if $i \notin C_t$, one has $\mathbb{Q}_0\{z_t|C_t\} = \mathbb{Q}_i\{z_t|C_t\}$.

On the other hand, if z_t is such that $i \in C_t$, then:

- under $\mathbb{Q}_0$ since all experts are identical and equal to the same $\mathrm{Ber}(1/2)$ variable (and Y_t is identically 0), $\mathbb{Q}_0(z_t|C_t)$ only charges the two points with all observed losses equal to 0 (denote this u_0) or all equal to 1 (denote this u_1), each with probability 1/2;
- under $\mathbb{Q}_i$, it holds $\mathbb{Q}_i(u_1|C_t) = \frac{1}{2} - \epsilon$ and $\mathbb{Q}_i(u_0|C_t) \geq \frac{1}{2}$. In fact, if $|C_t| \geq 2$, then $\mathbb{Q}_i(u_0|C_t) = \frac{1}{2}$ (since with probability ϵ under $\mathbb{Q}_i$, we observe a state that is neither u_0 nor u_1, namely when all observed experts err but F_i), and if $|C_t| = 1$, then $\mathbb{Q}_i(u_0|C_t) = \frac{1}{2} + \varepsilon$ (since F_i alone is observed then).

Therefore, in general

$$\mathrm{KL}\Big(\mathbb{Q}_0\Big\{Z_t|\boldsymbol{Z}^{t-1}\Big\}\|\mathbb{Q}_i\{Z_t|\boldsymbol{Z}^{t-1}\Big\}\Big) \leq \mathbb{P}_0(i \in C_t)\Bigg(\frac{1}{2}\log\Big(\frac{1/2}{1/2-\epsilon}\Big) + \frac{1}{2}\log\Big(\frac{1/2}{1/2}\Big)\Bigg)$$
$$\leq \frac{1}{2}\mathbb{P}_0(i \in C_t)\log(1-2\epsilon)^{-1}.$$

In the case where $|C_t| = 1$ for all t, we get the sharper bound

$$\mathrm{KL}\Big(\mathbb{Q}_0\Big\{Z_t|\boldsymbol{Z}^{t-1}\Big\}\|\mathbb{Q}_i\{Z_t|\boldsymbol{Z}^{t-1}\Big\}\Big) = \mathbb{P}_0(i \in C_t)\Bigg(\frac{1}{2}\log\Big(\frac{1/2}{1/2-\epsilon}\Big) + \frac{1}{2}\log\Big(\frac{1/2}{1/2+\epsilon}\Big)\Bigg)$$
$$= \frac{1}{2}\mathbb{P}_0(i \in C_t)\log\Big(1-4\epsilon^2\Big)^{-1}.$$

Plugging this into (47), we obtain

$\mathrm{KL}(\mathbb{Q}_0\|\mathbb{Q}_i) \leq -\frac{1}{2}\mathbb{E}_0[N_i]\log(1-2\epsilon)$, resp. $\mathrm{KL}(\mathbb{Q}_0\|\mathbb{Q}_i) \leq -\frac{1}{2}\mathbb{E}_0[N_i]\log(1-4\epsilon^2)$, if $|C_t| = 1$ for all t, leading to the claims.

□

Fourth step for Theorem 5.1: lower bounding the regret of $\hat{\mathcal{A}}$ in the case $|C_t| \geq 2$. Recall $\hat{i}_t$ denotes the single expert played by the "reduced" strategy $\hat{\mathcal{A}}$. At round t, the expected loss for the player playing $\hat{\mathcal{A}}$ is given by

$$\mathbb{E}_i\Big[l_{t,\hat{i}_t}\Big] = \Big(\frac{1}{2} - \epsilon\Big)\mathbb{P}_i\Big(\hat{i}_t = i\Big) + \frac{1}{2}\mathbb{P}_i\Big(\hat{i}_t \neq i\Big) = \frac{1}{2} - \epsilon\ \mathbb{P}_i\Big(\hat{i}_t = i\Big).$$

For each $j \in [\![K]\!]$ let $M_j := \sum_{t=1}^{T} \mathbb{1}\Big\{\hat{i}_t = j\Big\}$. Hence

$$\sum_{t=1}^{T} \mathbb{E}_i\Big[l_{t,\hat{i}_t}\Big] = \frac{T}{2} - \epsilon\, \mathbb{E}_i[M_i],$$

and the regret with respect to the optimal arm i under $\mathbb{P}_i$ is

$$\mathbb{E}_i\Big[\mathcal{R}_T(\hat{\mathcal{A}})\Big] = \epsilon(T - \mathbb{E}_i[M_i]). \tag{48}$$

We can apply Lemma I.1 to $F(\boldsymbol{Z}^t) = M_i$: since we assume the player follows a deterministic strategy, M_i is a function of the information $\boldsymbol{Z}^t$ available to the player, bounded by T. Thus it holds:

$$\mathbb{E}_i[M_i] \leq \mathbb{E}_0[M_i] + \frac{T}{2}\sqrt{\mathbb{E}_0[N_i]\log(1-2\epsilon)^{-1}}. \tag{49}$$

Observe that $\sum_{i=1}^{K} M_i = T$ and $\sum_{i=1}^{K} N_i = mT$. Hence

$$\begin{aligned}
\sum_{i=1}^{K} \mathbb{E}_i[M_i] &\leq \sum_{i=1}^{K} \mathbb{E}_0[M_i] + \frac{T}{2}\sum_{i=1}^{K}\sqrt{\mathbb{E}_0[N_i]\log(1-2\epsilon)^{-1}} \\
&\leq \mathbb{E}_0\left[\sum_{i=1}^{K} M_i\right] + \frac{TK}{2}\sqrt{\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_0[N_i]\log(1-2\epsilon)^{-1}} \\
&= T + T^{\frac{3}{2}}\sqrt{mK\epsilon},
\end{aligned}$$

where we used the fact that for $\epsilon \in (0, 1/4) : -\log(1-2\epsilon) \leq 4\epsilon$. Let $\mathbb{P}_* = \frac{1}{K}\sum_{i=1}^{K}\mathbb{P}_i$ the adversary choosing uniformly at random among the expert distributions $\mathbb{P}_i$ at the start of the game (i.e. choosing at random the optimal expert). From the above and (48) we deduce

$$\mathbb{E}_*\Big[\mathcal{R}_T(\hat{\mathcal{A}})\Big] \geq \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_i\Big[\mathcal{R}_T(\hat{\mathcal{A}})\Big] \geq \epsilon\left(T\left(1 - \frac{1}{K}\right) - T^{\frac{3}{2}}\sqrt{\frac{m\epsilon}{K}}\right)$$

Using inequality (46), we obtain

$$\mathbb{E}_*[\mathcal{R}_T(\mathcal{A})] \geq \epsilon\left(T\left(\frac{1}{4} - \frac{1}{K}\right) - T^{\frac{3}{2}}\sqrt{\frac{m\epsilon}{K}}\right) \geq \epsilon T\left(\frac{1}{20} - \sqrt{\frac{Tm\epsilon}{K}}\right),$$

if $K \geq 5$. Choosing $\epsilon = \frac{1}{900}\frac{K}{mT}$, we get

$$\mathbb{E}_*[\mathcal{R}_T(\mathcal{A})] \geq 10^{-5}\ \frac{K}{m}.$$

Recall that this lower bound was derived for deterministic players. Generalizing this bound to random players follows simply by applying Fubini's theorem. Also since the bound is in expectation over expert predictions drawn according to $\mathbb{P}_*$, for any strategy $\mathcal{A}$ there exists at least one deterministic sequence of expert forecasts with regret larger than its expectation.

Fourth step for Theorem 5.2: lower bounding the regret of $\hat{\mathcal{A}}$ in the case $|C_t| = 1$. The only difference between the proof in this case and the proof in the previous case is the bound given by Lemma I.1. The regret with respect to the optimal arm i under $\mathbb{P}_i$ is

$$\mathbb{E}_i\Big[\mathcal{R}_T(\hat{\mathcal{A}})\Big] = \epsilon(T - \mathbb{E}_i[M_i]). \tag{50}$$

We can apply Lemma I.1 to $F(\boldsymbol{Z}^t) = M_i$: since we assume the player follows a deterministic strategy, M_i is a function of the information $\boldsymbol{Z}^t$ available to the player, bounded by T. Thus it holds:

$$\mathbb{E}_i[M_i] \leq \mathbb{E}_0[M_i] + \frac{T}{2}\sqrt{\mathbb{E}_0[N_i]\log(1-4\epsilon^2)^{-1}}.$$

Observe that $\sum_{i=1}^K M_i = T$ and $\sum_{i=1}^K N_i = T$. Hence

$$\begin{aligned}
\sum_{i=1}^K \mathbb{E}_i[M_i] &\leq \sum_{i=1}^K \mathbb{E}_0[M_i] + \frac{T}{2}\sum_{i=1}^K \sqrt{\mathbb{E}_0[N_i]\log(1-4\epsilon^2)^{-1}} \\
&\leq \mathbb{E}_0\left[\sum_{i=1}^K M_i\right] + \frac{TK}{2}\sqrt{\frac{1}{K}\sum_{i=1}^K \mathbb{E}_0[N_i]\log(1-2\epsilon^2)^{-1}} \\
&= T + T^{\frac{3}{2}}\sqrt{2K\epsilon^2},
\end{aligned}$$

where we used the fact that for $\epsilon \in (0, 1/4) : -\log(1-4\epsilon^2) \leq 8\epsilon^2$. Let $\mathbb{P}_* = \frac{1}{K}\sum_{i=1}^K \mathbb{P}_i$ the adversary choosing uniformly at random among the expert distributions $\mathbb{P}_i$ at the start of the game (i.e. choosing at random the optimal expert). From the above and (50) we deduce

$$\mathbb{E}_*\Big[\mathcal{R}_T(\hat{\mathcal{A}})\Big] \geq \frac{1}{K}\sum_{i=1}^K \mathbb{E}_i\Big[\mathcal{R}_T(\hat{\mathcal{A}})\Big] \geq \epsilon\left(T\left(1 - \frac{1}{K}\right) - T^{\frac{3}{2}}\sqrt{2\frac{\epsilon^2}{K}}\right)$$

Using inequality (46), we obtain

$$\mathbb{E}_*[\mathcal{R}_T(\mathcal{A})] \geq \epsilon\left(T\left(\frac{1}{4} - \frac{1}{K}\right) - T^{\frac{3}{2}}\sqrt{2\frac{\epsilon^2}{K}}\right) \geq \epsilon T\left(\frac{1}{20} - \sqrt{2\frac{T\epsilon^2}{K}}\right),$$

if $K \geq 5$. Choosing $\epsilon = \frac{1}{30}\sqrt{\frac{K}{T}}$, we get

$$\mathbb{E}_*[\mathcal{R}_T(\mathcal{A})] \geq 10^{-5}\ \sqrt{KT}.$$

The generalization for the random players follows directly using the same argument as in the fourth step of the proof of Theorem 5.1.

J Proof of Theorem 5.3

Let ℓ be the squared loss: $l(x, y) = (x - y)^2$ on $\mathcal{X} = \mathcal{Y} = [0, 1]$. Consider the game protocol presented in Algorithm 1 with $p = 1$ and $m \in [\![K]\!]$. Suppose that the target variable y is identically equal to 0 ($y_t = 0$ for all $t \in [\![T]\!]$). Suppose that at each round $t \in [\![T]\!]$, for each

expert $i \in [\![K]\!]$, the prediction $F_{i,t}$ follows a Bernoulli distribution of a parameter denoted $\ell_{i,t}$. We have

$$\mathbb{E}[\mathcal{R}_T] = \sum_{t=1}^{T} \mathbb{E}[F_{I_t,t}] - \min_{i \in [\![K]\!]} \sum_{t=1}^{T} \mathbb{E}[F_{i,t}].$$

The game protocol presented in Algorithm 1 reduces to the K-armed bandit game with m feedbacks in each round, analysed in Seldin et al. [2014].

Theorem below presented in Seldin et al. [2014] (the full version including appendices) as Theorem 2, provides a lower bound for the regret.

Theorem J.1 (Seldin et al. [2014])**.** *For the K-armed bandit game with mT observed rewards and $T \geq \frac{3}{16}\frac{K}{m}$,*

$$\inf \sup \mathbb{E}[\mathcal{R}_T] \geq 0.03\sqrt{\frac{K}{m}T},$$

where the infinimum is over all playing strategies and the supremum is over all individual sequences.

The result stated in Theorem 5.3 is a direct consequence of the Theorem J.1 and the setting described above.

K Some implementation details and algorithmic complexity

We discuss here some details of the implementation of Algorithms 2, 3, 4, more specifically concerning the cost of keeping track of the distribution $\hat{p}_t$ and of sampling from it at each round. We concentrate on Algorithm 3 for simplicity, but the arguments below apply to all algorithms.

We start with a fundamental observation. While the definitions (6), (7) for $\hat{\ell}_{i,t}$ and $\hat{v}_{i,t}$ were written in order to emphasize the unbiased character of the loss estimates, the algorithm is unchanged if we use instead the shifted "pseudo-loss" estimates

$$\tilde{\ell}_{i,t} := \hat{\ell}_{i,t} - \ell_{I_t,t} = \frac{K}{\tilde{m}} \mathbb{1}(i \in \mathcal{U}_t)(\ell_{i,t} - \ell_{I_t,t}), \tag{51}$$

and further observe that it holds $\hat{v}_{i,t} = \tilde{\ell}^2_{i,t}$. Using the above pseudo-losses in place of the estimated losses does not change the sampling distribution $\hat{p}_t$, since all estimated losses are shifted by the *same* quantity $\ell_{I_t,t}$, which gets cancelled through the normalization in the definition (5) of the EW distribution $\widehat{p}_t$.

Observe that the pseudo-loss estimates $\tilde{\ell}_{i,t}$ (as well as the corresponding variance estimates $\hat{v}_{i,t}$) are equal to zero for all $i \notin \mathcal{U}_t$. Therefore, to keep track of the cumulative pseudo-loss estimates $\tilde{L}_{i,t} = \sum_{k \leq t} \tilde{\ell}_{i,k}$, only $|\mathcal{U}_t| = \max\{m-2, 1\}$ of them have to be updated at each round.

In order to keep track and sample efficiently from $\hat{p}_t$, we propose the following construction. Let T be a balanced binary tree of depth $\lceil \log_2(K) \rceil$, with K leaves, such that each leaf $i \in \partial T$ is identified to an expert index. Furthermore, assume that each internal node u of T stores the partial sum $S_{u,t} = \sum_{v \in \partial T_u} \exp\Big(-\lambda \tilde{L}_{v,t} + \lambda^2 \hat{V}_{v,t}\Big)$, where T_u is the subtree of T rooted at node

u. Then, by the above considerations, it holds that $S_{u,t} = D_t \sum_{v \in \partial T_u} \hat{p}_{u,t} = D_t \hat{p}_t(\partial T_u)$, where D_t is a factor depending only on t but not on the node u. Note also that $D_t = S_\emptyset$, where $\emptyset$ denotes the root note of T. It is then possible to sample efficiently $I_t \sim \hat{p}_t$ in a standard manner, as follows:

1. Generate $U \sim \mathrm{Unif}[0,1]$, and put $Z = S_\emptyset U$. Let $v = \emptyset$.
2. If v is a leaf of T, stop and output v.
3. Let $v_{\mathrm{left}}, v_{\mathrm{right}}$ denote the two descendent nodes of v.
4. If $Z < S_{v_{\mathrm{left}}}$, then let $v \leftarrow v_{\mathrm{left}}$ and go to step 2.
5. Otherwise, i.e. $Z \geq S_{v_{\mathrm{left}}}$, let $v \leftarrow v_{\mathrm{right}}$, $Z \leftarrow Z - S_{v_{\mathrm{left}}}$, and go to step 2.

It easy to check that the above sampling returns a random sample from the probability $\hat{p}_t$. (Namely, each time that step 2 is reached, conditionally to past steps Z is uniformly distributed in the interval $[0, S_v]$, and therefore the left or right descendent of u is picked with probability $\hat{p}_t(\partial T_{v_{\mathrm{left}}} | \partial T_v)$ resp. $\hat{p}_t(\partial T_{v_{\mathrm{right}}} | \partial T_v)$; the chain rule yields the claim.) Obviously, the computing complexity of the above is $\mathcal{O}(\log K)$ (the depth of the tree).

Furthermore, to update the quantities stored at the nodes of T at each round, since only the estimated cumulative pseudo-losses of experts $i \in \mathcal{U}_t$ have their value modified, it is sufficient to do the following for each $i \in \mathcal{U}_t$:

1. Let v be the leaf representing i. Update $S_v \leftarrow S_v \exp\Big(-\lambda \tilde{\ell}_{i,t} + \lambda^2 \hat{v}_{i,t}\Big)$.
2. Go up the tree to the root and sequentially update all ancestors w of v according to $S_w = S_{w_{\mathrm{left}}} + S_{w_{\mathrm{right}}}$.

Again, the computing complexity of this update operation is $\mathcal{O}(\log K)$.

All in all, the computational cost of the initialization of the tree is $\mathcal{O}(K)$, but then at each round the computational cost of the sampling and update operations is $\mathcal{O}(m \log(K))$.