Optimal learning with Bernstein Online Aggregation

Olivier Wintenberger

Optimal learning with Bernstein Online Aggregation

Olivier Wintenberger
olivier.wintenberger@upmc.fr
LSTA, Case 158 Université Pierre et Marie Curie 4 place Jussieu 75005 Paris FRANCE

April 4, 2014

Abstract

We introduce a new recursive aggregation procedure called Bernstein Online Aggregation (BOA). The exponential weights include an accuracy term and a second order term that is a proxy of the quadratic variation as in [17]. This second term stabilizes the procedure that is optimal in different senses. We first obtain optimal regret bounds in the deterministic context. Then, an adaptive version is proved to solve the so-called impossible tuning problem already solved in [15]. The second order bounds in the deterministic context are extended to a general stochastic context using the cumulative predictive risk. Such conversion provides the main result of the paper, an inequality of a novel type comparing the procedure with any deterministic aggregation procedure for an integrated criteria. It provides an observable confident interval on the excess of risk of the BOA procedure. To assert the optimality, we consider finally the iid case for strongly convex and Lipschitz continuous losses and we prove that the rate of convergence is of the optimal order given in [27]. The batch version of the BOA procedure is then the first adaptive solution satisfying an optimal oracle inequality with high probability.

Exponential weighted averages, Learning theory, Individual sequences.

1 Introduction and main results

We consider the online setting where observations $\mathcal{F}_t = \{(X_1, Y_1), \dots, (X_t, Y_t)\}$ are available recursively ($(X_0, Y_0) = (x_0, y_0)$ arbitrary). The goal of statistical learning is to predict $Y_{t+1} \in \mathbb{R}$ given $X_{t+1} \in \mathcal{X}$, for $\mathcal{X}$ a probability space, on the basis of $\mathcal{F}_t$. In this paper, we index with the subscript t any random element that is adapted with $\mathcal{F}_t$. A learner is a function $\mathcal{X} \mapsto \mathbb{R}$, denoted $\hat{f}_t$, that depends only on the past observations $\mathcal{F}_t$ and such that $\hat{f}_t(X_{t+1})$ is close to Y_{t+1}. This closeness at time $t+1$ is addressed by the predictive risk

$$\mathbb{E}[\ell(Y_{t+1}, \hat{f}_t(X_{t+1})) \mid \mathcal{F}_t]$$

where $\ell : \mathbb{R}^2 \to \mathbb{R}$ is a loss function. We define an online learner $\hat{f}$ as a recursive algorithm that produces at each time $t \geq 1$ a learner: $\hat{f} = (\hat{f}_0, \hat{f}_1, \hat{f}_2, \dots)$. The accuracy of an online

learner is quantified by the cumulative predictive risk

$$R_n(\hat{f}) = \frac{1}{n+1}\sum_{t=0}^{n} \mathbb{E}[\ell(Y_{t+1}, \hat{f}_t(X_{t+1})) \mid \mathcal{F}_t]. \tag{1}$$

Given a finite set $\mathcal{F} = \{f_1, \ldots, f_M\}$ of online learners, it is well known that any procedure that will select one learner is suboptimal. Thus, recursive aggregation procedures

$$\hat{f} = \sum_{j=1}^{M} \pi_j f_j = \Big(\sum_{j=1}^{M} \pi_{j,0} f_{j,0}, \sum_{j=1}^{M} \pi_{j,1} f_{j,1}, \sum_{j=1}^{M} \pi_{j,2} f_{j,2} \ldots \Big)$$

have been intensively studied, see the seminal book [10]. The predictive performance of the resulting online learner $\hat{f}$ can be compared with the best element of the dictionary $\mathcal{F}$ or with the best deterministic aggregation of the online learners of the dictionary. We denote $f_\pi = \mathbb{E}_\pi[f_j]$ any such deterministic aggregation procedures

$$f_\pi = \Big(\sum_{j=1}^{M} f_{j,0}, \sum_{j=1}^{M} f_{j,1}, \sum_{j=1}^{M} f_{j,2}, \ldots \Big)$$

with $\pi = (\pi_j)_{1 \leq \leq M}$ a measure on $\{1, \ldots, M\}$.

In this article, we provide a new recursive procedure, called Bernstein Online Aggregation (BOA), and we compare it with the best deterministic aggregation f_π. The weights $\pi_t = (\pi_{j,t})_{1 \leq j \leq M}$ are defined following a recursive rule. This rule, and the name of the Bernstein Online Aggregation procedure, derived from the study of the concentration properties of the difference between the cumulative predictive risk and the cumulative regret

$$\sum_{t=0}^{n} \mathbb{E}[\ell(Y_{t+1}, \hat{f}_t(X_{t+1})) \mid \mathcal{F}_t] - \mathbb{E}[\ell(Y_{t+1}, \hat{f}_{\pi^*}(X_{t+1})) \mid \mathcal{F}_t] - \ell(Y_{t+1}, \hat{f}_t(X_{t+1})) + \ell(Y_{t+1}, \hat{f}_{\pi^*}(X_{t+1}) \tag{2}$$

where π^* is the best element or the best deterministic weights. It is a martingale (M_t) adapted to the filtration $(\mathcal{F}_t)$. Recall that for any martingale (M_t) adapted to a filtration $(\mathcal{F}_t)$, we denote $\Delta M_t = M_t - M_{t-1}$ the difference of martingale, $< M >_t = \sum_{j=1}^{t} \mathbb{E}[\Delta M_j^2 \mid \mathcal{F}_t]$ and $[M]_t = \sum_{j=1}^{t} \Delta M_j^2$ its (predictable) quadratic variation. Instead of using the classical Bernstein inequality for martingales [14, 30], we develop its empirical counterpart that provides concentration of M via a variance term in $[M]$ instead of $< M >$:

Theorem 1.1. *Let M be a martingale such that*

$$\mathbb{E}(\Delta M_{t-}^4 \mid \mathcal{F}_{t-1}) \leq \mathbb{E}(\Delta M_{t-}^2 \mid \mathcal{F}_{t-1}), \qquad t > 0. \tag{3}$$

Then for any stopping time τ we have

$$\mathbb{P}(M_\tau \geq \sqrt{2\eta [M]_\tau x} + 7x/4) \leq e^{-x}, \qquad x > 0.$$

Empirical Bernstein's inequality have already been developed in [4, 25] to use successfully a variance proxy into, respectively, the multi-armed bandit and penalized ERM problems. Applying Theorem 1.1, we estimate the deviations of M_n in (2) via its quadratic variation $[M]$. An optimal aggregation procedure is a procedure that has a minimal cumulative regret and a minimal quadratic variation. However, in our context (2), the quadratic variation $[M]$ depends on π^* that is unknown and we will use a proxy of the quadratic variation denoted $V_{j,n+1} = \sum_{t=1}^{n+1} \ell_{j,t}^2$, where

$$\ell_{j,t} = \ell(Y_t, f_{j,t-1}(X_t)) - \mathbb{E}_{\pi_{t-1}}[\ell(Y_t, f_{j',t-1}(X_t))],$$

estimates ΔM_t, $1 \leq t \leq n+1$, as in (2) with π^* any Dirac mass at $\{j\}$, $1 \leq j \leq M$.

The BOA procedure is an exponential weights procedure that tends to minimize the quadratic variation through the terms $\ell_{j,t}$. This procedure favors online learners that predicts accurately and that are close to the last aggregative online learner, ensuring the stability in time and a small quadratic variation. It is derived in 3 different versions: the aggregation procedure itself described in Figure 1 and denoted $\hat{f}$, its randomized version of BOA, denoted $\bar{f}$, and defined as $\mathbb{P}(\bar{f}_t = f_{j,t}) = \pi_{j,t}$ and its batch version, denoted $\tilde{f}$, and defined as $\tilde{f} = \frac{1}{n+1} \sum_{t=0}^{n} \hat{f}_t$.

Parameters: Temperature parameter $\eta > 0$.
Initialization: Set $\pi_{j,0} > 0$ such that $\sum_{j=1}^{M} \pi_{j,0} = 1$ and $\hat{f}_0 = \sum_{j=1}^{M} \pi_{j,0} f_{j,0}$.
For: Each time round $1 \leq t \leq n$,

1. Compute the weight vector $\pi_t = (\pi_{j,t})_{1 \leq j \leq M}$ recursively from π_0: $\pi_{j,t}/\pi_{j,t-1} =$

$$\frac{\exp(-2^{-1}\eta\ell(Y_t, f_{j,t-1}(X_t)) - \eta^2 \ell_{j,t}^2)}{\sum_{j=1}^{M} \exp(-2^{-1}\eta\ell(Y_t, f_{j,t-1}(X_t)) - \eta^2 \ell_{j,t}^2)},$$

2. Output the aggregated prediction $\hat{f}_t = \mathbb{E}_{\pi_t}[f_{j,t}]$,

Figure 1: The BOA algorithm

With no convexity assumption on the loss, only the randomized version of BOA can be compared with the best element of the dictionary. The cumulative predictive risk associated with the randomized procedure is then

$$R_n(\bar{f}) = \sum_{t=1}^{n+1} \mathbb{E}[\mathbb{E}_{\pi_{t-1}}[\ell(Y_t, f_{j,t-1}(X_t))] \mid \mathcal{F}_{t-1}].$$

Thus it explains the term $\mathbb{E}_{\pi_{t-1}}[\ell(Y_t, f_{j',t-1}(X_t))]$ in the proxy of the quadratic variation of M_n in (2). Notice that such randomized algorithms that take into account a proxy of the variance have been studied in the iid context by [6]. In all the sequel, we focus on the less general context of a convex loss ℓ as the extension to the randomized version and

general loss is straightforward. We denote ℓ' its sub gradient with respect to its second argument. For convex losses, the aggregation procedure $\hat{f}$ provides sharper cumulative predictive risks than the randomized one $\bar{f}$ as, by Jensen's inequality, $R_n(\hat{f}) \leq R_n(\bar{f})$.

In the prediction from experts context, exponential weights aggregation with a proxy of the variance has been considered in [17]. In this article, we use the sub gradient trick, see [10], and replace in BOA the loss $\ell(Y_t, f_{j,t-1}(X_t))$ with its linearized approximation $\mathbb{E}_{\pi_{t-1}}[\ell'(Y_t, f_{j',t-1}(X_t))]f_{j,t-1}(X_t)$. The proxy of the quadratic variation $V_{j,n+1} = \sum_{t=1}^{n+1} \ell_{j,t}^2$ is then modify such that

$$\ell_{j,t} = \mathbb{E}_{\pi_{t-1}}[\ell'(Y_t, f_{j',t-1}(X_t))(f_{j,t-1}(X_t) - f_{j',t-1}(X_t))].$$

Linearizing the loss, we can compare the regret of the BOA procedure with the best deterministic aggregation of the elements in the dictionary. Working conditionally on the observations, we obtain as a first result a deterministic bound on the regret

$$\mathcal{R}(\hat{f}) \leq \min_{\pi} \Big\{ \mathcal{R}(f_\pi) + 2\eta \mathbb{E}_\pi[V_{j,n+1}] + \frac{2}{\eta} \mathcal{K}(\pi, \pi_0) \Big\}.$$

Here $\mathcal{R}(f)$ is the cumulative loss of any online learner $f = (f_0, f_1, f_2, \ldots)$:

$$\mathcal{R}(f) = \sum_{t=0}^{n} \ell(Y_{t+1}, f_t(X_{t+1})).$$

The optimality of such regret bounds is difficult to assert. Following the pioneer work of [11], we analyze the second order properties of a new adaptive version of exponential weights, see Figure 2 for its application on the BOA procedure. The novelty, compared with classical adaptive procedures developed in [11], is the dependence of the learning rates with respect to j. It solves the so called *impossible tuning question*, see [11, 15], by considering the rule for learning rates

$$\eta_{j,t} = \min \left\{ \frac{1}{E}, \sqrt{\frac{\log(M)}{\sum_{s=1}^{t} \ell_{j,s}^2}} \right\}, \qquad t \geq 0$$

where E is a known bound of the linearized losses $\mathbb{E}_{\pi_{t-1}}[\ell'(Y_t, f_{j',t-1}(X_t))]f_{j,t-1}(X_t)$. We also give a fully adaptive version of the algorithm for cases when the bound E is unknown. For these adaptive BOA procedures, we obtain regret bounds of the form

$$\mathcal{R}(\hat{f}) \leq \min_{\pi} \Big\{ \mathcal{R}(f_\pi) + C\mathbb{E}_\pi \Big[\sqrt{V_{j,n+1}} \Big] \sqrt{\log M} \Big\} + CE \log M,$$

for some "constant" (increasing in $\log \log n$) $C > 0$, see Theorems 3.2 and 3.3 for details. The optimality of such bounds is difficult to assert because it depends on the variance term $V_{j,n+1}$. For the square loss, as $\ell'(x,y)^2 \leq 4\ell(x,y)$, we derive optimal regret bounds of the form

$$\mathcal{R}(\hat{f}) \leq \min_{1 \leq j \leq M} \mathcal{R}(f_j) + CE \log M.$$

Parameter: a rule to sequentially pick the learning rates $(\eta_{j,t})$.
Initialization: Set $L_{j,0} = 0$, $\eta_{j,0} = \eta_0$, $\pi_{j,0} > 0$ such that $\sum_{j=1}^M \pi_{j,0} = 1$ and $\hat{f}_0 = \sum_{j=1}^M \pi_{j,0} f_{j,0}$.
For: each time round $t \geq 1$,

1. Compute recursively

$$L_{j,t} = L_{j,t-1} + \mathbb{E}_{\pi_{t-1}}[\ell'(Y_t, f_{j',t-1}(X_t))] f_{j,t-1}(X_t) + 2\eta_{j,t-1}\ell_{j,t}^2,$$

2. Compute the weights vector $\pi_t = (\pi_{j,t})_{1 \leq j \leq M}$:

$$\pi_{j,t} = \frac{\eta_{j,t} \exp(-2^{-1}\eta_{j,t} L_{j,t})\pi_{j,0}}{\sum_{j=1}^M \eta_{j,t} \exp(-2^{-1}\eta_{j,t} L_{j,t})\pi_{j,0}},$$

3. Output the aggregated prediction $\hat{f}_t = \mathbb{E}_{\pi_t}[f_{j,t}]$,

Figure 2: The adaptive BOA algorithm

Such bounds are also achieved by classical exponential weighting algorithms with no proxy of the quadratic variation, see [28, 16]. It is natural as the cumulative loss is not a risk and thus it only depends on the accuracy of the procedure, and not on its second order properties.

The proxy of the quadratic variation is necessary when we convert this results from the cumulative loss to the cumulative predictive risk. For the same adaptive BOA procedure, we obtain with probability $1 - e^{-x}$, $x > 0$

$$R_n(\hat{f}) \leq \min_{\pi} \left\{ R_n(f_\pi) + C\mathbb{E}_\pi \left[\sqrt{V_{j,n+1}} \right] \sqrt{\log M} (1 + x(\log M)^{-1}) \right\} + CE(\log M + x).$$

It is remarkable to obtain a result for an integrated criteria with no assumption on the dependence of the stochastic observations $\mathcal{F}_n$. It is the main result of the paper and the main motivation for the introduction of the BOA procedure; without the proxy of the quadratic variation, it seems impossible to convert a result on the regret to a result on a risk of this form. Formerly, such inequalities where derived under very restrictive dependent assumptions, see [2, 26, 3]. It is due to the use of the cumulative predictive risk. It is the correct criteria to assert the accuracy of predictive online algorithms as it coincides with the cumulative loss for deterministic observations and with the classical risk $R(f) = \mathbb{E}[\ell(Y, f(X))]$ for iid observations (where we can suppress the index t). Moreover, it appears naturally when using the minimax theory approach, see [1]. However, up to our knowledge, it is the first time that the cumulative predictive risk is used to compare online procedures with deterministic aggregation procedures. There is no warranty of the optimality of such results as lower bounds with similar random rates $\mathbb{E}_\pi \left[\sqrt{V_{j,n+1}} \right] \sqrt{\log M}$ are unknown.

The optimality of the BOA procedure is provided in a more restrictive context of iid observations when the online learners of the dictionary are constants: $f_{j,t} = f_j$, $t \geq 0$. In such context, applying Jensen's inequality we always have

$$\mathbb{E}\Big[\ell\Big(Y, \frac{1}{n+1}\sum_{t=1}^{n+1}\mathbb{E}_{\pi_{t-1}}[f_j](X)\Big)\Big] \leq R_n(\hat{f}) = \frac{1}{n+1}\sum_{t=1}^{n+1}\mathbb{E}[\ell(Y_t, \mathbb{E}_{\pi_{t-1}}[f_{j,t-1}](X_t))] \mid \mathcal{F}_{t-1}]$$

Then the batch conversion of BOA $\tilde{f} = (n+1)^{-1}\sum_{t=1}^{n+1}\mathbb{E}_{\pi_{t-1}}[f_{j'}]$ is always preferable. When the loss is Lipschitz continuous and strongly convex, we obtain an inequality in deviation; with probability $1 - e^{-x}$, $x > 0$, we have

$$R(\tilde{f}) \leq \min_{1\leq j\leq M} R(f_j) + C\frac{\log M + x + x^2}{n+1}.$$

The fast rate $\log M/(n+1)$ is optimal, see [27]. Notice that the proxy of the variance is necessary; without it, the BOA procedure coincides with the Progressive Mixture Rule of [8] that is optimal in expectation [18]. However, this procedure is suboptimal in deviation [5]. Thus, the stabilization term in the BOA procedure is necessary to control the deviations of exponential weights algorithms. There is few other optimal procedures in deviation in this iid context. The procedures in [5, 20] achieve the optimal rate using some prior information on the dictionary. In the Q-aggregation procedure of [21] as in the BOA procedure, no such extra-information is required. A priori, the Q-aggregation procedure is less explicit: it requires to calibrate an extra parameter and to optimize a non regular criteria. These practical issues have been solved in the context of quadratic loss with gaussian noise in [12]. On the opposite, the BOA procedure is explicit and fully adaptive in a general context.

2 Variational formula and Bernstein's inequalities

Classically, Bernstein's inequality is derived from an estimate of the Laplace transform and the Chernoff device. In order to derive empirical Bernstein's inequality of Theorem 1.1, we prefer to use another approach originally developed by Marton [24] and based on the variational formula of the entropy:

Lemma 2.1 (Donsker-Varadhan [13] variational formula)**.** *For any probability measures P on $\mathcal{X}$ and any measurable function $h : \mathcal{X} \to \mathbb{R}$ we have:*

$$\mathbb{E}_P[\exp(h - \mathbb{E}_P[h])] \leq 1 \quad \Longleftrightarrow \quad \mathbb{E}_Q[h] - \mathbb{E}_P[h] \leq \mathcal{K}(Q, P), \qquad \forall Q \tag{4}$$

that corresponds to the Gibbs measure $\mathbb{E}_P[e^h]dQ = e^h dP$.

That the Gibbs measure realizes the dual identity is at the core of the PAC-bayesian approach and proofs of optimality of exponential aggregation, see [9]. The novelty of the paper is to systematically consider the variational form of the Laplace transform to

linearize the concept of concentration of measures. In the following, the concentration of a measure P is expressed through the transport problem of its mass to any measure Q, see [29] for details and applications in mathematical statistics. The starting point of the proof of the empirical Bernstein inequality of Theorem 1.1 is the following Lemma

Lemma 2.2. *For any measures P and Q, for any random variable X the following relation holds*

$$\mathbb{E}_Q[X] \leq \mathbb{E}_P[X] + \sqrt{2(\mathbb{E}_Q[X_+^2] + \mathbb{E}_P[X_-^2])\mathcal{K}(Q,P)}.$$

Proof. By Young's inequality, it is equivalent that for any $\lambda > 0$ we have

$$\mathbb{E}_Q[X] \leq \mathbb{E}_P[X] + \lambda(\mathbb{E}_Q[X_+^2] + \mathbb{E}_P[X_-^2])/2 + \frac{\mathcal{K}(Q,P)}{\lambda}. \tag{5}$$

Multiplying this inequality by $\lambda > 0$ we obtain

$$\mathbb{E}_Q[\lambda(X - \mathbb{E}_P[X]) - \lambda^2(X_+^2 + \mathbb{E}_P[X_-^2])/2] \leq \mathcal{K}(Q,P).$$

By the variational form of the entropy, it is equivalent that the inequality holds for Q satisfying

$$\frac{dQ}{dP} = \frac{\exp(\lambda(X - \mathbb{E}_P[X]) - \lambda^2(X_+^2 + \mathbb{E}_P[X_-^2])/2)}{\mathbb{E}_P[\exp(\lambda(X - \mathbb{E}_P[X]) - \lambda^2(X_+^2 + \mathbb{E}_P[X_-^2])/2)]}.$$

We then obtain the dual form of the result as

$$\mathbb{E}_P[\exp(\lambda X - \lambda^2 X_+^2/2)] \leq \exp(\lambda \mathbb{E}_P[X] + \lambda^2 \mathbb{E}_P[X_-^2/2]).$$

This last inequality holds as for any real number $x \in \mathbb{R}$ we have the relation $\exp(x - x_+^2/2) \leq 1 + x + x_-^2/2$. □

Now we are ready to prove an exponential inequality of a random variable similar to the Bernstein's one with, instead of the variance, its own square.

Theorem 2.3. *Let X be any random variable such that $\mathbb{E}_P[X_-^4] \leq \mathbb{E}_P[X_-^2]$, then*

$$\mathbb{E}_P\Big[\exp\Big(\lambda(X - \mathbb{E}_P[X]) - \frac{\lambda^2}{2(1 - 7\lambda/4)}X^2\Big)\Big] \leq 1, \qquad 0 < \lambda < 4/7.$$

Proof. From the previous Lemma 2.2 applied to the non positive random variable $-X_-^2$ we obtain

$$\mathbb{E}_P[X_-^2] \leq \mathbb{E}_Q[X_-^2] + \sqrt{2\mathbb{E}_P[X_-^4]\mathcal{K}(Q,P)}.$$

By assumption, we obtain the estimate

$$\mathbb{E}_P[X_-^2] \leq \mathbb{E}_Q[X_-^2] + \sqrt{2\mathbb{E}_P[X_-^2]\mathcal{K}(Q,P)}.$$

By standard computation, using the Young inequality, we derive that for any $\lambda > 0$

$$\mathbb{E}_P[X_-^2] \leq 4^{-1}(\sqrt{2\mathcal{K}(Q,P)} + \sqrt{2\mathcal{K}(Q,P) + 4\mathbb{E}_Q[X_-^2]})^2 \leq \mathbb{E}_Q[X_-^2]\Big(1 + \frac{\lambda}{2}\Big) + \mathcal{K}(Q,P)\Big(1 + \frac{1}{2\lambda}\Big).$$

Plugging this estimate in the inequality (5), we obtain

$$\mathbb{E}_Q[X] - \mathbb{E}_P[X] \leq \frac{\lambda(1+\lambda/2)}{2}\mathbb{E}_Q[X^2] + \Big(\frac{2+\lambda^2}{2\lambda} + \frac{1}{4}\Big)\mathcal{K}(Q,P)$$

For $\lambda < 2$, we have $1 + \lambda/2 \leq (1-\lambda/2)^{-1}$ and $2 + \lambda^2 \leq 2(1-\lambda/2) + 3\lambda$. Thus, denoting $\eta = \lambda/(1-\lambda/2)$ we obtain

$$\mathbb{E}_Q[X] - \mathbb{E}_P[X] \leq \frac{\eta}{2}\mathbb{E}_Q[X^2] + \Big(\frac{1}{\eta} + \frac{7}{4}\Big)\mathcal{K}(Q,P), \qquad \eta > 0. \tag{6}$$

Using the variational form of the entropy we obtain

$$\mathbb{E}_P\Big[\exp\Big(\frac{4\eta}{4+7\eta}(X - \mathbb{E}_P[X]) - \frac{4\eta^2}{2(4+7\eta)}X^2\Big)\Big] \leq 1.$$

The desired result follows considering $\lambda = 4\eta/(4+7\eta)$. □

We are now ready to prove the empirical Bernstein inequality for martingales of Theorem 1.1. It follows the exponential inequality 2.3 by an application of classical submartingale arguments of [14]. Below, we detail another proof that uses only simple algebra and the decomposition of the entropy as the reasoning will be used in the proof of Theorem **??**:

Proof. of Theorem 1.1. We apply (6) to $P = P_t$, the distribution of δM_t and Q_t conditionally on $\mathcal{F}_{t-1}$

$$\mathbb{E}_{Q_t}[\Delta M_t] - \mathbb{E}_{P_t}[\Delta M_t] \leq \frac{\eta}{2}\mathbb{E}_{Q_t}[\Delta M_t^2] + \Big(\frac{1}{\eta} + \frac{7}{4}\Big)\mathcal{K}(Q_t,P_t).$$

As $\mathbb{E}_{P_t}[\Delta M_t] = 0$ by assumption, summing up for $1 \leq t \leq \tau$, we obtain:

$$\sum_{t=1}^{\tau}\mathbb{E}_{Q_t}[\Delta M_t] \leq \frac{\eta}{2}\sum_{t=1}^{\tau}\mathbb{E}_{Q_t}[\Delta M_t^2] + \Big(\frac{1}{\eta} + \frac{7}{4}\Big)\sum_{t=1}^{\tau}\mathcal{K}(Q_t,P_t).$$

Integrating with respect to Q, remarking that $\mathbb{E}_Q[\sum_{t=1}^{\tau}\mathcal{K}(Q_t,P_t)] = \mathcal{K}(Q,P)$ for P the distribution of M_τ and the decomposition of the entropy

$$\mathbb{E}_Q\Big[\sum_{t=1}^{\tau}\mathcal{K}(Q_t,P_t)\Big] = \mathbb{E}_Q\Big[\sum_{t=1}^{\tau}\log(dQ_t/dP_t)\Big] = \mathbb{E}_Q\Big[\log\Big(\frac{dQ_1\cdots dQ_\tau}{dP_1\cdots dP_\tau}\Big)\Big] = \mathcal{K}(Q,P) \tag{7}$$

we obtain

$$\mathbb{E}_Q[M_\tau] \leq \frac{\eta}{2}\mathbb{E}_Q[[M]_\tau] + \Big(\frac{1}{\eta} + \frac{7}{4}\Big)\mathcal{K}(Q,P).$$

Now we consider Q as restriction of P to the event

$$A = \Big\{M_\tau \geq \frac{\eta}{2}[M]_\tau + \Big(\frac{1}{\eta} + \frac{7}{4}\Big)x, \quad \eta > 0\Big\} \supseteq \Big\{M_\tau \geq \sqrt{2[M]_\tau x} + 7x/4\Big\}.$$

Then $\mathcal{K}(Q,P) = \log(1/P(A)) \geq x$ and the desired result follows. □

3 Second order bounds for the regret

3.1 First regret bound and link with the individual sequences framework

We work first conditionally on $\mathcal{F}_n$; it is the deterministic setting where $(X_t, Y_t) = (x_t, y_t)$ are provided recursively for $1 \leq t \leq n$. In this case, we consider the cumulative loss $\mathcal{R}(f)$ for any online learner $f = (f_0, f_1, f_2, \ldots)$ to assert the accuracy of the prediction. We focus on convex losses and then the sub gradient trick is useful to compare the BOA procedure with the best convex mixture of experts f_π rather than with the best expert. We estimate the regret of the BOA procedure with respect to the best possible deterministic aggregation of the online learners of the dictionary:

Theorem 3.1. *Assume that $\eta > 0$ satisfies*

$$\eta \max_{1\leq t\leq n+1} \max_{1\leq j\leq M} \ell_{j,t_+} \leq 1. \tag{8}$$

The cumulative loss of the BOA procedure satisfies

$$\mathcal{R}(\hat{f}) \leq \min_{\pi} \Big\{ \mathcal{R}(f_\pi) + 2\eta \sum_{t=0}^{n} \mathbb{E}_\pi[\ell_{j,t+1}^2] + 2\frac{\mathcal{K}(\pi, \pi_0)}{\eta} \Big\}.$$

Proof. We prove Theorem 3.1 with an application of the entropy based method described in the previous Section. A substantial advantage of this approach is also to simplify the proofs as it linearizes the concentration problem.

First, we study the concentration properties of the aggregation procedure. We work sequentially, conditionally on the observations of the sample $\mathcal{F}_n$. For any $1 \leq j \leq M$, $0 \leq t \leq n$, under (8),we have

$$(\eta \ell_{j,t_+})^4 \leq (\eta \ell_{j,t_+})^2, \qquad j \in \{1, \ldots, M\}.$$

We denote π_t the measure on the index space $\{1, \ldots, M\}$ such that $\pi_t(j) = \pi_{t,j}$, $1 \leq j \leq M$. We apply the transport inequality (6) to $-\eta \ell_{j,t+1}$ for $P = \pi_t$ and $Q = \pi_{t+1}$, two measures on $\{1, \ldots, M\}$. Taking $\eta = 4$ in (6) we obtain

$$\mathbb{E}_{\pi_t}[\eta \ell_{j,t+1}] \leq \mathbb{E}_{\pi_{t+1}}[\eta \ell_{j,t+1}] + 2\mathbb{E}_{\pi_{t+1}}[\eta^2 \ell_{j,t+1}^2] + 2\mathcal{K}(\pi_{t+1}, \pi_t).$$

By convexity, we can apply Jensen's inequality and we have $\mathbb{E}_{\pi_t}[\eta \ell_{j,t+1}] \geq 0$. By definition of the Kullback-Leibler divergence, we have

$$0 \leq \mathbb{E}_{\pi_{t+1}}\Big[2^{-1}\eta \ell_{j,t+1} + \eta^2 \ell_{j,t+1}^2 + \log(d\pi_{t+1}/d\pi_t) \Big]. \tag{9}$$

By the specific form of (π_t), we have for any $t \geq 1$

$$\begin{aligned}
2^{-1}\eta \ell_{j,t+1} + \eta^2 \ell_{j,t+1}^2 + \log(d\pi_{t+1}/d\pi_t) &= -\log\Big(\mathbb{E}_{\pi_t}\Big[\exp\Big(2^{-1}\eta \ell_{j,t+1} + \eta^2 \ell_{j,t+1}^2 \Big) \Big] \Big) \\
&= \log\Big(\mathbb{E}_{\pi_0}\Big[\exp\Big(\sum_{s=0}^{t-1} 2^{-1}\eta \ell_{s+1}(j) + \eta^2 \ell_{s+1}^2(j) \Big) \Big] \Big) \\
&\quad - \log\Big(\mathbb{E}_{\pi_0}\Big[\exp\Big(\sum_{s=0}^{t} 2^{-1}\eta \ell_{s+1}(j) + \eta^2 \ell_{s+1}^2(j) \Big) \Big] \Big).
\end{aligned}$$

Summing up for $t = 0, \dots, n$ we obtain

$$0 \leq -\log\Big(\mathbb{E}_{\pi_0}\Big[\exp\Big(\sum_{t=0}^{n} 2^{-1}\eta\ell_{j,t+1} + \eta^2\ell_{j,t+1}^2\Big)\Big]\Big).$$

Using the variational form of the entropy (4) we have

$$0 \leq \inf_{\pi}\Big\{\mathbb{E}_{\pi}\Big[\sum_{t=0}^{n} \eta\ell_{j,t+1} + 2\eta^2\ell_{j,t+1}^2 + 2\mathcal{K}(\pi,\pi_0)\Big]\Big\}.$$

In the sequel, we denote

$$\mathbb{E}_{\hat{\pi}}[\mathcal{R}(f_j)] = \sum_{t=1}^{n+1} \mathbb{E}_{\pi_t}[\ell(Y_t, f_{j,t-1}(X_t))]. \tag{10}$$

We derive the desired result using the classical sub-gradient trick, i.e. noticing that

$$\begin{aligned}
\mathbb{E}_{\hat{\pi}}[\mathcal{R}(f_j)] - \mathcal{R}(f_\pi) &\leq \sum_{t=0}^{n} \mathbb{E}_{\pi_t}[\ell(Y_{t+1}, f_{j',t}(X_{t+1}))] - \mathbb{E}_{\pi}[\ell(Y_{t+1}, f_{j,t}(X_{t+1}))] \\
&\leq \mathbb{E}_{\pi}\Big[\sum_{t=0}^{n} \mathbb{E}_{\pi_t}[\ell'(Y_{t+1}, f_{j',t}(X_{t+1}))(f_{j',t}(X_{t+1}) - f_{j,t}(X_{t+1}))]\Big] \\
&\leq -\mathbb{E}_{\pi}\Big[\sum_{t=0}^{n} \ell_{j,t}\Big].
\end{aligned}$$

We conclude applying the Jensen's inequality $\mathcal{R}(\hat{f}) \leq \mathbb{E}_{\hat{\pi}}[\mathcal{R}(f_j)]$. □

In the upper bound, we call proxy of the variance the term

$$\sum_{t=0}^{n} \mathbb{E}_{\pi}[\ell_{j,t+1}^2] \leq \sum_{t=0}^{n} \mathbb{E}_{\pi_t}[\ell'(Y_{t+1}, f_{j,t}(X_t))^2 \mathbb{E}_{\pi}[(f_{j,t}(X_t) - f_{j',t}(X_t))^2]].$$

This proxy can be small because the sub-gradient is small or because the aggregation strategy π is close to the BOA strategy. We will see at the end of Section 3.4 that the square of the sub-gradient is small because it is proportional to the loss when it is quadratic.

The first application of Theorem 3.1 is the context of individual sequences prediction [10]. We can consider that $Y_t = y_t$ for a deterministic sequence $y_0, \dots, y_n$. We have $\mathcal{F}_t = \{y_0, \dots, y_t\}$, $0 \leq t \leq n$, and the online learners $f_j = (y_{j,1}, y_{j,2}, y_{j,3}, \dots)$ of the dictionary are called the experts. The regret is now $\mathcal{R}(\hat{f}) = \sum_{t=1}^{n+1} \ell(y_t, \hat{y}_t)$ for learner predictions $\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)$ where $\hat{y}_t = \hat{f}_{t-1} = \sum_{j=1}^{M} \pi_{j,t-1} y_{j,t}$ where $\pi_{j,t-1}$ are measurable functions of the past $\{y_0, \dots, y_{t-1}\}$. The estimate obtained in Theorem 3.1 is called a second order bound after the seminal paper [11].

3.2 A new adaptive method for exponential weights

From Theorem 3.1, it is tempting to optimize the second order bound of the regret with respect to η:

$$\eta^* = \Big\{ \frac{1}{\max_{1\le t\le n+1} \max_{1\le j\le M} \ell_{j,t_+}}, \sqrt{\frac{\mathcal{K}(\pi,\pi_0)}{V_{j,n+1}}} \Big\},$$

where $V_{j,n+1} = \sum_{t=0}^{n} \ell_{j,t+1}^2$, to obtain the regret bound

$$\mathcal{R}(\hat{f}) \le \min_{\pi} \Big\{ \mathcal{R}(f_\pi) + 4\mathbb{E}_\pi \Big[\sqrt{V_{j,n+1}} \Big] \sqrt{\mathcal{K}(\pi,\pi_0)} \Big\}.$$

However, in practice, the optimal measure π is unknown and the term $\mathcal{K}(\pi,\pi_0)$ is not explicit and thus also η^*. Moreover, the resulting BOA procedure will not be recursive as η^* depends on the observations (X_t, Y_t) through $\ell_{j,t}$, $1 \le t \le n+1$. It is possible to adapt the BOA procedure by tuning the inverse temperature parameter η recursively with respect to the observations. We described in Figure 2 the adaptive version of the BOA algorithm. Notice that the adaptation of the exponential weights

$$\pi_{j,t} = \frac{\eta_{j,t} \exp(-2^{-1}\eta_{j,t} L_{j,t}) \pi_{j,0}}{\sum_{j=1}^{M} \eta_{j,t} \exp(-2^{-1}\eta_{j,t} L_{j,t}) \pi_{j,0}},$$

is new as the learning rates $\eta_{j,t}$ depends on the element of the dictionary j and appear into the exponential and as a factor. Adaptative procedures with such multiplicative forms have been studied in [15] to solve the impossible tuning problem, but with weights that are not exponential. Notice that the adaptive weights are only well defined when the learning rates are positive: $\eta_{j,t} > 0$, $1 \le j \le M$, $0 \le t \le n$ and that such multiplicative adaptive form can be investigated for other exponential weights than for those of BOA. We obtain the regret of this procedure:

Theorem 3.2. *If the learning rates are non increasing and satisfy*

$$\eta_{j,t-1} \ell_{j,t_+} \le 1, \qquad 1 \le t \le n+1, \quad 1 \le J \le M, \tag{11}$$

then the cumulative loss of the adaptive BOA procedure satisfies

$$\mathcal{R}(\hat{f}) \le \min_{\pi} \Big\{ \mathcal{R}(f_\pi) + 2\mathbb{E}_\pi \Big[\sum_{t=0}^{n} \eta_{j,t} \ell_{j,t+1}^2 \Big] + \mathbb{E}_\pi \Big[\frac{2\log(\pi_{j,0}^{-1})}{\eta_{j,n}} \Big] + \mathbb{E}_\pi \Big[\frac{2}{\eta_{j,n}} \log \Big(1 + \sum_{t=1}^{n} \sum_{j=1}^{M} \frac{\pi_{j,0}}{e} \Big(\frac{\eta_{j,t-1}}{\eta_{j,t}} - 1 \Big) \Big) \Big] \Big\}.$$

Proof. We denote $\tilde{\pi}_{j,t}$ the weights satisfying

$$\tilde{\pi}_{j,t} = \frac{\exp(-2^{-1}\eta_{j,t} \tilde{L}_{j,t}) \pi_{j,0}}{\sum_{j=1}^{M} \exp(-2^{-1}\eta_{j,t} \tilde{L}_{j,t}) \pi_{j,0}},$$

for $\tilde{L}_{j,t} = \sum_{s=0}^{t} \ell_{j,s} + 2\eta_{j,s}\ell_{j,s}^2$, and $\tilde{\pi}_t$ the measure on $\{1,\dots,M\}$ such that $\tilde{\pi}_t(j) = \tilde{\pi}_{j,t}$. We use the same notation than in the proof of Theorem 3.1. Under (11) we apply the transport inequality (6) to $-\eta_{j,t}\ell_{j,t+1}$ for $P = \tilde{\pi}_n$ and Q the Dirac mass on $\{j\}$ for any $1 \le j \le M$. For $\eta = 4$ in (6) we obtain

$$\mathbb{E}_{\tilde{\pi}_n}[2^{-1}\eta_{j,t}\ell_{j,n+1}] \le 2^{-1}\eta_{j,n}\ell_{j,n+1} + \eta_{j,n}^2\ell_{j,n+1}^2 - \log(\tilde{\pi}_{j,n}).$$

We remark that by definition we have $\mathbb{E}_{\tilde{\pi}_n}[2^{-1}\eta_{j,n}\ell_{j,n+1}] = 0$,

$$2^{-1}\eta_{j,n}\ell_{j,n+1} + \eta_{j,n}^2\ell_{j,n+1}^2 = 2^{-1}\eta_{j,n}(\tilde{L}_{j,n+1} - \tilde{L}_{j,n}) \qquad \text{and}$$
$$-\log(\tilde{\pi}_{j,n}) = 2^{-1}\eta_{j,n}\tilde{L}_{j,n} + \log(\pi_{j,0}^{-1}) + \log\Big(\sum_{j=1}^{M}\tilde{\pi}_{j,n}\Big).$$

Combining these identities, we derive that for any $1 \le j \le M$,

$$0 \le 2^{-1}\eta_{j,n}\tilde{L}_{j,n+1} + \log(\pi_{j,0}^{-1}) + \log\Big(\sum_{j=1}^{M}\tilde{\pi}_{j,n}\Big).$$

To estimate the last term of the upper bound, we will prove that for all $1 \le t \le n$ we have

$$\sum_{j=1}^{M}\tilde{\pi}_{j,t} \le \sum_{j=1}^{M}\tilde{\pi}_{j,t-1} + \frac{1}{e}\Big(\sum_{j=1}^{M}\frac{\eta_{j,t-1}}{\eta_{j,t}}\pi_{j,0} - 1\Big) \tag{12}$$

We remark that for any $1 \le j \le M$

$$\frac{\tilde{\pi}_{j,t}}{\pi_{j,0}} = \exp(-2^{-1}\eta_{j,t}\tilde{L}_{j,t}) = \exp(-2^{-1}\eta_{j,t}\ell_{j,t} - \eta_{j,t}\eta_{j,t-1}\ell_{j,t}^2)\exp(-2^{-1}\eta_{j,t}\tilde{L}_{j,t-1}) \tag{13}$$
$$= (\exp(-2^{-1}\eta_{j,t-1}\ell_{j,t} - \eta_{j,t-1}^2\ell_{j,t}^2)\exp(-2^{-1}\eta_{j,t-1}\tilde{L}_{j,t-1}))^{\eta_{j,t}/\eta_{j,t-1}}.$$

As $\eta_{j,t}$ is non increasing with t for any j, we have that $\alpha = \eta_{j,t-1}/\eta_{j,t} \ge 1$. Then, following the reasoning in [15], we use the inequality $x \le x^\alpha + (\alpha-1)/e$ for any $x \ge 0$ and $\alpha \ge 1$ to derive that

$$\tilde{\pi}_{j,t} \le \exp(-2^{-1}\eta_{j,t-1}\ell_{j,t} - \eta_{j,t-1}^2\ell_{j,t}^2)\tilde{\pi}_{j,t-1} + \frac{1}{e}\Big(\frac{\eta_{j,t-1}}{\eta_{j,t}} - 1\Big)\pi_{j,0}$$

Summing up this bound for $1 \le j \le M$, using the fact that $\sum_{j=1}^{M}\pi_{j,0} = 1$, we obtain that

$$\sum_{j=1}^{M}\tilde{\pi}_{j,t} \le \sum_{j=1}^{M}\exp(-2^{-1}\eta_{j,t-1}\ell_{j,t} - \eta_{j,t-1}^2\ell_{j,t}^2)\tilde{\pi}_{j,t-1} + \frac{1}{e}\Big(\sum_{j=1}^{M}\frac{\eta_{j,t-1}}{\eta_{j,t}}\pi_{j,0} - 1\Big)$$

But we remark that

$$\frac{\sum_{j=1}^{M}\exp(-2^{-1}\eta_{j,t-1}\ell_{j,t} - \eta_{j,t-1}^2\ell_{j,t}^2)\tilde{\pi}_{j,t-1}}{\sum_{j=1}^{M}\tilde{\pi}_{j,t-1}} = \mathbb{E}_{\tilde{\pi}_{t-1}}[\exp(-2^{-1}\eta_{j,t-1}\ell_{j,t} - \eta_{j,t-1}^2\ell_{j,t}^2)]$$

and the inequality (12) follows from an application of Theorem 2.3 with $\lambda = 2^{-1}$, $P = \tilde{\pi}_{t-1}$ and $X = \eta_{j,t-1}\ell_{j,t}$ that is centered by definition of $\hat{f}_t$. Using recursively (12) and noticing that $\sum_{j=1}^M \tilde{\pi}_{j,0} = \sum_{j=1}^M \pi_{j,0} = 1$ we obtain

$$\log\Big(\sum_{j=1}^M \tilde{\pi}_{j,n}\Big) \leq \log\Big(1 + \sum_{t=1}^n \frac{1}{e}\Big(\sum_{j=1}^M \frac{\eta_{j,t-1}}{\eta_{j,t}}\pi_{j,0} - 1\Big)\Big).$$

Combining the obtained bounds, by definition of $L_{j,n+1}$, we have for any $1 \leq j \leq M$

$$\sum_{t=0}^n \mathbb{E}_t[\ell'(Y_{t+1}, f_{j',t}(X_{t+1}))f_{j',t}(X_{t+1})] \leq \sum_{t=0}^n \mathbb{E}_t[\ell'(Y_{t+1}, \hat{f}_{j',t}(X_{t+1}))]f_j(X_{t+1})$$
$$+ 2\sum_{t=0}^n \eta_{j,t}\ell_{j,t+1}^2 + \frac{2\log(\pi_{j,0}^{-1})}{\eta_{j,n}} + \frac{2}{\eta_{j,n}}\log\Big(1 + \sum_{t=1}^n\sum_{j=1}^M \frac{\pi_{j,0}}{e}\Big(\frac{\eta_{j,t-1}}{\eta_{j,t}} - 1\Big)\Big).$$

The minimum for $1 \leq j \leq M$ of this upper bound is equal to the linear optimization problem in the measure π on $\{1, \ldots, M\}$

$$\mathbb{E}_t[\ell'(Y_{t+1}, \hat{f}_{j',t}(X_{t+1}))]f_\pi(X_{t+1}) + 2\mathbb{E}_\pi\Big[\sum_{t=0}^n \eta_{j,t}\ell_{j,t+1}^2\Big] + \mathbb{E}_\pi\Big[\frac{2\log(\pi_{j,0}^{-1})}{\eta_{j,n}}\Big]$$
$$+ \mathbb{E}_\pi\Big[\frac{2}{\eta_{j,n}}\log\Big(1 + \sum_{t=1}^n\sum_{j=1}^M \frac{\pi_{j,0}}{e}\Big(\frac{\eta_{j,t-1}}{\eta_{j,t}} - 1\Big)\Big)\Big]$$

We conclude by the sub-gradient trick as in the proof of Theorem 3.1. □

Notice that the proof and the upper bound of Theorem 3.2 has a different flavor than those of Theorem 3.1. The proof of Theorem 3.1 is based on a recursive argument. It asserts the optimality of the exponential weights for such sequential transport problem via the variational formula of the entropy (4). The upper bound involves the Kullback-Leibler divergence $\mathcal{K}(\pi, \pi_0)$ as the proof relies on a transport problem to any measure π on $\{1, \ldots, M\}$. The proof of Theorem 3.2 is rougher in the sense that the transport problem is now restricted to Dirac measures on $\{1, \ldots, M\}$. We can still compare the accuracy of the procedure with the best deterministic aggregation of the experts because of linearity. However, we cannot assert the optimality of the adaptive weights with such multiplicative form and the upper bound involves $\mathbb{E}_\pi[\log(\pi_{j,0}^{-1})]$ that is a rough upper bound of $\mathcal{K}(\pi, \pi_0)$. Finally notice that classical adaptive exponential procedures involve learning rates that do not depend on $\{j\}$ and thus the multiplicative form disappears:

$$\pi_{j,t} = \frac{\exp(-2^{-1}\eta_t L_{j,t})\pi_{j,0}}{\sum_{j=1}^M \exp(-2^{-1}\eta_t L_{j,t})\pi_{j,0}}.$$

Then a recursive argument similar to the proof of Theorem 3.1 can be used (see the Appendix D of the preliminary version of [6] available on arXiv:math/0703854). It assert

the optimality of such adaptive procedure via the variational formula of the entropy (4). Then we conjecture that the impossible tuning can be solved only by relaxing the transport problem to Dirac measures only and the regret bounds with a complexity term larger than $\mathcal{K}(\pi, \pi_0)$.

3.3 The adaptive BOA procedure when the range is known

First consider the case where the effective range of the linearized error is known: it exists $E \geq 1$ such that $|\ell_{j,t}| \leq E$, $1 \leq t \leq n+1$, $1 \leq J \leq M$. We tune the learning rates in the following way

$$\eta_{j,t} = \min\left\{\frac{1}{E}, \sqrt{\frac{\log(M)}{\sum_{s=1}^{t} \ell_{j,s}^2}}\right\}, \qquad t \geq 0. \tag{14}$$

The learning rates are similar than those of Section 4.1 in [11] except that they depend on j through the quadratic variation proxy $\sum_{s=1}^{t} \ell_{j,s}^2$. We restrict to the cases where $M > 1$ to consider only positive learning rates $\eta_{j,t} > 0$. Notice that here the constants in the bound should be sharper if adding a multiplicative constant to the quadratic variation proxy $\sum_{s=1}^{t} \ell_{j,s}^2$ as in [11].

Theorem 3.3. *If $|\ell_{j,t}| \leq E$, $1 \leq t \leq n+1$, $1 \leq J \leq M$ ($M > 1$) and the learning rates are tuned as in* (14) *then the adaptive BOA procedure achieves, for all $n \geq 1$,*

$$\mathcal{R}(\hat{f}) \leq \min_{\pi}\left\{\mathcal{R}(f_\pi) + 2\mathbb{E}_\pi\left[\sqrt{V_{j,n+1}}\right]\left(\frac{\sqrt{2\log M}}{\sqrt{2}-1} + \frac{B_{n,E}}{\sqrt{\log M}}\right)\right\} + E(2\log M + 2B_{n,E} + 1),$$

where $V_{j,n+1} = \sum_{t=0}^{n} \ell_{j,t+1}^2$ and $B_{n,E} = \log\left(1 + \frac{E(E+1)}{e\sqrt{\log(M)}} + \frac{\log n}{2e}\right)$.

Proof. We estimate

$$\log\left(1 + \sum_{t=1}^{n}\sum_{j=1}^{M} \frac{\pi_{j,0}}{e}\left(\frac{\eta_{j,t-1}}{\eta_{j,t}} - 1\right)\right) \leq B_{n,E}.$$

Using that $\sqrt{1+x} - 1 \leq x/2$, $x > 0$, we have

$$\begin{aligned}
\sum_{t=1}^{n}\left(\frac{\eta_{j,t-1}}{\eta_{j,t}} - 1\right) &\leq \frac{|\ell_{j,1}|E}{\sqrt{\log(M)}} - 1 + \sum_{t=2}^{n}\left(\sqrt{\frac{\sum_{s=1}^{t}\ell_{j,s}^2}{\max\{\sum_{s=1}^{t-1}\ell_{j,s}^2, E\sqrt{\log(M)}\}}} - 1\right) \\
&\leq \frac{|\ell_{j,1}|E}{\sqrt{\log(M)}} - 1 + \sum_{t=2}^{n}\left(\sqrt{1 + \frac{\ell_{j,t}^2}{\max\{\sum_{s=1}^{t-1}\ell_{j,s}^2, E\sqrt{\log(M)}\}}} - 1\right) \\
&\leq \frac{E^2}{\sqrt{\log(M)}} - 1 + \frac{1}{2}\sum_{t=2}^{n}\frac{\ell_{j,t}^2}{\max\{\sum_{s=1}^{t-1}\ell_{j,s}^2, E\sqrt{\log(M)}\}}.
\end{aligned}$$

Then we use similar arguments than in the proof of Theorem 5 of [11]; We denote by T the first time that $\sum_{s=1}^{t} \ell_{j,s}^2 > E^2$. Because $\eta_{j,T}^2 \leq E^2$ we obtain

$$\sum_{t=2}^{n} \frac{\ell_{j,t}^2}{\max\{\sum_{s=1}^{t-1} \ell_{j,s}^2, E\sqrt{\log(M)}\}} \leq \frac{2E}{\sqrt{\log(M)}} + \sum_{t=T+1}^{n} \frac{\ell_{j,t}^2}{\sum_{s=1}^{t-1} \ell_{j,s}^2}.$$

We use the Lemma 14 of [15] with $a_i = \ell_{j,T+i}^2/E^2$, $i \geq 1$, $a_0 = \sum_{s=1}^{T} \ell_{j,s}^2/E^2 > 1$ and $f(x) = 1/x$. We obtain

$$\sum_{t=T+1}^{n} \frac{\ell_{j,t}^2}{\sum_{s=1}^{t-1} \ell_{j,s}^2} \leq 1 + \log\Big(\sum_{t=1}^{n} \ell_{j,t}^2/E^2\Big)_+ \leq 1 + \log n.$$

We conclude the proof of Theorem similarly than the conclusion of the proof of Theorem 5 in [11]. □

3.4 The adaptive BOA procedure when the range is unknown

When the effective range of the linearized error is not known, we have to estimate it. To adapt the reasoning of [11], we consider the same kind of estimator E_t of the range: $E_t = 2^k$ where $k \in \mathbb{N}$ is the smallest integer such that $\max_{1\leq s\leq t} \max_{1\leq j\leq M} |\ell_{j,t}| \leq 2^k$. Then we define the learning rates as

$$\eta_{j,t} = \min\left\{\frac{1}{E_t}, \sqrt{\frac{\log M}{\sum_{s=1}^{t} \ell_{j,s}^2}}\right\}, \qquad t \geq 0. \tag{15}$$

This rule for updating learning rates is similar than the one in [11] except that it depends on j and that $E_t \geq 1$.

Theorem 3.4. *If $|\ell_{j,t}| \leq E$ with $E \geq 1$, $1 \leq t \leq n+1$, $1 \leq J \leq M$ ($M > 1$) and the learning rates are tuned as in* (15) *then the adaptive BOA procedure achieves, for all $n \geq 1$,*

$$\mathcal{R}(\hat{f}) \leq \min_{\pi}\Big\{\mathcal{R}(f_\pi) + 2\mathbb{E}_\pi\Big[\sqrt{V_{j,n+1}}\Big]\Big(\frac{\sqrt{2\log M}}{\sqrt{2}-1} + \frac{\tilde{B}_{n,E}}{\sqrt{\log M}}\Big)\Big\} + 4E(\log M + \tilde{B}_{n,E} + 1),$$

where $V_{j,n+1} = \sum_{t=0}^{n} \ell_{j,t+1}^2$ and $\tilde{B}_{n,E} = \log\Big(1 + \frac{E(E+1)}{e\sqrt{\log M}} + \frac{\log n}{2e}\Big) + \log E$.

Proof. With no loss of generality, we can assume that $\max_{1\leq j\leq M} |\eta_n \ell_{j,n+1}| \leq 1$. Then we can apply the same reasoning than in the proof of Theorem 3.2 and we have to estimate the term $\log\Big(\sum_{j=1}^{M} \tilde{\pi}_{j,n}\Big) \leq \tilde{B}_{n,E}$. We have to distinguish two cases.

First, we consider the set of indices that $\mathcal{T} = \{t_1, \dots t_R\}$ such that $E_{t_r-1} < E_{t_r}$. Then, we use the identity (13) to derive, as $\eta_{j,t-1} \geq \eta_{j,t}$, the inequality

$$\sum_{j=1}^{M} \frac{\tilde{\pi}_{j,t}}{\pi_{j,0}} \leq \sum_{j=1}^{M} \exp(-2^{-1}\eta_{j,t}\ell_{j,t} - \eta_{j,t}^2\ell_{j,t}^2) \exp(-2^{-1}\eta_{j,t}\tilde{L}_{j,t-1}).$$

Then we apply Theorem 2.3 with $\lambda = 2^{-1}$, $P(\{j\}) \propto \exp(-2^{-1}\eta_{j,t}\tilde{L}_{j,t-1})$ and $X = \eta_{j,t}\ell_{j,t}$ that is bounded by 1 but not centered to estimate the upper bound by

$$\exp(2^{-1}\mathbb{E}_P[\eta_{j,t}\ell_{j,t}]) \sum_{j=1}^{M} \exp(-2^{-1}\eta_{j,t}\tilde{L}_{j,t-1}).$$

But $\eta_{j,t}\ell_{j,t} \leq 1$ and thus for $t \in \mathcal{T}$, following the same reasoning than in the proof of Theorem 3.2, we obtain

$$\sum_{j=1}^{M} \tilde{\pi}_{j,t} \leq \sqrt{e}\Big(\sum_{j=1}^{M} \tilde{\pi}_{j,t-1} + \frac{1}{e}\Big(\sum_{j=1}^{M} \frac{\eta_{j,t-1}}{\eta_{j,t}} \pi_{j,0} - 1 \Big) \Big) \tag{16}$$

Second, we consider the the set of the indices t that do not belong to $\mathcal{T}$ and such that

$$E_t = E_{t_{r+1}} \qquad \text{and} \qquad \max_{1 \leq j \leq M} |\eta_{t-1}\ell_{j,t}| \leq 1, \qquad t \notin \mathcal{T}. \tag{17}$$

Then the same reasoning than in the proof of Theorem 2.3 apply and the recursive formula (12) holds.

To conclude, we apply recursive formulas (12) and (12) and we obtain the upper bound

$$\sum_{j=1}^{M} \tilde{\pi}_{j,n} \leq e^{R/2}\Big(1 + \sum_{t=1}^{n} \frac{1}{e}\Big(\sum_{j=1}^{M} \frac{\eta_{j,t-1}}{\eta_{j,t}} \pi_{j,0} - 1 \Big) \Big).$$

Thus, the logarithm $\log(\sum_{j=1}^{M} \tilde{\pi}_{j,n})$ is bounded $B_{n,E}$ and an additional term smaller than $R/2 \leq \lceil (\log_2 E)_+ \rceil / 2 \leq \log E$ and the Theorem is proved replacing $B_{n,E}$ with $\tilde{B}_{n,E}$ and using similar arguments than in the proof of Theorem 6 in [11]. □

The second order estimates provided in Theorem 3.4 is optimal in the sense that we solve the impossible tuning problem described in [15]. The advantage of the adaptive BOA procedure compared with the procedures studied in [15] s that it also adapts to the unknown range $E \geq 1$ of the linearized loss.

The bound on the regret obtained in Theorem 3.4 provides a confident interval for the excess loss in term of a proxy of the variance

$$V_{j,n+1} \leq \sum_{t=1}^{n+1} \mathbb{E}_\pi[\mathbb{E}_{\pi_t}[(\ell'(Y_t, f_{j,t-1}(X_t)))^2 (f_{j,t-1}(X_t) - f_{j',t-1}(X_t))^2]].$$

Let us give an example where this proxy is small and where . For the square loss $\ell(y, f(x)) = (y - f(x))^2$, we have $\ell'(y, f(x))^2 \leq 4\ell(y, f(x))$ and thus if $|f_{j,t-1}(X_t) - \hat{f}_{t-1}(X_t)| \leq b$ for $b > 0$ and any $1 \leq j \leq M$ and $1 \leq t \leq n+1$, using the notation of (10), we have

$$V_{j,n+1} \leq 4b^2 \mathbb{E}_{\hat{\pi}}[\mathcal{R}(f_j)].$$

Thus, abusively considering $\log\log n$ as a constant, it exists a constant $C > 0$ such that

$$\begin{aligned} 0 &\leq \sum_{t=1}^{n+1} \ell_{j,t} + 2b\sqrt{C\mathbb{E}_{\hat{\pi}}[\mathcal{R}(f_j)]\log M} + CE\log M, \\ &\leq \sum_{t=1}^{n+1} \ell_{j,t} + \eta\mathbb{E}_{\hat{\pi}}[\mathcal{R}(f_j)] + \frac{Cb^2\log M}{\eta} + CE\log M, \end{aligned}$$

for any $\eta > 0$. Then, the minimum in j of $\sum_{t=1}^{n+1}\ell_{j,t}$ coincides with the minimum in π of $\mathbb{E}_\pi[\sum_{t=1}^{n+1}\ell_{j,t}]$. By the sub-gradient trick $\mathbb{E}_{\hat{\pi}}[\mathcal{R}(f_j)] - \mathcal{R}(f_\pi) \leq -\mathbb{E}_\pi[\sum_{t=1}^{n+1}\ell_{j,t}]$ we obtain

$$(1-\eta)\mathbb{E}_{\hat{\pi}}[\mathcal{R}(f_j)] \leq \min_\pi \mathcal{R}(f_\pi) + \frac{Cb^2\log M}{\eta} + CE\log M.$$

Now we remark that as ℓ is the quadratic loss we have the decomposition

$$\mathbb{E}_{\hat{\pi}}[\mathcal{R}(f_j)] = \mathcal{R}(f_{\hat{\pi}}) + \sum_{t=0}^{n}\mathbb{E}_{\pi_t}[(f_{j,t}(X_t) - \hat{f}_t(X_t))^2].$$

As the Young inequality holds for any $\eta > 0$, if

$$\min_\pi \mathcal{R}(f_\pi) \leq (1-\eta^*)\min_{1\leq j\leq M}\mathcal{R}(f_j) + \sum_{t=0}^{n}\mathbb{E}_{\pi_t}[(f_{j,t}(X_t) - \hat{f}_t(X_t))^2]$$

for sufficiently small η^*, we obtain the regret bound

$$\mathcal{R}(\hat{f}) \leq \mathbb{E}_{\hat{\pi}}[\mathcal{R}(f_j)] \leq \min_{1\leq j\leq M}\mathcal{R}(f_j) + \frac{C\log(M)}{1-\eta^*}\Big(E + \frac{b^2}{\eta^*}\Big).$$

The rate is optimal [16] but the constant (not explicit here) are certainly not, see [28]. Notice that the stabilization term in the BOA procedure can be avoided as the simpler Exponential Averaging algorithm of [23] satisfies the inequality with a better constant in situation of exp-concavity. It is natural as the regret is not defined as the expectation of the loss. Only the accuracy of the learner predictions are taken into account by the regret criterion. The exp-concavity of the square loss on compact sets is enough to assert optimal procedure without any second order bounds. However, the BOA procedure does not depend on the exp-concavity properties of the quadratic loss and thus is adaptive to the unknown range which is not the case of the EA algorithm of [28, 23].

4 Optimality of the BOA procedure in a stochastic environment

4.1 Confidence interval on the cumulative predictive risk

We now turn to a stochastic setting where (X_t, Y_t) are random elements observed recursively with $1 \leq t \leq n$. The main motivation the introduction the BOA procedure the

presence of a proxy of the quadratic variation that is calibrated to extend the second order bounds on the cumulative error to the cumulative predictive risk. We are now ready to state the main result of the paper:

Theorem 4.1. *If $|\ell_{j,t}| \leq E$ with $E \geq 1$, $1 \leq t \leq n+1$, $1 \leq j \leq M$ ($M > 1$) and the learning rates are tuned as in* (15) *then the adaptive BOA procedure achieves, for all $n \geq 1$ and with probability $1 - e^{-x}$,*

$$R_n(\hat{f}) \leq \min_{\pi} \Big\{ R_n(f_\pi) + \frac{2\mathbb{E}_\pi[\sqrt{V_{j,n+1}}]}{n+1} \Big((\sqrt{2}+1)^2 \sqrt{\log M} + \frac{\tilde{B}_{n,E} + x}{\sqrt{\log M}} \Big) \Big\} + \frac{4E(\log M + \tilde{B}_{n,E} + 3 + x)}{n+1},$$

where $V_{j,n+1} = \sum_{t=0}^{n} \ell_{j,t+1}^2$ and $\tilde{B}_{n,E} = \log\Big(1 + \frac{E(E+1)}{e\sqrt{\log(M)}} + \frac{\log n}{2e}\Big) + \log E$ and π is any aggregative procedure that does not depend on the observations $\mathcal{F}_n$.

Proof. We analyze the concentration of the conditional excess of predictive risk $\mathbb{E}_t[\ell_{j,t+1}]$, where $\mathbb{E}_t$ denotes the expectation $\mathbb{E}_{P_t}$ where P_t is the law of (X_{t+1}, Y_{t+1}) conditionally on $\mathcal{F}_t$. For $t \notin \mathcal{T}$, we apply the transport inequality (6) to $-\eta_{j,t-1}\ell_{j,t}$ for P_{t-1} and any measure Q_{t-1} defined conditionally on $\mathcal{F}_{t-1}$. For $\eta = 4$ in (6), we obtain

$$\begin{aligned}\mathbb{E}_{t-1}[-\ell_{j,t}] &\leq \mathbb{E}_{Q_{t-1}}[-\ell_{j,t}] + 2\eta_{j,t-1}\mathbb{E}_{Q_{t-1}}[\ell_{j,t}^2] + \frac{2}{\eta_{j,t-1}}\mathcal{K}(Q_{t-1}, P_{t-1}) \\ &\leq \mathbb{E}_{Q_{t-1}}[-\ell_{j,t}] + 2\eta_{j,t-1}\mathbb{E}_{Q_{t-1}}[\ell_{j,t}^2] + \frac{2}{\eta_{j,n}}\mathcal{K}(Q_{t-1}, P_{t-1}).\end{aligned}$$

Here we use the fact that the $\eta_{j,t-1}$s are $\mathcal{F}_{t-1}$-measurable and constitute a non increasing sequence. For $t \in \mathcal{T}$, we simply use that $\mathbb{E}_{t-1}[-\ell_{j,t}] \leq E_t$. Summing up for $t = 1, \ldots, n+1$, integrating with respect to Q and using that $\sum_{t\in\mathcal{T}} E_t \leq 4E$ we obtain

$$\mathbb{E}_Q\Big[\sum_{t=0}^{n} \mathbb{E}_t[-\ell_{j,t+1}]\Big] \leq \mathbb{E}_Q\Big[\sum_{t=0}^{n} -\ell_{j,t+1} + 2\eta_{j,t}\ell_{j,t+1}^2 + \frac{2}{\eta_{j,n}}\sum_{t=0}^{n}\mathcal{K}(Q_t, P_t) + 4E\Big]. \quad (18)$$

Now we use the bound on $\sum_{t=0}^{n}\ell_{j,t+1}$ obtained in the core of the proof of Theorem 3.2:

$$-\sum_{t=0}^{n}\ell_{j,t+1} \leq 2\sum_{t=0}^{n}\eta_{j,t}\ell_{j,t+1}^2 + \frac{2}{\eta_{j,n}}\Big(\log(\pi_{j,0}^{-1}) + \log\Big(1 + \sum_{t=1}^{n}\sum_{j=1}^{M}\frac{\pi_{j,0}}{e}\Big(\frac{\eta_{j,t-1}}{\eta_{j,t}} - 1\Big)\Big)\Big).$$

We integrate it with respect to Q and we use the arguments in the proof of Theorem 3.4 to obtain for any $1 \leq j \leq M$

$$0 \leq \mathbb{E}_Q\Big[\sum_{t=0}^{n} E_t[\ell_{j,t+1}] + 4\sum_{t=0}^{n}\eta_{j,t}\ell_{j,t+1}^2 + \frac{2}{\eta_{j,n}}\Big(\log M + \tilde{B}_{n,E} + \sum_{t=0}^{n}\mathcal{K}(Q_t, P_t)\Big) + 4E\Big]. \quad (19)$$

Considering Q as the restriction of P to the event

$$A = \Big\{ \frac{\eta_{j,n}}{2} \Big(\sum_{t=0}^{n} \mathbb{E}_t[\ell_{j,t+1}] + 4 \sum_{t=0}^{n} \eta_{j,t} \ell_{j,t+1}^2 + 4E \Big) + \log M + \tilde{B}_{n,E} \leq -x \Big\}$$

we obtain $\mathbb{E}_Q[\sum_{t=0}^{n} \mathcal{K}(Q_t, P_t)] = \mathcal{K}(Q, P) = \log(1/P(A)) \geq x$ and the desired result follows using the computations in the proof of Theorem 3.4 and the sub-gradient trick applied to the cumulative predictive risk: $R_n(f) - R_n(f_\pi) \leq \mathbb{E}_\pi[\sum_{t=0}^{n} \mathbb{E}_t[\ell_{j,t+1}]]/(n+1)$. □

In the stochastic context, a proxy of the variance in the upper bound of any predictive risk is not avoidable for fast rates of convergence in n. It always appears during the online to batch conversion through the necessary Bernstein inequality, see [22]. Here we prefer to use the empirical Bernstein inequality for martingales given in Theorem 1.1 because it provides a confidence interval that is easily approximated. As an illustration, considering $\log \log n$ constant, for some $C > 0$ we have

$$R_n(\hat{f}) \leq \min_{\pi} R_n(f_\pi) + \frac{C\sqrt{\log M}}{n+1} \max_{1 \leq j \leq M} \sqrt{V_{j,n+1}} + \frac{CE \log M}{n+1}$$

As the term $\max_{1 \leq j \leq M} V_{j,n+1}$ can be estimated by $\max_{1 \leq j \leq M} \sum_{t=1}^{n} \ell_{j,t}^2$, it is a natural candidate to assert the complexity of the problem of aggregation; the more the $V_{j,n+1}$ are uniformly small and the more one can aggregate the elements of the dictionary optimally.

Notice that the generality of the result is remarkable; we do not assume any dependent structure nor boundedness on the observations. Indeed, in Theorem 4.1, E is not necessarily deterministic and can always be taken as equal to

$$E = \max_{1 \leq t \leq n+1} \max_{1 \leq j \leq M} |\ell_{j,t}|.$$

The range of the prediction E is also a good candidate to assert the complexity of the problem of aggregation. It is almost observable (one can estimate it by $\max_{1 \leq t \leq n} \max_{1 \leq j \leq M} |\ell_{j,t}|$) and is small for stationary $((\ell_{j,t})_{1 \leq j \leq M})_{t \in \mathbb{Z}}$ with light margin tails.

The complexity of the aggregation problem depends on the range and the proxy of the quadratic variation $V_{j,n+1}$. We will detail below the very restrictive context of bounded iid variables with strongly convex losses where the range and the proxy of the quadratic variation can be estimated easily. In more general contexts, as $\max_{1 \leq t \leq n} \max_{1 \leq j \leq M} |\ell_{j,t}|$ and $\sum_{t=1}^{n} \ell_{j,t}^2$ are observable, it is interesting to develop a parsimonious strategy that will only aggregate the elements of the dictionary with small complexity terms $\max_{1 \leq t \leq n} \max_{1 \leq j \leq M} |\ell_{j,t}|$ and $\sum_{t=1}^{n} \ell_{j,t}^2$. Estimating E, $V_{j,n+1}$ and M, the upper bound in 4.1 can be controlled at the price to restrict the dictionary and thus the corresponding best deterministic aggregation strategy $\min_\pi R_n(f_\pi)$. Such extensions of the present work will be developed in future researches.

4.2 Optimal learning in the iid case

As there is no warranty of the optimality of the general result given in Theorem 4.1, we restrict our study to the context of Lipschitz strongly convex losses with iid observations. in the iid framework where (X_t, Y_t) are iid copies of (X, Y), for any fixed constant f we have $R_n(f) = R(f) = \mathbb{E}[\ell(Y, f(X))]$. Thus it is always preferable to convert any online learner $\hat{f}$ to a batch learner by averaging

$$\bar{f} = \frac{1}{n+1} \sum_{t=0}^{n} \hat{f}_t$$

as an application of Jensen inequality gives $R(\bar{f}) \leq R_n(\hat{f})$. We have the following notion of optimality due to [27]:

Definition 4.1. *The batch learner $\tilde{f}$ is optimal if it exists some constant $c > 0$ such that*

$$R(\bar{f}) \leq \min_{f \in \mathcal{F}} R(f) + c \frac{\log M + x}{n+1}$$

with probability $1 - e^{-x}$, $x > 0$.

This optimality is called in deviation as it holds with high probability and by comparaison of the weakest notion of optimality in expectation where

$$\mathbb{E}_P[R(\bar{f})] \leq \min_{f \in \mathcal{F}} R(f) + c \frac{\log M}{n+1}.$$

Such fast rates cannot be obtained without regularity assumption on the loss ℓ, see [19, 6]. In the sequel $\ell : \mathbb{R}^2 \to \mathbb{R}$ is a loss function satisfying the following assumption called **(LIST)** after [22]

(LIST) the loss function ℓ is C_ℓ-smooth and C_b-Lipschitz continuous in its second coordinate on a convex set $\mathcal{C} \subset \mathbb{R}$.

Recall that a function g is c strongly convex on $\mathcal{C} \subset \mathbb{R}$ if there exists a constant $c > 0$ such that

$$g(\alpha a + (1-\alpha)a' \leq \alpha g(a) + (1-\alpha) g(a') - \frac{c}{2} \alpha(1-\alpha)(a-a')^2,$$

for any $a, a' \in C$, $0 < \alpha < 1$. Under the condition **(LIST)**, few algorithms are known to be optimal in expectation, see [5, 20, 21]. One of the most popular one is the Progressive Mixture Rule studied in detail in [8]. However PRM cannot be optimal in deviation, see [5].

Notice that Assumption **(LIST)** is restrictive and can hold only locally; on a compact set $\mathcal{C}$, the minimizer $f(y)^*$ of $f(y) \in \mathbb{R} \to \ell(y, f(y))$ exists and verifies, by strong convexity,

$$\ell(y, f(y)) \geq \ell(y, f(y)^*) + \frac{C_\ell}{2} (f(y) - f(y)^*)^2.$$

Moreover, by Lipschitz continuity, $\ell(y, f(y)) \le \ell(y, f(y)^*) + C_b|f(y) - f(y)^*|$. Thus, necessarily the diameter D of $\mathcal{C}$ is finite $C_\ell D \le 2C_b$. Then we deduce that the linearized loss is bounded by a deterministic constant $E = C_b D$:

$$\max_{1\le t\le n+1} \max_{1\le j\le M} |\ell_{j,t}| \le C_b D.$$

Theorem 4.2. *In the iid setting, under condition* **(LIST)**, *with probability* $1 - e^{-x}$ *we have*

$$R(\tilde{f}) \le \min_{1\le j\le M} R(f_j) + \frac{C_1 + C_2 \log M + C_3(\log(1+\log n) + 3\log(1+E)) + C_4 x + C_5 x^2}{n+1}$$

with $C_1 = 12C_bD + 144C_b^2/C_\ell$, $C_2 = 6C_bD + 2016C_b^2/C_\ell$, $C_3 = 6C_bD + 216C_b^2/C_\ell$, $C_4 = 7C_bD$ *and* $C_5 = 216C_b^2/C_\ell$.

Proof. We denote by $\mathbb{P}$ the measure of (X, Y) independent of $\mathcal{F}_n$. As we consider the batch version of BOA, we have the identities

$$\frac{1}{n+1}\sum_{t=1}^{n+1} \mathbb{E}_{\pi_t}[R(f_{j,t})] = \mathbb{E}_{\tilde{\pi}}[R(f_j)] \qquad \text{and} \qquad \frac{1}{n+1}\sum_{t=1}^{n+1} \mathbb{E}_{\mathbb{P}}[\ell_{j,t}^2] = \mathbb{E}_{\tilde{\pi}}[\mathbb{E}_{\mathbb{P}}[\ell_j^2]].$$

We start with the inequality (19) and we estimate the upper bound by its expectation; the second term in the sum of (19) can be bounded using Young's inequality

$$4\sum_{t=0}^{n} \eta_{j,t}\ell_{j,t+1}^2 \le \frac{4\sqrt{2}}{\sqrt{2}-1}\sqrt{V_{j,n+1}\log M} + 8E \le \frac{2}{\eta}\sum_{t=1}^{n+1}\frac{\ell_{j,t}^2}{E^2} + 20\eta E^2 \log M + 8E, \qquad \eta > 0.$$

The third term in the sum of (19), where $\sum_{t=0}^{n} \mathcal{K}(Q_t, P_t) = \mathcal{K}(Q, \mathbb{P})$, is bounded with

$$\begin{aligned}\frac{2}{\eta_{j,n}}(\log M + \tilde{B}_{n,E} + \mathcal{K}(Q,\mathbb{P})) \le& 4E(\log M + \tilde{B}_{n,E} + \mathcal{K}(Q,\mathbb{P})) + \frac{2}{\eta}\sum_{t=1}^{n+1}\frac{\ell_{j,t}^2}{E^2} \\ &+ \frac{3}{2}\eta E^2\Big(1 + \frac{\tilde{B}_{n,E}}{\log M} + \frac{\mathcal{K}(Q,\mathbb{P})^2}{\log M}\Big).\end{aligned}$$

Now we use the boundedness of $\mathbb{E}_\pi[\ell_{j,t}^2]/E^2 \le 1$ for any measure π on $\{1,\dots,M\}$ and the classical Bernstein inequality for (X, Y); via the variational form of the entropy (4), we have for any measure Q on (X, Y)

$$\begin{aligned}\mathbb{E}_Q\Big[\mathbb{E}_\pi\Big[\sum_{t=1}^{n+1}\frac{\ell_{j,t}^2}{E^2}\Big]\Big] &\le \mathbb{E}_{\mathbb{P}}\Big[\mathbb{E}_\pi\Big[\sum_{t=1}^{n+1}\frac{\ell_{j,t}^2}{E^2}\Big]\Big] + \mathbb{E}_{\mathbb{P}}\Big[\mathbb{E}_\pi\Big[\sum_{t=1}^{n+1}\frac{\ell_{j,t}^4}{E^4}\Big]\Big] + \mathcal{K}(Q,\mathbb{P}) \\ &\le 2\mathbb{E}_{\mathbb{P}}\Big[\mathbb{E}_\pi\Big[\sum_{t=1}^{n+1}\frac{\ell_{j,t}^2}{E^2}\Big]\Big] + \mathcal{K}(Q,\mathbb{P}). \qquad (20)\end{aligned}$$

Collecting all those identities and bounds in (19) we obtain

$$\mathbb{E}_Q[\mathbb{E}_{\tilde\pi}[R(f_j)]] \le \mathbb{E}_Q\Big[R(f_\pi)+\frac{8}{\eta E^2}\mathbb{E}_\pi[\mathbb{E}_{\tilde\pi}[\mathbb{E}_{\mathbb{P}}[\ell_j^2]]]\Big]+\frac{B_{n,E}(\eta)}{n+1}+4\Big(E+\frac{1}{\eta}\Big)\frac{\mathcal{K}(Q,\mathbb{P})}{n+1}+\frac{3\eta E^2\mathcal{K}(Q,\mathbb{P})^2}{2\log M(n+1)} \tag{21}$$

where Q is now any measure on $\mathcal{F}_n$ and

$$B_{n,E}(\eta) = 4E(2+\log M+\tilde B_{n,E})+\eta E^2\Big(\frac{3}{2}+20\log M+\frac{3\tilde B_{n,E}}{2\log M}\Big).$$

We estimate the proxy of the quadratic variation using the C_b-Lipschitz continuity of ℓ:

$$\mathbb{E}_\pi[\mathbb{E}_{\tilde\pi}[\mathbb{E}_{\mathbb{P}}[\ell_j^2]]] \le C_b^2\mathbb{E}_\pi[\mathbb{E}_{\tilde\pi}[\mathbb{E}_{\mathbb{P}}[(f_{j,t}-f_{j',t})^2]]] \le C_b^2(V(\pi)+V(\tilde\pi)+\mathbb{E}_{\mathbb{P}}[(\tilde f(X)-f_\pi(X))^2])$$

where $V(\pi)=\mathbb{E}_\pi[\mathbb{E}_{\mathbb{P}}[(f_j(X_t)-f_\pi(X_t))^2]]$. Then, we use as in [21] the convexity of the function $H\colon\ \pi\to R(f_\pi)+8C_b^2V(\pi)/(\eta E^2)$ when $\eta > 16C_b^2/(C_\ell E^2)$. Moreover, if one denotes π^* a minimizer of H, we have

$$R(f_\pi)+\frac{8C_b^2V(\pi)}{\eta E^2}-R(f_{\pi^*})-\frac{8C_b^2V(\pi^*)}{\eta E^2} \ge \Big(\frac{C_\ell}{2}-\frac{8C_b^2}{\eta E^2}\Big)\mathbb{E}_{\mathbb{P}}[(\tilde f(X)-f_\pi(X))^2]$$

Now, using C_ℓ-strong convexity as in Proposition 2 of [21], we have

$$R(f_\pi) \le \mathbb{E}_\pi[R(f_j)]-\frac{C_\ell V(\pi)}{2}.$$

Applying these inequalities to $\tilde\pi$, we obtain

$$\Big(\frac{C_\ell}{2}-\frac{16C_b^2}{\eta E^2}\Big)\mathbb{E}_{\mathbb{P}}[(\tilde f(X)-f_\pi(X))^2] \le \mathbb{E}_{\tilde\pi}[R(f_j)]-R(f_\pi)+\Big(\frac{16C_b^2}{\eta E^2}-\frac{C_\ell}{2}\Big)V(\tilde\pi)-\frac{8}{\eta E^2}\mathbb{E}_\pi[\mathbb{E}_{\tilde\pi}[\mathbb{E}_{\mathbb{P}}[\ell_j^2]]].$$

Choosing $\eta^* = 64C_b^2/(E^2C_\ell)$, integrating with respect to Q and using the estimate in (21) we derive that

$$\mathbb{E}_Q[\mathbb{E}_{\mathbb{P}}[(\tilde f(X)-f_\pi(X))^2]+V(\tilde\pi)] \le \frac{4}{C_\ell}\Big(B_{n,E}(\eta^*)+4\Big(E+\frac{1}{\eta^*}\Big)\frac{\mathcal{K}(Q,\mathbb{P})}{n+1}+\frac{3\eta^*E^2\mathcal{K}(Q,\mathbb{P})^2}{2\log M(n+1)}\Big).$$

Plugging in this estimate into (21) we obtain

$$\mathbb{E}_Q[\mathbb{E}_{\tilde\pi}[R(f_j)]] \le \mathbb{E}_Q\Big[R(f_\pi)+\frac{C_\ell}{4}V(\pi)\Big]+\frac{3}{2}\Big(B_{n,E}(\eta^*)+4\Big(E+\frac{1}{\eta^*}\Big)\frac{\mathcal{K}(Q,\mathbb{P})}{n+1}+\frac{3\eta^*E^2\mathcal{K}(Q,\mathbb{P})^2}{2\log M(n+1)}\Big).$$

We conclude by a crude estimate on the last term, noticing that

$$\tilde B_{n,E} \le \log(1+\log n)+3\log(1+E),$$

choosing Q similarly than at the end of the proof of Theorems 1.1 and 4.1 and noticing that $\inf_\pi\{R(f_\pi)+C_\ell V(\pi)/4\} \le \min_{1\le j\le M} R(f_j)$. □

The result in Theorem 4.2 is a direct consequence of Theorem 4.1 obtained by a rough estimate of the confident interval obtained there. Thus, the result in Theorem 4.1 is always more precise than the one in Theorem 4.2. The interest of Theorem 4.2 is that it is an online to batch conversion that expresses the condiment interval on the cumulative risk as an oracle inequality. This framework is very classical in mathematical statistics and lower bounds are given in [27]. The main advantage of the approach of the individual sequences consisting in

The awful constants and the $\log\log n$ terms seem to be the price to pay for adaptivity, i.e. the whole procedure does not depend on any constants involved in Assumption **(LIST)**. Moreover, the batch version of BOA is explicitly computed here with linear complexity $O(Mn)$. It is a real advantage of the BOA procedure compared with the Q-aggregation procedure given in [21]. We obtain much better constants for the batch version of the non-adaptive BOA procedure:

Theorem 4.3. *In the iid setting, under condition* **(LIST)**, *for any initial weights* π_0 *and any learning rate* $\eta = \gamma/(C_bD)^2$ *with* $0 < \gamma < C_\ell/(24C_b^2)$, *with probability* $1-e^{-x}$ *we have*

$$R(\tilde{f}) \le \min_{1\le j\le M}\Big\{R(f_j) + \frac{C_\gamma}{n+1}\Big(\frac{2(C_bD)^2\mathcal{K}(\pi_{j,0}^{-1})}{\gamma} + \Big(\frac{((C_bD)^2}{\gamma} + 2\gamma\Big)x\Big)\Big\}$$

where $C_\gamma = 1 + \frac{6\gamma}{(C_bD)^2(C_\ell/2-12C_b^2\gamma)}$.

Proof. The proof follows from the result in Theorem 3.2 and an application of the variational form of the Bernstein's inequality of [14] on $\ell_{j,t}$ (instead of the empirical one used in the proof of Theorem 4.1):

$$\mathbb{E}_Q[\mathbb{E}_{\hat\pi}[R(f_j)]] \le \mathbb{E}_Q[R(f_\pi)] + 2\gamma\sum_{t=0}^{n}\frac{\mathbb{E}_Q[\mathbb{E}_\pi[\ell_{j,t+1}^2] + \mathbb{E}_{\mathbb{P}}[\mathbb{E}_\pi[\ell_{j,t+1}^2]]]}{E^2(n+1)} + \frac{E^2(2\mathcal{K}(\pi,\pi_0)+\mathcal{K}(Q,\mathbb{P}))}{\gamma(n+1)}.$$

Then we use the variational form of the Bernstein's inequality on $\mathbb{E}_\pi[\ell_{j,t}^2]/E^2$ as in (20) to obtain

$$\mathbb{E}_Q[\mathbb{E}_{\hat\pi}[R_n(f_j)]] \le \mathbb{E}_Q\Big[R_n(f_\pi) + \frac{6\gamma\mathbb{E}_{\mathbb{P}}[\mathbb{E}_\pi[\ell_j^2]]}{E^2}\Big] + \frac{2E^2\mathcal{K}(\pi,\pi_0)}{\gamma(n+1)} + \Big(\frac{E^2}{\gamma}+2\gamma\Big)\frac{\mathcal{K}(Q,\mathbb{P})}{n+1}.$$

Following the same reasoning than in the proof of Theorem 4.2, we obtain

$$\mathbb{E}_Q[\mathbb{E}_P[\mathbb{E}_\pi[\ell_j^2]]] \le V(\pi) + \frac{1}{C_\ell/2-12C_b^2\gamma}\Big(\frac{2E^2\mathcal{K}(\pi,\pi_0)}{\gamma(n+1)} + \Big(\frac{E^2}{\gamma}+2\gamma\Big)\frac{\mathcal{K}(Q,\mathbb{P})}{n+1}\Big).$$

We conclude the proof using that $E = C_bD$ and choosing Q similarly than at the end of the proof of Theorems 1.1 and 4.1. □

The BOA procedure is here non adaptive, in the sense that it depends on the assumption **(LIST)** but it is still independent of the observations (except of their range E). For

instance, if we choose the square loss and we restrict us on the interval $[-1/2; 1/2]$ then we can choose $\gamma = \eta = 1/13 \leq 1/12$, $\pi_{j,0} = M^{-1}$ and we obtain, with probability $1 - e^{-x}$,

$$R(\tilde{f}) \leq \min_{1 \leq j \leq M} R(f_j) + \frac{14(13 \log M + 7x)}{n+1}.$$

The constant in front of $\log M$ is twice as large than in the result in [21] in the same context but we do not assume here that $|Y| \leq 1/2$ and we do not have an extra parameter to calibrate, as it is always the case for Q-aggregation, see [12] for practical issues.

References

[1] Abernethy, J., Agarwal, A., Bartlett, P. and Rakhlin, A. (2009) A Stochastic View of Optimal Regret through Minimax Duality. *COLT 2009*.

[2] Alquier, P., Li, X. and Wintenberger, O. (2013) Prediction of Time Series by Statistical Learning: General Losses and Fast Rates, *Dependence Modeling*, **1**, 65–93.

[3] Agarwal, A. and Duchi, J. C. (2013) The generalization ability of online algorithms for dependent data. *Information Theory, IEEE Trans.*, **59**, 573–587.

[4] Audibert, J. Y., Munos, R. and Szepesvari, C. (2006). Use of variance estimation in the multi-armed bandit problem. *NIPS (2006)*

[5] Audibert, J. Y. (2007) Progressive mixture rules are deviation suboptimal In *Advances in Neural Information Processing Systems*, 41–48.

[6] Audibert, J. Y. (2009) Fast learning rates in statistical inference through aggregation *The Annals of Statistics*, **37(4)**, 1591–1646.

[7] Boucheron, S., Lugosi, G. and Massart, P. (2013) *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press.

[8] Catoni, O. (2004) *Statistical Learning Theory and Stochastic Optimization, Lecture Notes in Mathematics (Saint-Flour Summer School on Probability Theory 2001).* Springer.

[9] Catoni, O. (2007) *Pac-Bayesian supervised classification: the thermodynamics of statistical learning. Institute of Mathematical Statistics Lecture Notes? Monograph Series, 56.* Institute of Mathematical Statistics, Beachwood, OH.

[10] Cesa-Bianchi, N. and Lugosi, G. (2006) *Prediction, Learning, and Games.* Cambridge University Press, New York, NY, USA.

[11] Cesa-Bianchi, N., Mansour, Y. and Stoltz, G. (2007) Improved second-order bounds for prediction with expert advice. *Machine Learning*, **66**, 321–352.

[12] Dai, D., Rigollet, P. and Zhang, T. (2012) Deviation optimal learning using greedy Q-aggregation. *The Annals of Statistics*, **40(3)**, 1878–1905.

[13] Donsker, M. D., and Varadhan, S. S. (1975) Asymptotic evaluation of certain Markov process expectations for large time, I. *Communications on Pure and Applied Mathematics*, **28(1)**, 1–47.

[14] Freedman, D. A. (1975) On tail probabilities for martingales. *Ann. Probab.*, 100–118.

[15] Gaillard, P., Stoltz, G. and Van Erven, T. (2014) A Second-order Bound with Excess Losses. *arXiv preprint arXiv:1402.2044*.

[16] Haussler, D., Kivinen, J., and Warmuth, M. K. (1998) Sequential prediction of individual sequences under general loss functions. *Information Theory, IEEE Transactions*, **44(5)**, 1906–1925.

[17] Hazan, E. and Kale, S. (2010) Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine learning*, **80(2-3)**, 165–188.

[18] Juditsky, A., Rigollet, P. and Tsybakov, A. B. (2008) Learning by mirror averaging. *The Annals of Statistics*, **36(5)**, 2183–2206.

[19] Lecué, G. (2007) Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, **13(4)**, 1000–1022.

[20] Lecué, G. and Mendelson, S. (2009) Aggregation via empirical risk minimization. *Probab. theory and related fields*, **145(3-4)**, 591–613.

[21] Lecué, G. and Rigollet, P. (2013) Optimal learning with Q-aggregation. *arXiv preprint* arXiv:1301.6080.

[22] Kakade, S. M. and Tewari, A. (2008) On the Generalization Ability of Online Strongly Convex Programming Algorithms. In NIPS (801–808).

[23] Kivinen, J. and Warmuth, M. K. (1999) Averaging expert predictions. In *Computational Learning Theory* (153–167). Springer Berlin Heidelberg.

[24] Marton K. (1996) Bounding $\overline{d}$-distance by informational divergence: a method to prove measure concentration. *Ann. Probab.* **24 (2)**, 857–866.

[25] Maurer, A. and Pontil, M. (2009) Empirical Bernstein bounds and sample variance penalization. *COLT (2009)*

[26] Mohri, M. and Rostamizadeh, A. (2010) Stability Bounds for Stationary φ-mixing and β-mixing Processes. *JMLR*, **11**, 789–814.

[27] Tsybakov, A. B. (2003) Optimal rates of aggregation. In *Learning Theory and Kernel Machines* (303–313). Springer Berlin Heidelberg.

[28] Vovk, V. G. (1990) Aggregating strategies. In *Proc. Third Workshop on Computational Learning Theory* (371–383). Morgan Kaufmann.

[29] Wintenberger, O. (2012) Weak transport inequalities and applications to exponential inequalities and oracle inequalities. *arXiv preprint* Arxiv:1207.4951

[30] Zhang, T. (2005) Data dependent concentration bounds for sequential prediction algorithms. In *Learning Theory* (173–187). Springer Berlin Heidelberg.