Optimal learning with Bernstein Online Aggregation

Olivier Wintenberger

Optimal learning with Bernstein Online Aggregation

Olivier Wintenberger
olivier.wintenberger@upmc.fr
Sorbonne Universités, UPMC Univ Paris 06
LSTA, Case 158, 4 place Jussieu
75005 Paris, FRANCE

April 20, 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 [HK10]. 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 the first exponential weights algorithm that exhibits a second order bound with excess losses that appears first in [GSVE14]. 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. Then we obtain an observable estimate of 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 optimal rate of aggregation of [Tsy03] is achieved. The batch version of the BOA procedure is then the first adaptive explicit algorithm that satisfies an optimal oracle inequality with high probability.

Keywords 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: $\hat{f} = (\hat{f}_0, \hat{f}_1, \hat{f}_2, \ldots)$. 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

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

have been intensively studied; see for instance the seminal book [CBL06]. 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

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

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 (BOA) procedure, derive from the study of the concentration properties of the difference between the excess of cumulative predictive risk and the regret:

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

where π^* is the measure on $\{1, \ldots, M\}$ with the best deterministic weights. It constitutes a martingale (M_t) adapted to the filtration $(\mathcal{F}_t)$. For any such martingale (M_t), we denote $\Delta M_t = M_t - M_{t-1}$ the difference of martingale, $< M >_t = \sum_{j=1}^{t} \mathbb{E}[\Delta M_j^2 \mid \mathcal{F}_{j-1}]$ and $[M]_t = \sum_{j=1}^{t} \Delta M_j^2$ its quadratic variation and predictable quadratic variation, respectively. Instead of using the classical Bernstein inequality for martingales as in [Fre75, Zha05], we develop its empirical counterpart that provides concentration of M via the predictable quadratic variation $[M]$ 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 [AMS06, MP09] 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. We will use a proxy of the quadratic variation $[M]$ 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 predicted accurately and which past predictions losses are close to the loss of the last aggregative online learner, ensuring the stability in time and a small quadratic variation. We introduce 3 different versions of the algorithm: the aggregation procedure itself described in Figure 1 and denoted $\hat{f}$, its randomized version, 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:

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

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. Its cumulative predictive risk is then

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

One can thus explain 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) as an unbiased estimator of the predictive risk. Such randomized algorithms that take into account a proxy of the quadratic variation (variance in the iid context considered there) have been studied by [Aud09]. In the prediction from experts context, exponential weights aggregation with a proxy of the variance has already been considered in [HK10].

In all the sequel, we focus on the less general context of a convex loss ℓ. We denote ℓ' the sub gradient of the loss 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, $R_n(\hat{f}) \leq R_n(\bar{f})$. We also use the sub gradient trick, see [CBL06], and replace in BOA the loss $\ell(Y_t, f_{j,t-1}(X_t))$ with its linearized version

$$\mathbb{E}_{\pi_{t-1}}[\ell'(Y_t, f_{j',t-1}(X_t))] f_{j,t-1}(X_t). \tag{4}$$

The proxy of the quadratic variation $V_{j,n+1} = \sum_{t=1}^{n+1} \ell_{j,t}^2$ is modified with

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

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 in Theorem 3.1 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 $f = (f_0, f_1, f_2, \ldots)$:

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

Such second order bounds with excess losses have been obtained by [GSVE14] for other procedures. They obtain better constants in their bound available for losses bounded by 1 only. Following the pioneer work of [CBMS07], we also analyze the second order properties of a new adaptive version of exponential weights, see Figure 2. The novelty, compared with classical adaptive procedures developed in [CBMS07], is the dependence of the learning rates with respect to j and we consider 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 estimate of the range of the linearized loss (4). 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 such as

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

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

for some "constant" (increasing in $\log\log n$) $C > 0$, see Theorems 3.2 and 3.3 for details. Such second order bounds involving excess losses terms as the $\ell_{j,t}$s have been proved for other algorithms in [GSVE14], and we refer to this article for nice consequences of such bounds. The optimality of such bounds is difficult to assert because it depends on the quadratic variation 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

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

Such bounds are also achieved by classical exponential weights algorithms with no proxy of the quadratic variation, see [Vov90, HKW98]. 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 to convert a second order bound 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. Formerly, such inequalities where derived under very restrictive dependent assumptions, see [ALW13, MR10, AD13]. The generality of our results 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 and constant elements of the dictionary $f_{j,t} = f_j$, $t \geq 0$ (we can suppress the index t). Moreover, it appears naturally when using the minimax theory approach, see [AABR09]. 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 second order bounds with excess losses are unknown. Thus we provide the optimality of the BOA procedure 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

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

and 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 than the online version $\hat{f}$. When the loss is Lipschitz continuous and strongly convex, we obtain an oracle inequality in deviation; with probability $1 - e^{-x}$, $x > 0$, we have

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

The fast rate $\log M/(n+1)$ is optimal, see [Tsy03]. Notice that the proxy of the variance is necessary; without it, the batch version of the BOA procedure coincides with the Progressive Mixture Rule of [Cat04]. This procedure is optimal in expectation; see [Cat04, JRT08], but suboptimal in deviation; see [Aud07]. Thus, the stabilization term in the batch version of the BOA procedure is necessary to control the deviations of the exponential weights algorithms. Few other optimal learning procedures are known. The procedures in [Aud07, LM09] achieve the optimal rate using some prior information on the dictionary. In the Q-aggregation procedure of [LR13] 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 [DRXZ12]. On the opposite, the BOA procedure is the first explicit algorithm that achieves the optimal rate of convergence in deviation. Such optimal learning results should also be extended to the other algorithms provided in [GSVE14] that achieved similar second order bounds with excess losses (under the boundedness of the losses).

We present the second order bounds with excess losses of the regret of BOA in Section 3 and the main results in a stochastic environment are provided in Section 4. All the proofs are given using the formalism of the transport of measure developed in [Win12]. Some arguments of the proofs are directly translated in this new formalism from former

works like [Aud09, CBMS07, GSVE14, LR13] and are provided here for completeness. In the next Section, we introduce this new formalism and prove a new empirical Bernstein inequality that represents the main probabilistic tool of our paper provided in Theorem 1.1. It is proved in full generality as it will be useful in future work.

2 Transport of measure and proof of the empirical Bernstein inequalities

Classically, Bernstein's inequality is derived from an estimate of the Laplace transform and the Chernoff device. Here, we derive empirical Bernstein's inequality of Theorem 1.1 using another approach originally developed by [Mar96] and based on the variational formula of the entropy $\mathcal{K}(Q,P) = \mathbb{E}_Q[\log(dP/dQ)]$:

Lemma 2.1 ([DV75] variational formula of the entropy)**.** *For any probability measures P on $\mathcal{X}$ and any measurable function $h : \mathcal{X} \to \mathbb{R}$ we have:*

$$\mathbb{E}_P[\exp(h - \mathbb{E}_P[h])] \leq 1 \quad \Longleftrightarrow \quad \mathbb{E}_Q[h] - \mathbb{E}_P[h] \leq \mathcal{K}(Q,P), \qquad \text{for any measure } Q. \tag{5}$$

The left hand side corresponds to the right hand side with Q equals 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 the proofs of the optimality of exponential weights aggregation procedure, see [Cat07]. 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 to any measure Q, see [Win12] 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

$$\mathbb{E}_Q[X] \leq \mathbb{E}_P[X] + \lambda(\mathbb{E}_Q[X_+^2] + \mathbb{E}_P[X_-^2])/2 + \frac{\mathcal{K}(Q,P)}{\lambda}. \tag{6}$$

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 quadratic variation, 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. Applying Lemma 2.2 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 derive 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 for any $\lambda > 0$, we have

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

Plugging this estimate in the inequality (6), 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 $\gamma = \lambda/(1-\lambda/2)$ we obtain

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

Using the variational form of the entropy we obtain

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

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

We are now ready to prove the empirical Bernstein inequality for martingales of Theorem 1.1. It follows from the exponential inequality provided in 2.3 by an application of classical Doob's submartingale argument; see [Fre75]. Below, we provide a new proof that uses only simple algebra and the decomposition of the entropy. It follows the formalism of transport of measures due to [Mar96, Win12]. The reasoning is detailed as it will be used several time in the proofs of the paper:

Proof. of Theorem 1.1. We apply (7) 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 and remarking that for P the distribution of M_τ the entropy decomposes as

$$\mathbb{E}_Q\Big[\sum_{t=1}^{\tau} \mathcal{K}(Q_t, P_t)\Big] = \mathbb{E}_Q\Big[\sum_{t=1}^{\tau} \log(dQ_t/dP_t)\Big] = \mathbb{E}_Q\Big[\log\Big(\frac{dQ_1 \cdots dQ_\tau}{dP_1 \cdots dP_\tau}\Big)\Big] = \mathcal{K}(Q, P) \quad (8)$$

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 the 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 with excess losses of 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, similar than in [Ger13], where $(X_t, Y_t) = (x_t, y_t)$ are provided recursively for $1 \leq t \leq n$. In that case, the cumulative loss $\mathcal{R}(f)$ quantify the prediction of $f = (f_0, f_1, f_2, \ldots)$. We focus on convex losses and then the sub gradient trick is useful to compare the BOA procedure with the best deterministic aggregation $f_\pi = \sum_{j=1}^{M} \pi_j f_j$:

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

$$\eta \max_{1 \leq t \leq n+1} \max_{1 \leq j \leq M} \ell_{j,t_+} \leq 1. \quad (9)$$

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 work sequentially on the weights $(\pi_{j,t})$: for any $1 \leq j \leq M$, $0 \leq t \leq n$, under (9), 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 (7) to $-\eta \ell_{j,t+1}$ for $P = \pi_t$ and $Q = \pi_{t+1}$, two measures on $\{1, \ldots, M\}$. Taking $\gamma = 4$ in (7) 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 derive that

$$0 \leq \mathbb{E}_{\pi_{t+1}}[2^{-1}\eta \ell_{j,t+1} + \eta^2 \ell_{j,t+1}^2 + \log(d\pi_{t+1}/d\pi_t)]. \tag{10}$$

By the specific form of (π_t), we use the classical telescoping sum reasoning

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

Summing up for $t = 0, \ldots, n$ 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 (5) 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

$$\mathbb{E}_{\hat{\pi}}[\mathcal{R}(f_j)] = \sum_{t=1}^{n+1} \mathbb{E}_{\pi_t}[\ell(Y_t, f_{j,t-1}(X_t))]. \tag{11}$$

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 recognize the proxy of the proxy of the quadratic variation

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

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 the loss is quadratic.

The first application of Theorem 3.1 is the context of individual sequences prediction [CBL06]. We consider that $Y_t = y_t$ for a deterministic sequence $y_0, \ldots, y_n$ ((X_t) is useless in this context). We have $\mathcal{F}_t = \{y_0, \ldots, y_t\}$, $0 \leq t \leq n$, and the online learners $f_j = (y_{j,1}, y_{j,2}, y_{j,3}, \ldots)$ of the dictionary are called the experts. The cumulative loss is now $\mathcal{R}(\hat{f}) = \sum_{t=1}^{n+1} \ell(y_t, \hat{y}_t)$ for any aggregative strategy $\hat{y}_t = \hat{f}_{t-1} = \sum_{j=1}^{M} \pi_{j,t-1} y_{j,t}$ where $\pi_{j,t-1}$ are measurable functions of the past $\{y_0, \ldots, y_{t-1}\}$. The estimate obtained in Theorem 3.1 on the regret $\mathcal{R}(\hat{f}) - \mathcal{R}(f_\pi^*)$ is called a second order bound after the seminal paper [CBMS07]. The excess losses similarly appears in the second order bounds of [GSVE14] for the Prod algorithm when losses are bounded by 1.

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:

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

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

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

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 $\ell_{j,t}$, $1 \leq t \leq n+1$. 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 adaptive version 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 different from [CBMS07] as the learning rates $\eta_{j,t}$ depends on the element of the dictionary j and appear into the exponential and as a product. Adaptive procedures with such multiplicative forms have been studied in [GSVE14]. Notice that the adaptive weights are only well defined when the learning rates are positive: $\eta_{j,t} > 0$, $1 \leq j \leq M$, $0 \leq t \leq n$.

Remark also that such multiplicative adaptive form can be investigated for other exponential weights than for those of BOA. We obtain a second order bound with excess losses for this adaptive BOA procedure similar than in [GSVE14]:

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

$$\eta_{j,t-1}\ell_{j,t_+} \le 1, \qquad 1 \le t \le n+1, \quad 1 \le J \le M, \tag{12}$$

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. Some elements of the proofs are translated from [GSVE14] to the transport of measure formalism described in Section 2. 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 (12) we apply the transport inequality (7) 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 $\gamma = 4$ in (7) 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

$$\sum_{j=1}^{M}\tilde{\pi}_{j,t} \le \sum_{j=1}^{M}\tilde{\pi}_{j,t-1} + \frac{1}{e}\Big(\sum_{j=1}^{M}\frac{\eta_{j,t-1}}{\eta_{j,t}}\pi_{j,0} - 1\Big) \tag{13}$$

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

$$\frac{\tilde{\pi}_{j,t}}{\pi_{j,0}} = \exp(-2^{-1}\eta_{j,t}\tilde{L}_{j,t}) = \exp(-2^{-1}\eta_{j,t}\ell_{j,t} - \eta_{j,t}\eta_{j,t-1}\ell_{j,t}^2)\exp(-2^{-1}\eta_{j,t}\tilde{L}_{j,t-1}) \tag{14}$$
$$= (\exp(-2^{-1}\eta_{j,t-1}\ell_{j,t} - \eta_{j,t-1}^2\ell_{j,t}^2)\exp(-2^{-1}\eta_{j,t-1}\tilde{L}_{j,t-1}))^{\eta_{j,t}/\eta_{j,t-1}}.$$

As $\eta_{j,t}$ is non increasing with t for any $1 \leq j \leq M$, we have that $\alpha = \eta_{j,t-1}/\eta_{j,t} \geq 1$. Then, following the reasoning in [GSVE14], we use the inequality $x \leq x^\alpha + (\alpha - 1)/e$ for any $x \geq 0$ and $\alpha \geq 1$ to derive that

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

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

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

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 (13) follows from an application of Theorem 2.3 with $\lambda = 2^{-1}$, $P = \tilde{\pi}_{t-1}$ and $X = \eta_{j,t-1}\ell_{j,t}$ satisfying $\mathbb{E}_P[X] = 0$ by definition of $\hat{f}_t$. Using recursively (13) 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 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 the successive conditional transport problems via the variational formula of the entropy (5). 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 the sub-gradient trick. However, the upper bound is less share as it 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:

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

For such weights, a recursive argument similar to the proof of Theorem 3.1 can be used (see the Appendix D of the preliminary version of [Aud09] available on arXiv:math/0703854). It asserts the optimality of such adaptive procedure via the variational formula of the entropy (5) for the transport problem to any measure π on $\{1, \ldots, M\}$.

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 $|\ell_{j,t}| \leq E$, $1 \leq t \leq n+1$, $1 \leq J \leq M$. The fact that E is larger than one is a technical assumption that is non restrictive as one can always renormalize $\ell_{j,t}$ by the first observed loss $\ell_{1,1}$. We tune the learning rates in the following way

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

The learning rates are similar than those of Section 4.1 in [CBMS07] except that they depend on j through the quadratic variation proxy $\sum_{s=1}^t \ell_{j,s}^2$; see [GSVE14] for similar multiple learning rates. We restrict to the cases where $M > 1$ to consider only positive learning rates $\eta_{j,t} > 0$. We provide below a second order bound with excess losses on the regret of adaptive BOA:

Theorem 3.3. *If $|\ell_{j,t}| \leq E$, $1 \leq t \leq n+1$, $1 \leq J \leq M$ (E, $M > 1$) and the learning rates are tuned as in* (15) *then the adaptive BOA procedure achieves, for all $n \geq 1$,*

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

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

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

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

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

Now we use similar arguments than in the proof of Theorem 5 of [CBMS07]; 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 [GSVE14] 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 3.3 similarly than the proof of Theorem 5 in [CBMS07]. □

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 [CBMS07], 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

$$\eta_{j,t}=\min\left\{\frac{1}{E_t},\sqrt{\frac{\log M}{\sum_{s=1}^{t}\ell_{j,s}^2}}\right\},\qquad t\geq 0. \tag{16}$$

This rule for updating learning rates is similar than the one in [CBMS07] except that it depends on j and that E_t is larger than 1. This restriction allows us to estimate by $\log E$ the number of different values of E_t. We have

Theorem 3.4. *If $|\ell_{j,t}| \leq E$, $1 \leq t \leq n+1$, $1 \leq J \leq M$ (E, $M > 1$) and the learning rates are tuned as in* (16) *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 (14) and $\eta_{j,t-1} \geq \eta_{j,t}$ to derive that

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

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

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

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

$$\sum_{j=1}^{M} \tilde{\pi}_{j,t} \leq \sqrt{e}\Big(\sum_{j=1}^{M} \tilde{\pi}_{j,t-1} + \frac{1}{e}\Big(\sum_{j=1}^{M} \frac{\eta_{j,t-1}}{\eta_{j,t}} \pi_{j,0} - 1 \Big) \Big) \tag{17}$$

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

$$E_t = E_{t_{r+1}} \qquad \text{and} \qquad \max_{1\leq j\leq M} |\eta_{t-1}\ell_{j,t}| \leq 1, \qquad t \notin \mathcal{T}. \tag{18}$$

Then the same reasoning than in the proof of Theorem 2.3 apply and the recursive formula (13) holds.

To conclude, we apply recursive formulas (13) and (13) and we obtain the upper bound

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

The logarithm $\log(\sum_{j=1}^{M} \tilde{\pi}_{j,n})$ is bounded by $B_{n,E}$ and an additional term smaller than $R/2 \leq \lceil (\log_2 E)_+ \rceil / 2 \leq \log E$. Theorem 3.4 is proved replacing $B_{n,E}$ with $\tilde{B}_{n,E}$ and using similar arguments than in the proof of Theorem 6 in [CBMS07]. □

The advantage of the adaptive BOA procedure compared with the procedures studied in [GSVE14] is to be adaptive to the unknown range that can be unbounded (E can be random in Theorem 3.4).

The second order bound of the regret obtained in Theorem 3.4 provides a confident interval for the regret in term of a proxy of the quadratic variation

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

Let us give an example where this proxy is small. 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 $b > 0$, $1 \leq j \leq M$ and $1 \leq t \leq n+1$, using the notation of (11), 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

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

As ℓ is the quadratic loss we have the decomposition

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

If $\min_\pi \mathcal{R}(f_\pi) \leq (1 - \eta^*) \min_{1 \leq j \leq M} \mathcal{R}(f_j)$ for sufficiently small $\eta^* > 0$, using the Young inequality for this 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; see [HKW98], but the constants (not explicit here) are certainly not, see [Vov90]. Notice that the stabilization term in the BOA procedure can be avoided as the simpler Exponential Weights Averaging algorithm of [KW99] satisfies such optimal regret bounds. 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 the optimality of averaging procedures without the need of 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 [Vov90, KW99].

4 Optimality of the BOA procedure in a stochastic environment

4.1 Second order bounds 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 motivation of the introduction of the BOA procedure with a proxy of the quadratic variation in the exponential weights is the extension of the second order bounds on the regret to the excess of cumulative predictive risk. We are now ready to state the main result of the paper:

Theorem 4.1. *If $|\ell_{j,t}| \leq E$, $1 \leq t \leq n+1$, $1 \leq j \leq M$ (E, $M > 1$) and the learning rates are tuned as in* (16) *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 measure on $\{1 \ldots, M\}$ independent of $\mathcal{F}_n$.

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 (7) 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 $\gamma = 4$ in (7), 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

$$\mathbb{E}_Q\Big[\sum_{t=0}^{n} \mathbb{E}_t[-\ell_{j,t+1}]\Big] \leq \mathbb{E}_Q\Big[\sum_{t=0}^{n} -\ell_{j,t+1} + 2\eta_{j,t}\ell_{j,t+1}^2 + \frac{2}{\eta_{j,n}}\sum_{t=0}^{n}\mathcal{K}(Q_t, P_t) + 4E\Big]. \tag{19}$$

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 of the proof of Theorem 3.4 to obtain

$$0 \leq \mathbb{E}_Q\Big[\sum_{t=0}^n E_t[\ell_{j,t+1}] + 4\sum_{t=0}^n \eta_{j,t}\ell_{j,t+1}^2 + \frac{2}{\eta_{j,n}}\Big(\log M + \tilde{B}_{n,E} + \sum_{t=0}^n \mathcal{K}(Q_t,P_t)\Big) + 4E\Big]. \quad (20)$$

Consider 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 have $\mathbb{E}_Q[\sum_{t=0}^n \mathcal{K}(Q_t,P_t)] = \mathcal{K}(Q,P) = \log(1/P(A)) \geq x$ and the desired result follows using the computations provided 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 quadratic variation appears in any upper bound on the excess of risk. It is due to the online to batch conversion and the use of a Bernstein inequality; see for instance [KT08]. Here, we use the empirical Bernstein inequality for martingales given in Theorem 1.1 because it provides a confidence interval that can be 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.

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 would be interesting to develop a parsimonious strategy that would 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 and $V_{j,n+1}$, reducing the size M of the dictionary, the last terms in the second order bound of Theorem 4.1 are controlled at the price to increase the accuracy of the best deterministic aggregation strategy $\min_\pi R_n(f_\pi)$.

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 constant learner $f = f_t$, $t \geq 0$, 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 [Tsy03]:

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. By comparison, the weaker notion of optimality in expectation is defined as

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

Such fast rates cannot be obtained without regularity assumption on the loss ℓ, see [Lec07, Aud09]. In the sequel $\ell : \mathbb{R}^2 \to \mathbb{R}$ is a loss function satisfying the following assumption called **(LIST)** after [KT08]

(LIST) the loss function ℓ is C_ℓ-strongly convex 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 [Aud07, LM09, LR13]. One of the most popular one is the Progressive Mixture Rule studied in detail in [Cat04]. However PRM cannot be optimal in deviation, see [Aud07].

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 $|\ell_{j,t}| \le C_b D.$, $1 \le t \le n+1$, $1 \le j \le M$, and the range can be fixed as $E = C_b D$.

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 (20) and we estimate the upper bound by its expectation; the second term in the sum of (20) can be bounded using Young's inequality

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

The third term in the sum of (20), where $\sum_{t=0}^{n} \mathcal{K}(Q_t, P_t) = \mathcal{K}(Q, \mathbb{P})$, is bounded with

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

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 (5), we have for any measure Q on (X, Y)

$$\begin{aligned}\mathbb{E}_Q\Big[\mathbb{E}_\pi\Big[\sum_{t=1}^{n+1}\frac{\ell_{j,t}^2}{E^2}\Big]\Big] &\le \mathbb{E}_{\mathbb{P}}\Big[\mathbb{E}_\pi\Big[\sum_{t=1}^{n+1}\frac{\ell_{j,t}^2}{E^2}\Big]\Big] + \mathbb{E}_{\mathbb{P}}\Big[\mathbb{E}_\pi\Big[\sum_{t=1}^{n+1}\frac{\ell_{j,t}^4}{E^4}\Big]\Big] + \mathcal{K}(Q,\mathbb{P}) \\ &\le 2\mathbb{E}_{\mathbb{P}}\Big[\mathbb{E}_\pi\Big[\sum_{t=1}^{n+1}\frac{\ell_{j,t}^2}{E^2}\Big]\Big] + \mathcal{K}(Q,\mathbb{P}). \qquad (21)\end{aligned}$$

Collecting all those identities and bounds in (20) we obtain

$$\mathbb{E}_Q[\mathbb{E}_{\tilde{\pi}}[R(f_j)]] \le \mathbb{E}_Q\Big[R(f_\pi) + \frac{8}{\gamma E^2}\mathbb{E}_\pi[\mathbb{E}_{\tilde{\pi}}[\mathbb{E}_{\mathbb{P}}[\ell_j^2]]]\Big] + \frac{B_{n,E}(\gamma)}{n+1} + 4\Big(E + \frac{1}{\gamma}\Big)\frac{\mathcal{K}(Q,\mathbb{P})}{n+1} + \frac{3\gamma E^2\mathcal{K}(Q,\mathbb{P})^2}{2\log M(n+1)} \qquad (22)$$

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

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

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 [LR13] the convexity of the function H: $\pi \to R(f_\pi) + 8C_b^2V(\pi)/(\gamma E^2)$ when $\gamma > 16C_b^2/(C_\ell E^2)$. Moreover, if one denotes π^* a minimizer of H, we have

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

Now, using C_ℓ-strong convexity as in Proposition 2 of [LR13], 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

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

Choosing $\gamma^* = 64C_b^2/(E^2C_\ell)$, integrating with respect to Q and using the estimate in (22) we derive that

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

Plugging in this estimate into (22) we obtain

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

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 second order bound of . 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 to derive an oracle inequality from a second order bound on the cumulative predictive risk. This framework is very classical in mathematical statistics and lower bounds are provided by [Tsy03].

The constants $(C_i)_{i=1,\dots,5}$, are very large. It is the price to pay for adaptivity of the procedure. We obtain much smaller 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* $0 < \eta < C_\ell/(24C_b^2)$, *with probability* $1 - e^{-x}$ *we have*

$$R(\tilde{f}) \leq \min_{1\leq j\leq M} \Big\{ R(f_j) + \frac{C_\eta}{n+1}\Big(\frac{2\log(\pi_{j,0}^{-1})}{\eta} + \Big(\frac{1}{\eta} + 2\eta(C_bD)^2\Big)x\Big)\Big\}$$

where $C_\eta = 1 + \frac{6C_b^2\eta}{C_\ell/2 - 12C_b^2\eta}$.

Proof. The proof starts from the result of Theorem 3.1 and an application of the variational form of the Bernstein's inequality of [Fre75] on $\ell_{j,t}$ (instead of the empirical one used in the proof of Theorem 4.1), for any $\eta < E^{-1}$:

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

Here we used the notation $\gamma = \eta E^2$. Then we use the variational form of the Bernstein's inequality on $\mathbb{E}_\pi[\ell_{j,t}^2]/E^2$ as in (21) to obtain

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

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

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

We conclude the proof using that $E \leq C_bD$, $\eta < C_\ell/(24C_b^2) \leq (12C_bD)^{-1} < (C_bD)^{-1}$ and choosing Q similarly than at the end of the proof of Theorem 4.1. □

The batch version of the BOA procedure is non adaptive in the sense that it depends on the constants appearing in the condition **(LIST)** via the restriction on the learning rate η but it is still independent of the observations (except that the range of the predictions D must be known). The parameter η can be considered as the inverse of the temperature β of the Q-aggregation procedure studied in [LR13]. In the Q-aggregation, the extra parameter β is required to be larger than $40C_b^2/C_\ell$. It is a condition similar to our restriction $0 < \eta < C_\ell/(24C_b^2)$. Then, we see that the batch version of BOA is comparable with the Q-aggregation and the oracle inequalities satisfied by both procedures have constants of similar order. Indeed, under the more restrictive assumptions on $\beta = 1/\eta$ of [LR13] we derive the same oracle inequality than in their Theorem A with different numerical constants:

$$R(\tilde{f}) \leq \min_{1\leq j\leq M} \Big\{ R(f_j) + 3.5\frac{\beta}{n+1}\log(\pi_{j,0}^{-1}))\Big\} + 1.76\frac{\beta x}{n+1}.$$

Finally, we remark that the three versions of BOA are explicitly computed with complexity $O(Mn)$. It is a practical advantage of BOA, even in its non adaptive form, compared with the Q-aggregation procedure studied in [LR13] that requires an optimization routine.

Aknowledgments

I would like to thank Gilles Stoltz for valuable comments on a preliminary version.

References

[AABR09] J. Abernethy, A. Agarwal, P.L. Bartlett, and A. Rakhlin, *A stochastic view of optimal regret through minimax duality*, COLT, 2009.

[AD13] A. Agarwal and J. C. Duchi, *The generalization ability of online algorithms for dependent data*, Information Theory, IEEE Trans. **59** (2013), 573–587.

[ALW13] P. Alquier, X. Li, and O. Wintenberger, *Prediction of time series by statistical learning: General losses and fast rates*, Dependence Modeling **1** (2013), 65–93.

[AMS06] J. Y. Audibert, R. Munos, and C. Szepesvari, *Use of variance estimation in the multi-armed bandit problem*, NIPS, 2006.

[Aud07] J.-Y. Audibert, *Progressive mixture rules are deviation suboptimal*, Advances in Neural Information Processing Systems, 2007, pp. 41–48.

[Aud09] ———, *Fast learning rates in statistical inference through aggregation*, Ann. Stat. **37** (2009), 1591–1646.

[Cat04] O. Catoni, *Statistical learning theory and stochastic optimization*, Notes in Mathematics (Saint-Flour Summer School on Probability Theory 2001), Springer, 2004.

[Cat07] ———, *Pac-bayesian supervised classification: the thermodynamics of statistical learning*, Institute of Mathematical Statistics, Beachwood, OH, 2007.

[CBL06] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*, Cambridge University Press, NY, USA, 2006.

[CBMS07] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz, *Improved second-order bounds for prediction with expert advice*, Mach Learn **66** (2007), 321–352.

[DRXZ12] D. Dai, P. Rigollet, L. Xia, and T. Zhang, *Deviation optimal learning using greedy Q-aggregation.*, Ann. Stat. **40** (2012), 1878–1905.

[DV75] M. D. Donsker and S. S. Varadhan, *Asymptotic evaluation of certain markov process expectations for large time, I*, Commun. Pure Appl. Math. **28** (1975), 1–47.

[Fre75] D. A. Freedman, *On tail probabilities for martingales*, Ann. Probab. **3** (1975), 100–118.

[Ger13] S. Gerchinovitz, *Sparsity regret bounds for individual sequences in online linear regression*, The Journal of Machine Learning Research **14** (2013), 729–769.

[GSVE14] P. Gaillard, G. Stoltz, and T. Van Erven, *A second-order bound with excess losses*, arXiv preprint arXiv:1402.2044, 2014.

[HK10] E. Hazan and S. Kale, *Extracting certainty from uncertainty: Regret bounded by variation in costs*, Mach Learn **80** (2010), 165–188.

[HKW98] D. Haussler, J. Kivinen, and M. K. Warmuth, *Sequential prediction of individual sequences under general loss functions*, Information Theory, IEEE Transactions **44** (1998), 1906–1925.

[JRT08] A. Juditsky, P. Rigollet, and A. B. Tsybakov, *Learning by mirror averaging*, Ann. Stat. **36** (2008), 2183–2206.

[KT08] S. M. Kakade and A. Tewari, *On the generalization ability of online strongly convex programming algorithms*, NIPS, 2008.

[KW99] J. Kivinen and M.K. Warmuth, *Averaging expert predictions*, COLT, 1999.

[Lec07] G. Lecué, *Optimal rates of aggregation in classification under low noise assumption*, Bernoulli **13** (2007), 1000–1022.

[LM09] G. Lecué and S. Mendelson, *Aggregation via empirical risk minimization*, Probab. theory and related fields **145** (2009), 591–613.

[LR13] G. Lecué and P. Rigollet, *Optimal learning with Q-aggregation*, arXiv preprint arXiv:1301.6080., 2013.

[Mar96] K. Marton, *Bounding $\overline{d}$-distance by informational divergence: a method to prove measure concentration.*, Ann. Probab. **24** (1996), 857–866.

[MP09] A. Maurer and M. Pontil, *Empirical bernstein bounds and sample variance penalization*, COLT, 2009.

[MR10] M. Mohri and A. Rostamizadeh, *Stability bounds and for ϕ-mixing and β-mixing processes*, JMLR **4** (2010), 1–26.

[Tsy03] A. B. Tsybakov, *Optimal rates of aggregation*, Learning Theory and Kernel Machines, Springer Berlin Heidelberg, 2003.

[Vov90] V.G. Vovk, *Aggregating strategies*, Proc. Third Workshop on Computational Learning Theory, 1990.

[Win12] O. Wintenberger, *Weak transport inequalities and applications to exponential inequalities and oracle inequalities*, arXiv preprint Arxiv:1207.4951, 2012.

[Zha05] T. Zhang, *Data dependent concentration bounds for sequential prediction algorithms*, Learning Theory, Springer Berlin Heidelberg, 2005.