# A randomized online learning algorithm for better variance control

Jean-Yves Audibert

CERTIS - Ecole des Ponts
19, rue Alfred Nobel - Cité Descartes
77455 Marne-la-Vallée - France
audibert@certis.enpc.fr

**Abstract.** We propose a sequential randomized algorithm, which at each step concentrates on functions having both low risk and low variance with respect to the previous step prediction function. It satisfies a simple risk bound, which is sharp to the extent that the standard statistical learning approach, based on supremum of empirical processes, does not lead to algorithms with such a tight guarantee on its efficiency. Our generalization error bounds complement the pioneering work of Cesa-Bianchi et al. [12] in which standard-style statistical results were recovered with tight constants using worst-case analysis.
A nice feature of our analysis of the randomized estimator is to put forward the links between the probabilistic and worst-case viewpoint. It also allows to recover recent model selection results due to Juditsky et al. [16] and to improve them in least square regression with heavy noise, i.e. when no exponential moment condition is assumed on the output.

## 1 Introduction

We are given a family $\mathcal{G}$ of functions and we want to learn from data a function that predicts as well as the best function in $\mathcal{G}$ up to some additive term called the rate of convergence. When the set $\mathcal{G}$ is finite, this learning task is often referred to as model selection aggregation.
This learning task has rare properties. First, in general an algorithm picking functions in the set $\mathcal{G}$ is not optimal (see e.g. [10, p.14]). This means that the estimator has to looked at an enlarged set of prediction functions. Secondly, in the statistical community, the only known optimal algorithms are all based on a Cesaro mean of Bayesian estimators (also referred to as progressive mixture rule). And thirdly, the proof of their optimality is not achieved by the most prominent tool in statistical learning theory: bounds on the supremum of empirical processes.
The idea of the proof, which comes back to Barron [5], is based on a chain rule and appeared to be successful for least square and entropy losses [9, 10, 6, 22, 7] and for general loss in [16].
In the online prediction with expert advice setting, without any probabilistic assumption on the generation of the data, appropriate weighting methods have been showed to behave as well as the best expert up to a minimax-optimal additive remainder term (see [18] and references

within). In this worst-case context, amazingly sharp constants have been found (see in particular [15, 12, 13, 23]). These results are expressed in cumulative loss and can be transposed to model selection aggregation to the extent that the expected risk of the randomized procedure based on sequential predictions is proportional to the expectation of the cumulative loss of the sequential procedure (see Lemma 3 for precise statement). This work presents a sequential algorithm, which iteratively updates a prior distribution put on the set of prediction functions. Contrarily to previously mentioned works, these updates take into account the variance of the task. As a consequence, posterior distributions concentrate on simultaneously low risk functions and functions close to the previous prediction. This conservative law is not surprising in view of previous works on high dimensional statistical tasks, such as wavelet thresholding, shrinkage procedures, iterative compression schemes ([3]), iterative feature selection ([1]).
The paper is organized as follows. Section 2 introduces the notation and the existing algorithms. Section 3 proposes a unifying setting to combine worst-case analysis tight results and probabilistic tools. It details our randomized estimator and gives a sharp expectation bound. In Sections 4 and 5, we show how to apply our main result under assumptions coming respectively from sequential prediction and model selection aggregation. Section 6 contains an algorithm that satisfies a sharp standard-style generalization error bound. To the author's knowledge, this bound is not achievable with classical statistical learning approach based on supremum of empirical processes. Section 7 presents an improved bound for least square regression regression when the noise has just a bounded moment of order $s \geq 2$.

## 2 Notation and existing algorithms

We assume that we observe $n$ pairs $Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n)$ of input-output and that each pair has been independently drawn from the same unknown distribution denoted $\mathbb{P}$. The input and output space are denoted respectively $\mathcal{X}$ and $\mathcal{Y}$, so that $\mathbb{P}$ is a probability distribution on the product space $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$. The target of a learning algorithm is to predict the output $Y$ associated to an input $X$ for pairs $(X, Y)$ drawn from the distribution $\mathbb{P}$. The quality of a prediction function $g : \mathcal{X} \to \mathcal{Y}$ is measured by the risk

$$R(g) \triangleq \mathbb{E}_{\mathbb{P}(dZ)} L(Z, g),$$

where $L(Z, g)$ assesses the loss of considering the prediction function $g$ on the data $Z \in \mathcal{Z}$. We use $L(Z, g)$ rather than $L[Y, g(X)]$ to underline that our results are not restricted to non-regularized losses, where we call non-regularized loss a loss that can be written as $l[Y, g(X)]$ for some function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.
We will say that the loss function is convex when the function $g \mapsto L(z, g)$ is convex for any $z \in \mathcal{Z}$. In this work, we do not assume the loss function to be convex except when it is explicitly mentioned.

For any $i \in \{0, \ldots, n\}$, the *cumulative loss* suffered by the prediction function $g$ on the first $i$ pairs of input-output, denoted $Z_1^i$ for short, is

$$\Sigma_i(g) \triangleq \sum_{j=1}^{i} L(Z_j, g),$$

where by convention we take $\Sigma_0$ identically equal to zero ($\Sigma_0 \equiv 0$). Throughout this work, without loss of generality, we assume that $\mathcal{Y}$ is convex so that convex combination of prediction functions are prediction functions. The symbol $C$ will denote some positive constant whose value may differ from line to line.

To handle possibly continuous set $\mathcal{G}$, we consider that $\mathcal{G}$ is a measurable space and that we have some prior distribution $\pi$ on it. The set of probability distributions on $\mathcal{G}$ will be denoted $\mathcal{M}$. The Kullback-Leibler divergence between a distribution $\rho \in \mathcal{M}$ and the prior distribution $\pi$ is

$$K(\rho, \pi) \triangleq \begin{cases} \mathbb{E}_{\rho(dg)} \log\left(\frac{\rho}{\pi}(g)\right) & \text{if } \rho \ll \pi, \\ +\infty & \text{otherwise} \end{cases}$$

where $\frac{\rho}{\pi}$ denotes the density of $\rho$ w.r.t. $\pi$ when it exists (i.e. $\rho \ll \pi$). For any $\rho \in \mathcal{M}$, we have $K(\rho, \pi) \geq 0$ and when $\pi$ is the uniform distribution on a finite set $\mathcal{G}$, we also have $K(\rho, \pi) \leq \log |\mathcal{G}|$. The Kullback-Leibler divergence satisfies the duality formula (see e.g. [11, p.10]): for any real-valued measurable function $h$ defined on $\mathcal{G}$,

$$\inf_{\rho \in \mathcal{M}} \left\{ \mathbb{E}_{\rho(dg)} h(g) + K(\rho, \pi) \right\} = -\log \mathbb{E}_{\pi(dg)} e^{-h(g)}. \tag{1}$$

and that the infimum is reached for the Gibbs distribution

$$\pi_{-h} \triangleq \frac{e^{-h(g)}}{\mathbb{E}_{\pi(dg')} e^{-h(g')}} \cdot \pi(dg). \tag{2}$$

The algorithm used to prove optimal convergence rates for several different losses (see e.g. [9, 10, 6, 22, 7, 16]) is the following:

**Algorithm** $A$**:** Let $\lambda > 0$. Predict according to $\frac{1}{n+1} \sum_{i=0}^{n} \mathbb{E}_{\pi_{-\lambda \Sigma_i}(dg)} g$, where we recall that $\Sigma_i$ maps a function $g \in \mathcal{G}$ to its cumulative loss up to time $i$.

In other words, for a new input $x$, the prediction of the output given by Algorithm $A$ is $\frac{1}{n+1} \sum_{i=0}^{n} \frac{\int g(x) e^{-\lambda \Sigma_i(g)} \pi(dg)}{\int e^{-\lambda \Sigma_i(g)} \pi(dg)}$.

From Vovk, Haussler, Kivinen and Warmuth works ([20, 15, 21]) and the link between cumulative loss in online setting and expected risk in the batch setting (see later Lemma 3), an "optimal" algorithm is:

**Algorithm** $B$**:** Let $\lambda > 0$. For any $i \in \{0, \ldots, n\}$, let $\hat{h}_i$ be a prediction function such that

$$\forall\, z \in \mathcal{Z} \qquad L(z, \hat{h}_i) \leq -\tfrac{1}{\lambda} \log \mathbb{E}_{\pi_{-\lambda \Sigma_i}(dg)} e^{-\lambda L(z,g)}.$$

If one of the $\hat{h}_i$ does not exist, the algorithm is said to fail. Otherwise it predicts according to $\frac{1}{n+1} \sum_{i=0}^{n} \hat{h}_i$.

In particular, for appropriate $\lambda > 0$, this algorithm does not fail when the loss function is the square loss (i.e. $L(z, g) = [y - g(x)]^2$) and when the output space is bounded. Algorithm $B$ is based on the same Gibbs distribution $\pi_{-\lambda\Sigma_i}$ as Algorithm $A$. Besides, in [15, Example 3.13], it is shown that Algorithm $A$ is not in general a particular case of Algorithm $B$, and that Algorithm $B$ will not generally produce a prediction function in the convex hull of $\mathcal{G}$ unlike Algorithm $A$. In Sections 4 and 5, we will see how both algorithms are connected to our generic algorithm.
We assume that the set, denoted $\bar{\mathcal{G}}$, of all measurable prediction functions has been equipped with a $\sigma$-algebra. Let $\mathcal{D}$ be the set of all probability distributions on $\bar{\mathcal{G}}$. By definition, a randomized algorithm produces a prediction function drawn according to a probability in $\mathcal{D}$. Let $\mathcal{P}$ be a set of probability distributions on $\mathcal{Z}$ in which we assume that the true unknown distribution generating the data is.

## 3 The algorithm and its generalization error bound

The aim of this section is to build an algorithm with the best possible convergence rate regardless of computational issues. For any $\lambda > 0$, let $\delta_\lambda$ be a real-valued function defined on $\mathcal{Z} \times \mathcal{G} \times \bar{\mathcal{G}}$ that satisfies

$$\forall\, \rho \in \mathcal{M} \quad \exists\, \hat{\pi}(\rho) \in \mathcal{D}$$
$$\sup_{\mathbb{P} \in \mathcal{P}} \left\{ \mathbb{E}_{\hat{\pi}(\rho)(dg')} \mathbb{E}_{\mathbb{P}(dZ)} \log \mathbb{E}_{\rho(dg)} e^{\lambda \left[ L(Z,g') - L(Z,g) - \delta_\lambda(Z,g,g') \right]} \right\} \leq 0. \tag{3}$$

Condition (3) is our probabilistic version of the generic algorithm condition in the online prediction setting (see [20, proof of Theorem 1] or more explicitly in [15, p.11]), in which we added the variance function $\delta_\lambda$. Our results will be all the sharper as this variance function is small. To make (3) more readable, let us say for the moment that

- without any assumption on $\mathcal{P}$, for several usual strongly convex loss functions, we may take $\delta_\lambda \equiv 0$ provided that $\lambda$ is a small enough constant (see Section 4).
- Inequality (3) can be seen as a "small expectation" inequality. The usual viewpoint is to control the quantity $L(Z, g)$ by its expectation with respect to (w.r.t.) $Z$ and a variance term. Here, roughly, $L(Z, g)$ is mainly controlled by $L(Z, g')$ where $g'$ is appropriately chosen through the choice of $\hat{\pi}(\rho)$, plus the additive term $\delta_\lambda$. By definition this additive term does not depend on the particular probability distribution generating the data and leads to empirical compensation.
- in the examples we will be interested in throughout this work, $\hat{\pi}(\rho)$ will be either equal to $\rho$ or to a Dirac distribution on some function, which is *not necessarily in* $\mathcal{G}$.
- for any loss function $L$, any set $\mathcal{P}$ and any $\lambda > 0$, one may choose $\delta_\lambda(Z, g, g') = \frac{\lambda}{2} \left[ L(Z, g) - L(Z, g') \right]^2$ (see Section 6).

Our results concern the following algorithm, in which we recall that $\pi$ is a prior distribution put on the set $\mathcal{G}$.

**Generic Algorithm:**

1. Let $\lambda > 0$. Define $\hat{\rho}_0 \triangleq \hat{\pi}(\pi)$ in the sense of (3) and draw a function $\hat{g}_0$ according to this distribution. Let $S_0(g) = 0$ for any $g \in \mathcal{G}$.
2. For any $i \in \{1, \dots, n\}$, iteratively define

$$S_i(g) \triangleq S_{i-1}(g) + L(Z_i, g) + \delta_\lambda(Z_i, g, \hat{g}_{i-1}) \qquad \text{for any } g \in \mathcal{G}.$$

   and

$$\hat{\rho}_i \triangleq \hat{\pi}(\pi_{-\lambda S_i}) \qquad \text{(in the sense of (3))}$$

   and draw a function $\hat{g}_i$ according to the distribution $\hat{\rho}_i$.
3. Predict with a function drawn according to the uniform distribution on $\{\hat{g}_0, \dots, \hat{g}_n\}$.

*Remark 1.* When $\delta_\lambda(Z, g, g')$ does not depend on $g$, we recover a more standard-style algorithm to the extent that we then have $\pi_{-\lambda S_i} = \pi_{-\lambda \Sigma_i}$. Precisely our algorithm becomes the randomized version of Algorithm $A$. When $\delta_\lambda(Z, g, g')$ depends on $g$, the posterior distributions tend to concentrate on functions having small risk and small variance term.

For any $i \in \{0, \dots, n\}$, the quantities $S_i$, $\hat{\rho}_i$ and $\hat{g}_i$ depend on the training data only through $Z_1, \dots, Z_i$. Let $\Omega_i$ denote the joint distribution of $\hat{g}_0^i \triangleq (\hat{g}_0, \dots, \hat{g}_i)$ conditional to $Z_1^i$, where we recall that $Z_1^i$ denotes $(Z_1, \dots, Z_i)$. Our randomized algorithm produces a prediction function which has three causes of randomness: the training data, the way $\hat{g}_i$ is obtained (step 2) and the uniform draw (step 3). So the expected risk of our iteratively randomized generic procedure is

$$\mathcal{E} \triangleq \mathbb{E}_{\mathbb{P}(dZ_1^n)} \mathbb{E}_{\Omega_n(d\hat{g}_0^n)} \frac{1}{n+1} \sum_{i=0}^n R(\hat{g}_i) = \frac{1}{n+1} \sum_{i=0}^n \mathbb{E}_{\mathbb{P}(dZ_1^i)} \mathbb{E}_{\Omega_i(d\hat{g}_0^i)} R(\hat{g}_i)$$

Our main result is

**Theorem 1.** *Let $\Delta_\lambda(g, g') \triangleq \mathbb{E}_{\mathbb{P}(dZ)} \delta_\lambda(Z, g, g')$ for $g \in G$ and $g' \in \bar{\mathcal{G}}$. The expected risk of the generic algorithm satisfies*

$$\mathcal{E} \leq \min_{\rho \in \mathcal{M}} \left\{ \mathbb{E}_{\rho(dg)} R(g) + \mathbb{E}_{\rho(dg)} \mathbb{E}_{\mathbb{P}(dZ_1^n)} \mathbb{E}_{\Omega_n(d\hat{g}_0^n)} \frac{\sum_{i=0}^n \Delta_\lambda(g, \hat{g}_i)}{n+1} + \frac{K(\rho, \pi)}{\lambda(n+1)} \right\} \tag{4}$$

*In particular, when $\mathcal{G}$ is finite and when the loss function $L$ and the set $\mathcal{P}$ are such that $\delta_\lambda \equiv 0$, by taking $\pi$ uniform on $\mathcal{G}$, we get*

$$\mathcal{E} \leq \min_{\mathcal{G}} R + \frac{\log |\mathcal{G}|}{\lambda(n+1)} \tag{5}$$

*Proof.* Let $Z_{n+1} \in \mathcal{Z}$ be drawn according to $\mathbb{P}$ and independent from $Z_1, \dots, Z_n$. To shorten formulae, let $\hat{\pi}_i \triangleq \pi_{-\lambda S_i}$ so that by definition we have $\hat{\rho}_i = \hat{\pi}(\hat{\pi}_i)$ in the sense of (3). Inequality (3) implies that

$$\mathbb{E}_{\hat{\pi}(\rho)(dg')} R(g') \leq -\frac{1}{\lambda} \mathbb{E}_{\hat{\pi}(\rho)(dg')} \mathbb{E}_{\mathbb{P}(dZ)} \log \mathbb{E}_{\rho(dg)} e^{-\lambda[L(Z,g) + \delta_\lambda(Z,g,g')]},$$

so by Fubini's theorem for any $i \in \{0, \dots, n\}$,

$$\mathbb{E}_{\mathbb{P}(dZ_1^i)} \mathbb{E}_{\Omega_i(d\hat{g}_0^i)} R(\hat{g}_i) \leq -\frac{1}{\lambda} \mathbb{E}_{\mathbb{P}(dZ_1^{i+1})} \mathbb{E}_{\Omega_i(d\hat{g}_0^i)} \log \mathbb{E}_{\hat{\pi}_i(dg)} e^{-\lambda[L(Z_{i+1}, g) + \delta_\lambda(Z_{i+1}, g, \hat{g}_i)]}.$$

Consequently, by the chain rule (i.e. cancellation in the sum of logarithmic terms; [5]) and by intensive use of Fubini's theorem, we get

$$\begin{aligned}
\mathcal{E} &= \tfrac{1}{n+1}\sum_{i=0}^{n}\mathbb{E}_{\mathbb{P}(dZ_1^i)}\mathbb{E}_{\Omega_i(d\hat{g}_0^i)}R(\hat{g}_i)\\
&\leq -\tfrac{1}{\lambda(n+1)}\sum_{i=0}^{n}\mathbb{E}_{\mathbb{P}(dZ_1^{i+1})}\mathbb{E}_{\Omega_i(d\hat{g}_0^i)}\log\mathbb{E}_{\hat{\pi}_i(dg)}e^{-\lambda[L(Z_{i+1},g)+\delta_\lambda(Z_{i+1},g,\hat{g}_i)]}\\
&= -\tfrac{1}{\lambda(n+1)}\mathbb{E}_{\mathbb{P}(dZ_1^{n+1})}\mathbb{E}_{\Omega_n(d\hat{g}_0^n)}\sum_{i=0}^{n}\log\mathbb{E}_{\hat{\pi}_i(dg)}e^{-\lambda[L(Z_{i+1},g)+\delta_\lambda(Z_{i+1},g,\hat{g}_i)]}\\
&= -\tfrac{1}{\lambda(n+1)}\mathbb{E}_{\mathbb{P}(dZ_1^{n+1})}\mathbb{E}_{\Omega_n(d\hat{g}_0^n)}\sum_{i=0}^{n}\log\left(\frac{\mathbb{E}_{\pi(dg)}e^{-\lambda S_{i+1}(g)}}{\mathbb{E}_{\pi(dg)}e^{-\lambda S_i(g)}}\right)\\
&= -\tfrac{1}{\lambda(n+1)}\mathbb{E}_{\mathbb{P}(dZ_1^{n+1})}\mathbb{E}_{\Omega_n(d\hat{g}_0^n)}\log\left(\frac{\mathbb{E}_{\pi(dg)}e^{-\lambda S_{n+1}(g)}}{\mathbb{E}_{\pi(dg)}e^{-\lambda S_0(g)}}\right)\\
&= -\tfrac{1}{\lambda(n+1)}\mathbb{E}_{\mathbb{P}(dZ_1^{n+1})}\mathbb{E}_{\Omega_n(d\hat{g}_0^n)}\log\mathbb{E}_{\pi(dg)}e^{-\lambda S_{n+1}(g)}
\end{aligned}$$

Now from the following lemma, we obtain

$$\begin{aligned}
\mathcal{E} &\leq -\tfrac{1}{\lambda(n+1)}\log\mathbb{E}_{\pi(dg)}e^{-\lambda\mathbb{E}_{\mathbb{P}(dZ_1^{n+1})}\mathbb{E}_{\Omega_n(d\hat{g}_0^n)}S_{n+1}(g)}\\
&= -\tfrac{1}{\lambda(n+1)}\log\mathbb{E}_{\pi(dg)}e^{-\lambda\left[(n+1)R(g)+\mathbb{E}_{\mathbb{P}(dZ_1^n)}\mathbb{E}_{\Omega_n(d\hat{g}_0^n)}\sum_{i=0}^{n}\Delta_\lambda(g,\hat{g}_i)\right]}\\
&= \min_{\rho\in\mathcal{M}}\left\{\mathbb{E}_{\rho(dg)}R(g)+\mathbb{E}_{\rho(dg)}\mathbb{E}_{\mathbb{P}(dZ_1^n)}\mathbb{E}_{\Omega_n(d\hat{g}_0^n)}\frac{\sum_{i=0}^{n}\Delta_\lambda(g,\hat{g}_i)}{n+1}+\tfrac{K(\rho,\pi)}{\lambda(n+1)}\right\}.
\end{aligned}$$

**Lemma 1.** *Let $\mathcal{W}$ be a real-valued measurable function defined on a product space $\mathcal{A}_1\times\mathcal{A}_2$ and let $\mu_1$ and $\mu_2$ be probability distribtutions on respectively $\mathcal{A}_1$ and $\mathcal{A}_2$ such that $\mathbb{E}_{\mu_1(da_1)}\log\mathbb{E}_{\mu_2(da_2)}e^{-\mathcal{W}(a_1,a_2)}<+\infty$. We have*

$$-\mathbb{E}_{\mu_1(da_1)}\log\mathbb{E}_{\mu_2(da_2)}e^{-\mathcal{W}(a_1,a_2)}\leq-\log\mathbb{E}_{\mu_2(da_2)}e^{-\mathbb{E}_{\mu_1(da_1)}\mathcal{W}(a_1,a_2)}.$$

*Proof.* It mainly comes from (1) (used twice) and Fubini's theorem.

Inequality (5) is a direct consequence of (4).

Theorem 1 bounds the expected risk of a randomized procedure, where the expectation is taken w.r.t. both the training set distribution and the randomizing distribution. From the following lemma, for convex loss functions, (5) implies

$$\mathbb{E}_{\mathbb{P}(dZ_1^n)}R\Big(\mathbb{E}_{\Omega_n(d\hat{g}_0^n)}\tfrac{1}{n+1}\sum_{i=0}^{n}\hat{g}_i\Big)\leq\min_{\mathcal{G}}R+\tfrac{\log|\mathcal{G}|}{\lambda(n+1)},\tag{6}$$

where we recall that $\Omega_n$ is the distribution of $\hat{g}_0^n=(\hat{g}_0,\ldots,\hat{g}_n)$ and $\lambda$ is a parameter whose typical value is the largest $\lambda>0$ such that $\delta_\lambda\equiv 0$.

**Lemma 2.** *For convex loss functions, the doubly expected risk of a randomized algorithm is greater than the expected risk of the deterministic version of the randomized algorithm, i.e. if $\hat{\rho}$ denotes the randomizing distribution,*

$$\mathbb{E}_{\mathbb{P}(Z_1^n)}R(\mathbb{E}_{\hat{\rho}(dg)}g)\leq\mathbb{E}_{\mathbb{P}(Z_1^n)}\mathbb{E}_{\hat{\rho}(dg)}R(g).$$

*Proof.* The result is a direct consequence of Jensen's inequality.

In [12], the authors rely on worst-case analysis to recover standard-style statistical results such as Vapnik's bounds [19]. Theorem 1 can be seen as a complement to this pioneering work. Inequality (6) is the model

selection bound that is well-known for least square regression and entropy loss, and that has been recently proved for general losses in [16].
Let us discuss the generalized form of the result. The r.h.s. of (4) is a classical regularized risk, which appears naturally in the PAC-Bayesian approach (see e.g. [8, 11, 4, 24]). An advantage of stating the result this way is to be able to deal with uncountable infinite $\mathcal{G}$. Even when $\mathcal{G}$ is countable, this formulation has some benefit to the extent that for any measurable function $h : \mathcal{G} \to \mathbb{R}$, $\min_{\rho \in \mathcal{M}} \{\mathbb{E}_{\rho(dg)} h(g) + K(\rho, \pi)\} \leq \min_{g \in \mathcal{G}} \{h(g) + \log \pi^{-1}(g)\}$.
Our generalization error bounds depend on two quantities $\lambda$ and $\pi$ which are the parameters of our algorithm. Their choice depends on the precise setting. Nevertheless, when $\mathcal{G}$ is finite and with no special structure a priori, a natural choice for $\pi$ is the uniform distribution on $\mathcal{G}$.
Once the distribution $\pi$ is fixed, an appropriate choice for the parameter $\lambda$ is the minimizer of the r.h.s. of (4). This minimizer is unknown by the statistician, and it is an open problem to adaptively choose $\lambda$ close to it.

## 4 Link with sequential prediction

This section aims at illustrating condition (3) and at clearly stating in our batch setting results coming from the online learning community. In [20, 15, 21], the loss function is assumed to satisfy: there are positive numbers $\eta$ and $c$ such that

$$\forall \rho \in \mathcal{M} \quad \exists g_\rho : \mathcal{X} \to \mathcal{Y} \quad \forall x \in \mathcal{X} \quad \forall y \in \mathcal{Y} \quad L[(x,y), g_\rho] \leq -\tfrac{c}{\eta} \log \mathbb{E}_{\rho(dg)} e^{-\eta L[(x,y),g]} \tag{7}$$

Then (3) holds both for $\lambda = \eta$ and $\delta_\lambda(Z, g, g') = -(1 - 1/c)L(Z, g')$ and for $\lambda = \eta/c$ and $\delta_\lambda(Z, g, g') = (c-1)L(Z, g)$, and we may take in both cases $\hat{\pi}(\rho)$ as the Dirac distribution at $g_\rho$. This leads to the *same* procedure which is described in the following straightforward corollary of Theorem 1.

**Corollary 1.** *Let $g_{\pi_{-\eta\Sigma_i}}$ be defined in the sense of (7). Consider the algorithm which predicts by drawing a function in $\{g_{\pi_{-\eta\Sigma_0}}, \ldots, g_{\pi_{-\eta\Sigma_n}}\}$ according to the uniform distribution. Under assumption (7), the expected risk of this procedure satisfies*

$$\mathcal{E} \leq c \min_{\rho \in \mathcal{M}} \left\{ \mathbb{E}_{\rho(dg)} R(g) + \tfrac{K(\rho,\pi)}{\eta(n+1)} \right\}. \tag{8}$$

This result is not surprising in view of the following two results. The first one comes from worst case analysis in sequential prediction.

**Theorem 2 (Haussler et al. [15], Theorem 3.8).** *Let $\mathcal{G}$ be countable. For any $g \in \mathcal{G}$, let $\Sigma_i(g) = \sum_{j=1}^{i} L(Z_j, g)$ (still) denote the cumulative loss up to time $i$ of the expert which always predict according to function $g$. The cumulative loss on $Z_1^n$ of the strategy in which the prediction at time $i$ is done according to $g_{\pi_{-\eta\Sigma_{i-1}}}$ in the sense of (7) is bounded by*

$$\inf_{g \in \mathcal{G}} \{c\Sigma_n(g) + \tfrac{c}{\eta} \log \pi^{-1}(g)\}. \tag{9}$$

The second result shows how the previous bound can be transposed into our model selection context by the following lemma.

**Lemma 3.** *Let $Z_{n+1}$ be a random variable independent from $Z_1^n$ and with the same distribution $\mathbb{P}$. Let $\mathcal{A}$ be a learning algorithm which produces the prediction function $\mathcal{A}(Z_1^i)$ at time $i+1$, i.e. from the data $Z_1^i = (Z_1, \ldots, Z_i)$. Let $\mathcal{L}$ be the randomized algorithm which produces a prediction function $\mathcal{L}(Z_1^n)$ drawn according to the uniform distribution on $\{\mathcal{A}(\emptyset), \mathcal{A}(Z_1), \ldots, \mathcal{A}(Z_1^n)\}$. The (doubly) expected risk of $\mathcal{L}$ is equal to $\frac{1}{n+1}$ times the expectation of the cumulative loss of $\mathcal{A}$ on the sequence $Z_1^{n+1}$.*

*Proof.* By Fubini's theorem, we have

$$\begin{aligned}\mathbb{E}R[\mathcal{L}(Z_1^n)] &= \tfrac{1}{n+1}\textstyle\sum_{i=0}^n \mathbb{E}_{\mathbb{P}(dZ_1^n)} R[\mathcal{A}(Z_1^i)] \\ &= \tfrac{1}{n+1}\textstyle\sum_{i=0}^n \mathbb{E}_{\mathbb{P}(dZ_1^{i+1})} L[Z_{i+1}, \mathcal{A}(Z_1^i)] \\ &= \tfrac{1}{n+1}\mathbb{E}_{\mathbb{P}(dZ_1^{n+1})} \textstyle\sum_{i=0}^n L[Z_{i+1}, \mathcal{A}(Z_1^i)].\end{aligned}$$

For any $\eta > 0$, let $c(\eta)$ denote the infimum of the $c$ for which (7) holds. Under weak assumptions, Vovk ([21]) proved that the infimum exists and studied the behaviour of $c(\eta)$ and $a(\eta) = c(\eta)/\eta$, which are key quantities of (8) and (9). Under weak assumptions, and in particular in the examples given in the table, the optimal constants in (9) are $c(\eta)$ and $a(\eta)$ ([21, Theorem 1]) and we have $c(\eta) \geq 1$, $\eta \mapsto c(\eta)$ nondecreasing and $\eta \mapsto a(\eta)$ nonincreasing. From these last properties, we understand the trade-off which occurs to choose the optimal $\eta$. Table 1 specifies (8) in different well-known learning tasks. For instance, for bounded least square regression (i.e. when $|Y| \leq B$ for some $B > 0$), the generalization error of the algorithm described in Corollary 1 when $\eta = 1/(2B^2)$ is bounded with $\min_{\rho \in \mathcal{M}} \left\{ \mathbb{E}_{\rho(dg)} R(g) + 2B^2 \frac{K(\rho,\pi)}{n+1} \right\}$.

| | Output space | Loss L(Z,g) | $c(\eta)$ |
|---|---|---|---|
| Entropy loss [15, Example 4.3] | $\mathcal{Y} = [0;1]$ | $Y \log\left(\frac{Y}{g(X)}\right)$ $+(1-Y)\log\left(\frac{1-Y}{1-g(X)}\right)$ | $c(\eta) = 1$ if $\eta \leq 1$ $c(\eta) = \infty$ if $\eta > 1$ |
| Absolute loss game [15, Section 4.2] | $\mathcal{Y} = [0;1]$ | $\lvert Y - g(X)\rvert$ | $\frac{\eta}{2\log[2/(1+e^{-\eta})]}$ $= 1 + \eta/4 + o(\eta)$ |
| Square loss [15, Example 4.4] | $\mathcal{Y} = [-B, B]$ | $[Y - g(X)]^2$ | $c(\eta) = 1$ if $\eta \leq 1/(2B^2)$ $c(\eta) = +\infty$ if $\eta > 1/(2B^2)$ |

**Table 1.** Value of $c(\eta)$ for different loss functions. Here $B$ denotes a positive real.

## 5 Model selection aggregation under Juditsky, Rigollet and Tsybakov assumptions ([16])

The main result of [16] relies on the following assumption on the loss function $L$ and the set $\mathcal{P}$ of probability distributions on $\mathcal{Z}$ in which we

assume that the true distribution is. There exist $\lambda > 0$ and a real-valued function $\psi$ defined on $\mathcal{G} \times \mathcal{G}$ such that for any $\mathbb{P} \in \mathcal{P}$

$$\begin{cases} \mathbb{E}_{\mathbb{P}(dZ)} e^{\lambda[L(Z,g')-L(Z,g)]} \leq \psi(g',g) & \text{for any } g, g' \in \mathcal{G} \\ \psi(g,g) = 1 \qquad \text{for any } g \in \mathcal{G} \\ \text{the function } \big[g \mapsto \psi(g',g)\big] \text{ is concave for any } g' \in \mathcal{G} \end{cases} \tag{10}$$

Theorem 1 gives the following result.

**Corollary 2.** *Under assumption* (10)*, the algorithm which draws uniformly its prediction function in the set* $\{\mathbb{E}_{\pi_{-\lambda\Sigma_0}(dg)} g, \ldots, \mathbb{E}_{\pi_{-\lambda\Sigma_n}(dg)} g\}$ *satisfies*

$$\mathcal{E} \leq \min_{\rho \in \mathcal{M}} \big\{ \mathbb{E}_{\rho(dg)} R(g) + \tfrac{K(\rho,\pi)}{\lambda(n+1)} \big\}. \tag{11}$$

*Besides for convex losses,*

$$R\Big( \tfrac{1}{n+1} \textstyle\sum_{i=0}^{n} \mathbb{E}_{\pi_{-\lambda\Sigma_i}(dg)} g \Big) \leq \min_{\rho \in \mathcal{M}} \big\{ \mathbb{E}_{\rho(dg)} R(g) + \tfrac{K(\rho,\pi)}{\lambda(n+1)} \big\}. \tag{12}$$

*Proof.* We start by proving that condition (3) holds with $\delta_\lambda \equiv 0$, and that we may take $\pi(\rho)$ as the Dirac distribution at the function $\mathbb{E}_{\rho(dg)} g$. By using Jensen's inequality and Fubini's theorem, assumption (10) implies that

$$\begin{aligned} \mathbb{E}_{\pi(\rho)(dg')} \mathbb{E}_{\mathbb{P}(dZ)} \log \mathbb{E}_{\rho(dg)} & e^{\lambda[L(Z,g')-L(Z,g)]} \\ &= \mathbb{E}_{\mathbb{P}(dZ)} \log \mathbb{E}_{\rho(dg)} e^{\lambda[L(Z,\mathbb{E}_{\rho(dg')} g')-L(Z,g)]} \\ &\leq \log \mathbb{E}_{\rho(dg)} \mathbb{E}_{\mathbb{P}(dZ)} e^{\lambda[L(Z,\mathbb{E}_{\rho(dg')} g')-L(Z,g)]} \\ &\leq \log \mathbb{E}_{\rho(dg)} \psi(\mathbb{E}_{\rho(dg')} g', g) \\ &\leq \log \psi(\mathbb{E}_{\rho(dg')} g', \mathbb{E}_{\rho(dg)} g) \\ &= 0, \end{aligned}$$

so that we can apply Theorem 1. It remains to note that in this context our generic algorithm is the one described in the corollary.

In this context, our generic algorithm reduces to the randomized version of Algorithm $A$. From Lemma 2, for convex loss functions, (11) also holds for the risk of Algorithm $A$. Corollary 2 also shows that the risk bounds of [16, Theorem 3.2 and the examples of Section 4.2] hold with the same constants for our randomized algorithm (provided that the expected risk w.r.t. the training set distribution is replaced by the expected risk w.r.t. both training set and randomizing distributions).
On assumption (10) we should say that it does not a priori require the function $L$ to be convex. Nevertheless, any known relevant examples deal with strongly convex loss functions and we know that in general the assumption will not hold for SVM loss function and for absolute loss function (since $1/n$ model selection rate are in general not achievable for these loss functions).
One can also recover the results in [16, Theorem 3.1 and Section 4.1] by taking $\delta_\lambda(Z,g,g') = \mathbf{1}_{Z \in S}[\sup_{g \in \mathcal{G}} L(Z,g) - \inf_{g \in \mathcal{G}} L(Z,g)]$ with appropriate set $S \subset \mathcal{Z}$. Once more the aggregation procedure is different because of the randomization step but the generalization error bounds are identical.

## 6 A standard-style statistical bound

This section proposes new results of a different kind. In the previous sections, under convexity assumptions, we were able to achieve fast rates. Here we have assumption neither on the loss function nor on the probability generating the data. Nevertheless we show that our generic algorithm applied for $\delta_\lambda(Z,g,g') = \lambda[L(Z,g) - L(Z,g')]^2/2$ satisfies a sharp standard-style statistical bound.

**Theorem 3.** *Let $V(g,g') = \mathbb{E}_{\mathbb{P}(dZ)}\big\{[L(Z,g) - L(Z,g')]^2\big\}$. Our generic algorithm applied with $\delta_\lambda(Z,g,g') = \lambda[L(Z,g) - L(Z,g')]^2/2$ and $\hat{\pi}(\rho) = \rho$ satisfies*

$$\mathcal{E} \leq \min_{\rho\in\mathcal{M}} \left\{ \mathbb{E}_{\rho(dg)} R(g) + \tfrac{\lambda}{2} \mathbb{E}_{\rho(dg)} \mathbb{E}_{\mathbb{P}(dZ_1^n)} \mathbb{E}_{\Omega_n(d\hat{g}_0^n)} \tfrac{\sum_{i=0}^n V(g,\hat{g}_i)}{n+1} + \tfrac{K(\rho,\pi)}{\lambda(n+1)} \right\} \tag{13}$$

*Proof.* To check that (3) holds, it suffices to prove that for any $z \in \mathcal{Z}$, $\mathbb{E}_{\rho(dg')} \log \mathbb{E}_{\rho(dg)} e^{\lambda[L(z,g') - L(z,g)] - \frac{\lambda^2}{2}[L(z,g') - L(z,g)]^2} \leq 0$.
To shorten formulae, let $\alpha(g',g) \triangleq \lambda[L(z,g') - L(z,g)]$ By Jensen's inequality and the following symmetrization trick, the previous expectation is bounded with

$$\begin{aligned}
&\mathbb{E}_{\rho(dg')} \mathbb{E}_{\rho(dg)} e^{\alpha(g',g) - \frac{\alpha^2(g',g)}{2}} \\
&\quad \leq \tfrac{1}{2} \mathbb{E}_{\rho(dg')} \mathbb{E}_{\rho(dg)} e^{\alpha(g',g) - \frac{\alpha^2(g',g)}{2}} + \tfrac{1}{2} \mathbb{E}_{\rho(dg')} \mathbb{E}_{\rho(dg)} e^{-\alpha(g',g) - \frac{\alpha^2(g',g)}{2}} \\
&\quad \leq \mathbb{E}_{\rho(dg')} \mathbb{E}_{\rho(dg)} \cosh\big(\alpha(g,g')\big) e^{-\frac{\alpha^2(g',g)}{2}} \\
&\quad \leq 1
\end{aligned} \tag{14}$$

where in the last inequality we used the inequality $\cosh(t) \leq e^{t^2/2}$ for any $t \in \mathbb{R}$. The first result then follows from Theorem 1.

To make (13) more explicit and to obtain a generalization error bound in which the randomizing distribution does not appear in the r.h.s. of the bound, the following corollary considers a widely used assumption that relates the variance term to the excess risk. Precisely, from Theorem 3, we obtain (proof omitted of this extended abstract)

**Corollary 3.** *Under the generalized Mammen and Tsybakov's assumption which states that there exist $0 \leq \gamma \leq 1$ and a prediction function $\tilde{g}$ (not necessarily in $\mathcal{G}$) such that $V(g,\tilde{g}) \leq c[R(g) - R(\tilde{g})]^\gamma$ for any $g \in \mathcal{G}$, the expected risk of the generic algorithm used in Theorem 3 satisfies*

- *When $\gamma = 1$,*

$$\mathcal{E} - R(\tilde{g}) \leq \min_{\rho\in\mathcal{M}} \left\{ \tfrac{1+c\lambda}{1-c\lambda} \big[\mathbb{E}_{\rho(dg)} R(g) - R(\tilde{g})\big] + \tfrac{K(\rho,\pi)}{(1-c\lambda)\lambda(n+1)} \right\}$$

  *In particular, for $\mathcal{G}$ finite, $\pi$ the uniform distribution, $\lambda = 1/2c$, when $\tilde{g}$ belongs to $\mathcal{G}$, we get $\mathcal{E} \leq \min_{g\in\mathcal{G}} R(g) + \frac{4c\log|\mathcal{G}|}{n+1}$.*

- *When $\gamma < 1$, for any $0 < \beta < 1$ and for $\tilde{R}(g) \triangleq R(g) - R(\tilde{g})$,*

$$\mathcal{E} - R(\tilde{g}) \leq \left\{ \tfrac{1}{\beta} \left( \mathbb{E}_{\rho(dg)} [\tilde{R}(g) + c\lambda \tilde{R}^\gamma(g)] + \tfrac{K(\rho,\pi)}{\lambda(n+1)} \right) \right\} \vee \left( \tfrac{c\lambda}{1-\beta} \right)^{\frac{1}{1-\gamma}}.$$

To understand the sharpness of Theorem 3, we have to compare this result to the one coming from the traditional (PAC-Bayesian) statistical learning approach which relies on supremum of empirical processes.

**Theorem 4.** *We still use* $V(g,g') = \mathbb{E}_{\mathbb{P}(dZ)}\left\{[L(Z,g) - L(Z,g')]^2\right\}$. *The generalization error of the algorithm which draws its prediction function according to the Gibbs distribution* $\pi_{-\lambda\Sigma_n}$ *satisfies*

$$\begin{aligned}&\mathbb{E}_{\mathbb{P}(dZ_1^n)}\mathbb{E}_{\pi_{-\lambda\Sigma_n}(dg')}R(g')\\&\quad\le \min_{\rho\in\mathcal{M}}\Big\{\mathbb{E}_{\rho(dg)}R(g) + \tfrac{K(\rho,\pi)+1}{\lambda n} + \lambda\mathbb{E}_{\rho(dg)}\mathbb{E}_{\mathbb{P}(dZ_1^n)}\mathbb{E}_{\pi_{-\lambda\Sigma_n}(dg')}V(g,g')\\&\quad+\lambda\tfrac{1}{n}\textstyle\sum_{i=1}^n \mathbb{E}_{\rho(dg)}\mathbb{E}_{\mathbb{P}(dZ_1^n)}\mathbb{E}_{\pi_{-\lambda\Sigma_n}(dg')}[L(Z_i,g) - L(Z_i,g')]^2\Big\}.\end{aligned} \tag{15}$$

*Let* $\varphi$ *be the positive convex increasing function defined as* $\varphi(t) \triangleq \frac{e^t-1-t}{t^2}$ *and* $\varphi(0) = \frac{1}{2}$ *by continuity. When* $\sup_{g\in\mathcal{G},g'\in\mathcal{G}}|L(Z,g') - L(Z,g)| \le B$, *we also have*

$$\begin{aligned}\mathbb{E}_{\mathbb{P}(dZ_1^n)}\mathbb{E}_{\pi_{-\lambda\Sigma_n}(dg')}R(g') \le \min_{\rho\in\mathcal{M}}\Big\{&\mathbb{E}_{\rho(dg)}R(g)\\&+\lambda\varphi(\lambda B)\mathbb{E}_{\rho(dg)}\mathbb{E}_{\mathbb{P}(dZ_1^n)}\mathbb{E}_{\pi_{-\lambda\Sigma_n}(dg')}V(g,g') + \tfrac{K(\rho,\pi)+1}{\lambda n}\Big\}.\end{aligned} \tag{16}$$

*Proof.* Let us prove (16). Let $r(g)$ denote the empirical risk of $g \in \mathcal{G}$, that is $r(g) = \frac{\Sigma_n(g)}{n}$. Let $\rho \in \mathcal{M}$ be some fixed distribution on $\mathcal{G}$. From [3, Section 8.1], with probability at least $1-\epsilon$, for any $\mu \in \mathcal{M}$, we have

$$\begin{aligned}&\mathbb{E}_{\mu(dg')}R(g') - \mathbb{E}_{\rho(dg)}R(g)\\&\qquad\le \mathbb{E}_{\mu(dg')}r(g') - \mathbb{E}_{\rho(dg)}r(g)\\&\qquad\quad+\lambda\varphi(\lambda B)\mathbb{E}_{\mu(dg')}\mathbb{E}_{\rho(dg)}V(g,g') + \tfrac{K(\mu,\pi)+\log(\epsilon^{-1})}{\lambda n}.\end{aligned}$$

Since $\pi_{-\lambda\Sigma_n}$ minimizes $\mu \mapsto \mathbb{E}_{\mu(dg')}r(g') + \frac{K(\mu,\pi)}{\lambda n}$, we have

$$\begin{aligned}&\mathbb{E}_{\pi_{-\lambda\Sigma_n}(dg')}R(g')\\&\quad\le \mathbb{E}_{\rho(dg)}R(g) + \lambda\varphi(\lambda B)\mathbb{E}_{\pi_{-\lambda\Sigma_n}(dg')}\mathbb{E}_{\rho(dg)}V(g,g') + \tfrac{K(\rho,\pi)+\log(\epsilon^{-1})}{\lambda n}.\end{aligned}$$

Then we apply the following inequality: for any random variable $W$, $\mathbb{E}W \le \mathbb{E}(W\vee 0) = \int_0^{+\infty}\mathbb{P}(W\ge u)du = \int_0^1 \epsilon^{-1}\mathbb{P}(W\ge\log(\epsilon^{-1}))d\epsilon$. At last we may choose the distribution $\rho$ minimizing the upper bound to obtain (16). Similarly using [3, Section 8.3], we may prove (15).

*Remark 2.* By comparing (16) and (13), we see that the classical approach requires the quantity $\sup_{g\in\mathcal{G},g'\in\mathcal{G}}|L(Z,g')-L(Z,g)|$ to be bounded and the unpleasing function $\varphi$ appears. In fact, using technical small expectations theorems (see [2, Lemma 7.1]), exponential moments conditions on the above quantity would be sufficient.
The symmetrization trick used to prove Theorem 3 is performed in the prediction functions space. We do not call on the second virtual training set currently used in statistical learning theory (see [19]). Nevertheless both symmetrization tricks end up to the same nice property: we need no boundedness assumption on the loss functions. In our setting, symmetrization on training data leads to an unwanted expectation and to a constant four times larger (see the two variance terms of (15) and the discussion in [3, Section 8.3.3]). In particular, deducing from Theorem 4 a corollary similar to Corollary 3 is only possible through (16), because of the last variance term in (15) (since $\Sigma_n$ depends on $Z_i$).

## 7 Application to least square regression

This section shows that Theorem 1 used jointly with the symmetrization idea developed in the previous section allows to obtain improved convergence rates in heavy noise situation. We start with the following theorem concerning twice differentiable convex loss functions.

**Theorem 5.** *Let $B \geq b > 0$. Consider a loss function $L$ which can be written as $L[(x,y),g] = l[y,g(x)]$, where the function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is twice differentiable and convex w.r.t. the second variable. Let $l'$ and $l''$ denote respectively the first and second derivative of the function $l$ w.r.t. the second variable. Let $\Delta(y) = \sup_{|\alpha|\leq b, |\beta|\leq b} \big[l(y,\alpha) - l(y,\beta)\big]$. Assume that $\lambda_0 \triangleq \inf_{|y|\leq B, |y'|\leq b} \frac{l''(y,y')}{[l'(y,y')]^2} > 0$ and that $\sup_{g\in\mathcal{G}, x\in\mathcal{X}} |g(x)| \leq b$. For any $0 < \lambda \leq \lambda_0$, the algorithm which draws uniformly its prediction function among $\mathbb{E}_{\pi_{-\lambda\Sigma_0}(dg)} g, \ldots, \mathbb{E}_{\pi_{-\lambda\Sigma_n}(dg)} g$ satisfies*

$$\mathcal{E} \leq \min_{\rho\in\mathcal{M}} \Big\{ \mathbb{E}_{\rho(dg)} R(g) + \tfrac{K(\rho,\pi)}{\lambda(n+1)} \Big\} + \mathbb{E}\Big\{ \tfrac{\lambda\Delta^2(Y)}{2} \mathbf{1}_{\lambda\Delta(Y)<1;|Y|>B} + \big[\Delta(Y) - \tfrac{1}{2\lambda}\big] \mathbf{1}_{\lambda\Delta(Y)\geq 1;|Y|>B} \Big\}.$$

*Proof.* According to Theorem 1, it suffices to check that condition (3) holds for $0 < \lambda \leq \lambda_0$, $\hat{\pi}(\rho)$ the Dirac distribution at $\mathbb{E}_{\rho(dg)} g$ and

$$\begin{aligned}\delta_\lambda[(x,y),g,g'] = \delta_\lambda(y) &\triangleq \min_{0\leq\zeta\leq 1} \Big[\zeta\Delta(y) + \tfrac{(1-\zeta)^2\lambda\Delta^2(y)}{2}\Big] \mathbf{1}_{|y|>B} \\ &= \tfrac{\lambda\Delta^2(y)}{2} \mathbf{1}_{\lambda\Delta(y)<1;|y|>B} + \big[\Delta(y) - \tfrac{1}{2\lambda}\big] \mathbf{1}_{\lambda\Delta(y)\geq 1;|y|>B}.\end{aligned}$$

– For any $z = (x,y) \in \mathcal{Z}$ such that $|y| \leq B$, for any probability distribution $\rho$ and for the above values of $\lambda$ and $\delta_\lambda$, we have

$$\begin{aligned}\mathbb{E}_{\rho(dg)} e^{\lambda[L(z,\mathbb{E}_{\rho(dg')}g') - L(z,g) - \delta_\lambda(z,g,g')]} &= e^{\lambda L(z,\mathbb{E}_{\rho(dg')}g')} \mathbb{E}_{\rho(dg)} e^{-\lambda l[y,g(x)]} \\ &\leq e^{\lambda l[y,\mathbb{E}_{\rho(dg')}g'(x)] - \lambda l[y,\mathbb{E}_{\rho(dg)}g(x)]} = 1,\end{aligned}$$

where the inequality comes from the concavity of $y' \mapsto e^{-\lambda l(y,y')}$ for $\lambda \leq \lambda_0$. This concavity argument goes back to [17, Section 4], and was also used in [7] and in some of the examples given in [16].

– For any $z = (x,y) \in \mathcal{Z}$ such that $|y| > B$, for any $0 \leq \zeta \leq 1$, by using twice Jensen's inequality and then by using the symmetrization trick presented in Section 6, we have

$$\begin{aligned}&\mathbb{E}_{\rho(dg)} e^{\lambda[L(z,\mathbb{E}_{\rho(dg')}g') - L(z,g) - \delta_\lambda(z,g,g')]} \\ &= e^{-\delta_\lambda(y)} \mathbb{E}_{\rho(dg)} e^{\lambda[L(z,\mathbb{E}_{\rho(dg')}g') - L(z,g)]} \\ &\leq e^{-\delta_\lambda(y)} \mathbb{E}_{\rho(dg)} e^{\lambda[\mathbb{E}_{\rho(dg')}L(z,g') - L(z,g)]} \\ &\leq e^{-\delta_\lambda(y)} \mathbb{E}_{\rho(dg)} \mathbb{E}_{\rho(dg')} e^{\lambda[L(z,g') - L(z,g)]} \\ &= e^{-\delta_\lambda(y)} \mathbb{E}_{\rho(dg)} \mathbb{E}_{\rho(dg')} \exp\Big\{\lambda(1-\zeta)[L(z,g') - L(z,g)] \\ &\qquad -\tfrac{1}{2}\lambda^2(1-\zeta)^2[L(z,g') - L(z,g)]^2 \\ &\qquad +\lambda\zeta[L(z,g') - L(z,g)] + \tfrac{1}{2}\lambda^2(1-\zeta)^2[L(z,g') - L(z,g)]^2\Big\} \\ &\leq e^{-\delta_\lambda(y)} \mathbb{E}_{\rho(dg)} \mathbb{E}_{\rho(dg')} \exp\big\{\lambda(1-\zeta)[L(z,g') - L(z,g)] \\ &-\tfrac{1}{2}\lambda^2(1-\zeta)^2[L(z,g') - L(z,g)]^2 + \lambda\zeta\Delta(y) + \tfrac{1}{2}\lambda^2(1-\zeta)^2\Delta^2(y)\big\} \\ &\leq e^{-\delta_\lambda(y)} e^{\lambda\zeta\Delta(y) + \frac{1}{2}\lambda^2(1-\zeta)^2\Delta^2(y)}\end{aligned}$$

Taking $\zeta \in [0;1]$ minimizing the last r.h.s., we obtain that

$$\mathbb{E}_{\rho(dg)} e^{\lambda[L(z,\mathbb{E}_{\rho(dg')}g')-L(z,g)-\delta_\lambda(z,g,g')]} \leq 1$$

From the two previous computations, we obtain that for any $z \in \mathcal{Z}$,

$$\log \mathbb{E}_{\rho(dg)} e^{\lambda[L(z,\mathbb{E}_{\rho(dg')}g')-L(z,g)-\delta_\lambda(z,g,g')]} \leq 0,$$

so that condition (3) holds for the above values of $\lambda$, $\hat{\pi}(\rho)$ and $\delta_\lambda$, and the result follows from Theorem 1.

In particular, for least square regression, Theorem 5 can be stated as:

**Theorem 6.** *Assume that* $\sup_{g\in\mathcal{G},x\in\mathcal{X}}|g(x)| \leq b$ *for some* $b > 0$. *For any* $0 < \lambda \leq 1/(8b^2)$, *the algorithm which draws uniformly its prediction function among* $\mathbb{E}_{\pi_{-\lambda\Sigma_0}(dg)}g, \ldots, \mathbb{E}_{\pi_{-\lambda\Sigma_n}(dg)}g$ *satisfies:*

$$\begin{aligned}\mathcal{E} \leq \min_{\rho\in\mathcal{M}} \Big\{\mathbb{E}_{\rho(dg)}R(g) + \tfrac{K(\rho,\pi)}{\lambda(n+1)}\Big\} + \mathbb{E}\big\{\big(4b|Y| - \tfrac{1}{2\lambda}\big)\mathbf{1}_{|Y|\geq(4b\lambda)^{-1}}\big\} \\ + \mathbb{E}\big\{8\lambda b^2|Y|^2\mathbf{1}_{(2\lambda)^{-1/2}-b<|Y|<(4b\lambda)^{-1}}\big\}.\end{aligned} \tag{17}$$

*Proof.* The result follows from Theorem 5, computations of $\lambda_0 = \frac{1}{2(B+b)^2}$ and $\Delta(y) = 4b|y|$, and from the optimization of the parameter $B$.

Theorem 6 improves [16, Corollary 4.1] and [7, Theorem 1]. From it, we can deduce the following improvement of [16, Corollary 4.2].

**Corollary 4.** *Under the assumptions*

$$\begin{cases} \sup_{g\in\mathcal{G},x\in\mathcal{X}}|g(x)| \leq b & \text{for some } b > 0 \\ \mathbb{E}|Y|^s \leq A & \text{for some } s \geq 2 \text{ and } A > 0 \\ \mathcal{G} \text{ finite} \end{cases}$$

*for* $\lambda = C_1\big(\frac{\log|\mathcal{G}|}{n}\big)^{2/(s+2)}$ *where* $C_1 > 0$, *the algorithm which draws uniformly its prediction function among* $\mathbb{E}_{\pi_{-\lambda\Sigma_0}(dg)}g, \ldots, \mathbb{E}_{\pi_{-\lambda\Sigma_n}(dg)}g$ *satisfies*

$$\mathcal{E} \leq \min_{g\in\mathcal{G}} R(g) + C\big(\tfrac{\log|\mathcal{G}|}{n}\big)^{s/(s+2)} \tag{18}$$

*for a quantity* $C$ *which depends only on* $C_1$, $b$, $A$ *and* $s$.

*Proof.* The moment assumption on $Y$ implies $\alpha^{s-q}\mathbb{E}|Y|^q\mathbf{1}_{|Y|\geq\alpha} \leq A$ for any $0 \leq q \leq s$ and $\alpha \geq 0$. As a consequence, the second and third term of the r.h.s. of (17) are respectively bounded with $4bA(4b\lambda)^{s-1}$ and $8\lambda b^2 A(2\lambda)^{(s-2)/2}$, so that (17) can be weakened into $\mathcal{E} \leq \min_{g\in\mathcal{G}} R(g) + \frac{\log|\mathcal{G}|}{\lambda n} + C'\lambda^{s-1} + C''\lambda^{s/2}$ for $C' = A(4b)^s$ and $C'' = A2^{2+s/2}b^2$. This gives the desired result.

In particular, with the minimal assumption $\mathbb{E}|Y|^2 \leq A$ (i.e. $s = 2$), the convergence rate is of order $n^{-1/2}$, and at the opposite, when $s$ goes to infinity, we recover the $n^{-1}$ rate we have under exponential moment condition on the output.

## 8 Conclusion and open problems

A learning task can be defined by a set of reference prediction functions and a set of probability distributions in which we assume that the distribution generating the data is. In this work, we propose to summarize this learning problem by the variance function of the key condition (3). We have proved that our generic algorithm based on this variance function leads to optimal rates of convergence on the model selection aggregation problem, and that it gives a nice unified view to results coming from different communities. Our results concern expected risks and it is an open problem to provide corresponding tight exponential inequalities.
Besides without any assumption on the learning task, we proved a Bernstein's type bound which has no known equivalent form when the loss function is not assumed to be bounded. Nevertheless much work still has to be done to propose algorithms having better generalization error bounds that the ones based on supremum of empirical processes. For instance, in several learning tasks, Dudley's chaining trick [14] is the only way to prove risk convergence with the optimal rate. So a natural question and another open problem is whether it is possible to combine the better variance control presented here with the chaining argument (or other localization argument used while exponential inequalities are available).

**Acknowledgement.** I would like to thank Nicolas Vayatis, Alexandre Tsybakov, Gilles Stoltz and the referees for their very helpful comments.

## References

1. P. Alquier. Iterative feature selection in least square regression estimation. 2005. Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7.
2. J.-Y. Audibert. Aggregated estimators and empirical complexity for least square regression. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 40(6):685–736, 2004.
3. J.-Y. Audibert. A better variance control for PAC-Bayesian classification. Preprint n.905, http://www.proba.jussieu.fr/mathdoc/preprints/index.html, 2004. Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7.
4. J.-Y. Audibert. *PAC-Bayesian statistical learning theory.* PhD thesis, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2004.
5. A. Barron. Are bayes rules consistent in information? In T.M. Cover and B. Gopinath, editors, *Open Problems in Communication and Computation*, pages 85–91. Springer, 1987.
6. A. Barron and Y. Yang. Information-theoretic determination of minimax rates of convergence. *Ann. Stat.*, 27(5):1564–1599, 1999.
7. F. Bunea and A. Nobel. Sequential procedures for aggregating arbitrary estimators of a conditional mean, 2005. Technical report, Available from http://stat.fsu.edu/%7Eflori/ps/bnapril2005IEEE.pdf.

8. O. Catoni. *Statistical Learning Theory and Stochastic Optimization: Ecole d'été de Probabilités de Saint-Flour XXXI - 2001.* Lecture Notes in Mathematics. Springer Verlag.
9. O. Catoni. A mixture approach to universal model selection. preprint LMENS 97-30, Available from http://www.dma.ens.fr/edition/preprints/Index.97.html, 1997.
10. O. Catoni. Universal aggregation rules with exact bias bound. Preprint n.510, http://www.proba.jussieu.fr/mathdoc/preprints/index.html#1999, 1999.
11. O. Catoni. A PAC-Bayesian approach to adaptive classification. Preprint n.840, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2003.
12. N. Cesa-Bianchi, Y. Freund, D. Haussler, D.P. Helmbold, R.E. Schapire, and M.K. Warmuth. How to use expert advice. *J. ACM*, 44(3):427–485, 1997.
13. N. Cesa-Bianchi and G. Lugosi. On prediction of individual sequences. *Ann. Stat.*, 27(6):1865–1895, 1999.
14. R.M. Dudley. Central limit theorems for empirical measures. *Ann. Probab.*, 6:899–929, 1978.
15. D. Haussler, J. Kivinen, and M. K. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Trans. on Information Theory*, 44(5):1906–1925, 1998.
16. A. Juditsky, P. Rigollet, and A.B. Tsybakov. Learning by mirror averaging, 2005. Available from arxiv website.
17. J. Kivinen and M. K. Warmuth. Averaging expert predictions. *Lecture Notes in Computer Science*, 1572:153–167, 1999.
18. Merhav and Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44, 1998.
19. V. Vapnik. *The nature of statistical learning theory.* Springer-Verlag, second edition, 1995.
20. V.G. Vovk. Aggregating strategies. In *COLT '90: Proceedings of the third annual workshop on Computational learning theory*, pages 371–386, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc.
21. V.G. Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, pages 153–173, 1998.
22. Y. Yang. Combining different procedures for adaptive regression. *Journal of multivariate analysis*, 74:135–161, 2000.
23. R. Yaroshinsky, R. El-Yaniv, and S.S. Seiden. How to better use expert advice. *Mach. Learn.*, 55(3):271–309, 2004.
24. T. Zhang. Information theoretical upper and lower bounds for statistical estimation. *IEEE Transaction on Information Theory*, 2006. to appear.