No fast exponential deviation inequalities for the progressive mixture rule

Jean-Yves Audibert

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

Abstract. We consider the learning task consisting in predicting as well as the best function in a finite reference set $\mathcal{G}$ up to the smallest possible additive term. If $R(g)$ denotes the generalization error of a prediction function g, under reasonable assumptions on the loss function (typically satisfied by the least square loss when the output is bounded), it is known that the progressive mixture rule $\hat{g}$ satisfies

$$\mathbb{E}R(\hat{g}) \leq \min_{g\in\mathcal{G}} R(g) + C\frac{\log|\mathcal{G}|}{n}, \tag{1}$$

where n denotes the size of the training set, $\mathbb{E}$ denotes the expectation w.r.t. the training set distribution and C denotes a positive constant. This work mainly shows that for any training set size n, there exist $\epsilon > 0$, a reference set $\mathcal{G}$ and a probability distribution generating the data such that with probability at least ϵ

$$R(\hat{g}) \geq \min_{g\in\mathcal{G}} R(g) + c\sqrt{\frac{\log(|\mathcal{G}|\epsilon^{-1})}{n}},$$

where c is a positive constant. In other words, surprisingly, for appropriate reference set $\mathcal{G}$, the deviation convergence rate of the progressive mixture rule is only of order $1/\sqrt{n}$ while its expectation convergence rate is of order $1/n$. The same conclusion holds for the progressive indirect mixture rule. This work also emphasizes on the suboptimality of algorithms based on penalized empirical risk minimization on $\mathcal{G}$.

1 Setup and notation

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

$$R(g) = \mathbb{E}_{(X,Y)\sim P}\ \ell[Y, g(X)],$$

where $\ell[Y, g(X)]$ denotes the loss (possibly infinite) incurred by predicting $g(X)$ when the true output is Y. We work under the following assumptions for the data space and the loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$.

Main assumptions. The input space is assumed to be infinite: $|\mathcal{X}| = +\infty$. The output space is a non-trivial (i.e. infinite) interval of $\mathbb{R}$ symmetrical w.r.t. some $a \in \mathbb{R}$: for any $y \in \mathcal{Y}$, we have $2a - y \in \mathcal{Y}$. The loss function is

- *uniformly exp-concave:* there exists $\lambda > 0$ such that for any $y \in \mathcal{Y}$, the set $\{y' \in \mathbb{R} : \ell(y, y') < +\infty\}$ is an interval containing a on which the function $y' \mapsto e^{-\lambda \ell(y,y')}$ is concave.
- *symmetrical:* for any $y_1, y_2 \in \mathcal{Y}$, $\ell(y_1, y_2) = \ell(2a - y_1, 2a - y_2)$,
- *admissible:* for any $y, y' \in \mathcal{Y} \cap]a; +\infty[$, $\ell(y, 2a - y') > \ell(y, y')$,
- *well behaved at center:* for any $y \in \mathcal{Y} \cap]a; +\infty[$, the function $\ell_y : y' \mapsto \ell(y, y')$ is twice continuously differentiable on a neighborhood of a and $\ell'_y(a) < 0$.

These assumptions imply that

- $\mathcal{Y}$ has necessarily one of the following form: $]-\infty; +\infty[$, $[a - \zeta; a + \zeta]$ or $]a - \zeta; a + \zeta[$ for some $\zeta > 0$.
- for any $y \in \mathcal{Y}$, from the exp-concavity assumption, the function $\ell_y : y' \mapsto \ell(y, y')$ is convex on the interval on which it is finite[1]. As a consequence, the risk R is also a convex function (on the convex set of prediction functions for which it is finite).

The assumptions were motivated by the fact that they are satisfied in the following settings:

- least square loss with bounded outputs: $\mathcal{Y} = [y_{\min}; y_{\max}]$ and $\ell(y_1, y_2) = (y_1 - y_2)^2$. Then we have $a = (y_{\min} + y_{\max})/2$ and may take $\lambda = 1/[2(y_{\max} - y_{\min})^2]$.
- entropy loss: $\mathcal{Y} = [0; 1]$ and $\ell(y_1, y_2) = y_1 \log\left(\frac{y_1}{y_2}\right) + (1 - y_1) \log\left(\frac{1-y_1}{1-y_2}\right)$. Note that $\ell(0, 1) = \ell(1, 0) = +\infty$. Then we have $a = 1/2$ and may take $\lambda = 1$.
- exponential (or AdaBoost) loss: $\mathcal{Y} = [-y_{\max}; y_{\max}]$ and $\ell(y_1, y_2) = e^{-y_1 y_2}$. Then we have $a = 0$ and may take $\lambda = e^{-y_{\max}^2}$.
- logit loss: $\mathcal{Y} = [-y_{\max}; y_{\max}]$ and $\ell(y_1, y_2) = \log(1 + e^{-y_1 y_2})$. Then we have $a = 0$ and may take $\lambda = e^{-y_{\max}^2}$.

Progressive indirect mixture rule. Let $\mathcal{G}$ be a finite reference set of prediction functions. Under the previous assumptions, the only known algorithms satisfying (1) are the progressive indirect mixture rules defined below.

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

$$\Sigma_i(g) \triangleq \sum_{j=1}^{i} \ell[Y_j, g(X_j)],$$

[1] Indeed, if ξ denotes the function $e^{-\lambda \ell_y}$, from Jensen's inequality, for any probability distribution, $\mathbb{E}\ell_y(Y) = \mathbb{E}\big(-\frac{1}{\lambda}\log \xi(Y)\big) \geq -\frac{1}{\lambda}\log \mathbb{E}\xi(Y) \geq -\frac{1}{\lambda}\log \xi(\mathbb{E}Y) = \ell_y(\mathbb{E}Y)$.

where by convention we take $\varSigma_0 \equiv 0$. Let π denote the uniform distribution on $\mathcal{G}$. We define the probability distribution $\hat{\pi}_i$ on $\mathcal{G}$ as

$$\hat{\pi}_i \propto e^{-\lambda \varSigma_i} \cdot \pi$$

equivalently for any $g \in \mathcal{G}$, $\hat{\pi}_i(g) = e^{-\lambda \varSigma_i(g)} / (\sum_{g' \in \mathcal{G}} e^{-\lambda \varSigma_i(g')})$. This distribution concentrates on functions having low cumulative loss up to time i. For any $i \in \{0, \dots, n\}$, let $\hat{h}_i$ be a prediction function such that

$$\forall\, (x,y) \in \mathcal{Z} \qquad \ell[y, \hat{h}_i(x)] \leq -\tfrac{1}{\lambda} \log \mathbb{E}_{g \sim \hat{\pi}_i}\, e^{-\lambda \ell[y, g(x)]}. \tag{2}$$

The *progressive indirect mixture rule* produces the prediction function

$$\hat{g}_{\mathrm{pim}} = \tfrac{1}{n+1} \textstyle\sum_{i=0}^{n} \hat{h}_i.$$

From the uniform exp-concavity assumption and Jensen's inequality, $\hat{h}_i$ does exist since one may take $\hat{h}_i = \mathbb{E}_{g \sim \hat{\pi}_i}\, g$. This particular choice leads to the *progressive mixture rule*, for which the predicted output for any $x \in \mathcal{X}$ is

$$\hat{g}_{\mathrm{pm}}(x) = \textstyle\sum_{g \in \mathcal{G}} \left(\frac{1}{n+1} \sum_{i=0}^{n} \frac{e^{-\lambda \varSigma_i(g)}}{\sum_{g' \in \mathcal{G}} e^{-\lambda \varSigma_i(g')}} \right) g(x).$$

Consequently, any result that holds for any progressive indirect mixture rule in particular holds for the progressive mixture rule.

The idea of a progressive mean of estimators has been introduced by Barron ([3]) in the context of density estimation with Kullback-Leibler loss. The form $\hat{g}_{\mathrm{pm}}$ is due to Catoni ([7]). It was also independently proposed in [4]. The study of this procedure was made in density estimation and least square regression in [8,5,15,6]. Results for general losses can be found in [12,2]. Finally, the progressive indirect mixture rule is inspired by the work of Vovk, Haussler, Kivinen and Warmuth [13,11,14] on sequential prediction and was studied in the "batch" setting in [2].

The symbol C will denote some positive constant whose value may differ from line to line. The logarithm in base 2 is denoted by $\log_2$ (i.e. $\log_2 t = \log t / \log 2$) and $\lfloor x \rfloor$ denotes the largest integer k such that $k \leq x$.

2 Expectation convergence rate

First let us define the expectation convergence rate of a learning algorithm.

Definition 1. *For a given reference set $\mathcal{G}$ of prediction functions and a set $\mathcal{P}$ of probability distributions on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, a positive sequence $(\varDelta_n)_{n \geq 2}$ is said to be an* expectation convergence rate *of a learning algorithm producing the prediction function $\hat{g}$ iff there exist $C > c > 0$ such that*

1. *for any distribution $P \in \mathcal{P}$ and any $n \geq 2$, we have*

$$\mathbb{E} R(\hat{g}) - \inf_{g \in \mathcal{G}} R(g) \leq C \varDelta_n \tag{3}$$

2. for large enough n, there exists $P \in \mathcal{P}$ for which

$$\mathbb{E}R(\hat{g}) - \inf_{g\in\mathcal{G}} R(g) \geq c\Delta_n.$$

We say that the rate Δ_n is optimal *iff the previous item 2 is also satisfied for any other algorithm, in other words iff there is no algorithm having an expectation convergence rate $\tilde{\Delta}_n$ satisfying $\lim_{n\to+\infty} \tilde{\Delta}_n/\Delta_n = 0$.*

The following theorem shows that the expectation convergence rate of any progressive indirect mixture rule is at least $(\log|\mathcal{G}|)/n$ and that for any positive integer d, there exists a set $\mathcal{G}$ of d prediction functions such that this rate is optimal whether we take $\mathcal{P}$ as the set of all probability distributions on $\mathcal{Z}$ or the set of all probability distributions on $\mathcal{Z}$ for which the output has almost surely two symmetrical values (e.g. {-1;+1}-classication with exponential or logit losses).

Theorem 1. *Any progressive indirect mixture rule satisfies*

$$\mathbb{E}R(\hat{g}_{pim}) \leq \min_{g\in\mathcal{G}} R(g) + \tfrac{\log|\mathcal{G}|}{\lambda(n+1)}.$$

Let $y_1 \in \mathcal{Y} - \{a\}$ and d be a positive integer. There exists a set $\mathcal{G}$ of d prediction functions such that: for any learning algorithm, there exists a probability distribution generating the data for which

- *the output marginal is supported by $2a - y_1$ and y_1: $P(Y \in \{2a - y_1; y_1\}) = 1$,*
- *$\mathbb{E}R(\hat{g}) \geq \min_{g\in\mathcal{G}} R(g) + e^{-1}\kappa\big(1 \wedge \frac{\lfloor\log_2|\mathcal{G}|\rfloor}{n+1}\big)$, with $\kappa \triangleq \sup_{y\in\mathcal{Y}} [\ell(y_1, a) - \ell(y_1, y)] > 0$.*

Proof. See Appendix A.

The second part of Theorem 1 has the same $(\log|\mathcal{G}|/n)$-rate as the lower bounds obtained in sequential prediction ([11]). From the link between sequential predictions and our "batch" setting with i.i.d. data (see e.g. [2, Lemma 3]), upper bounds for sequential prediction lead to upper bounds for i.i.d. data, and lower bounds for i.i.d. data leads to lower bounds for sequential prediction. The converse of this last assertion is not true, so that the second part of Theorem 1 is not a consequence of the lower bounds of [11].

The following theorem shows that for appropriate set $\mathcal{G}$:

- the empirical risk minimizer has a $\sqrt{\log|\mathcal{G}|/n}$-expectation convergence rate.
- any empirical risk minimizer and any of its penalized variants are really poor algorithms in our learning task since their expectation convergence rate cannot be faster than $\sqrt{\log|\mathcal{G}|/n}$. This last point explains the interest we have in progressive mixture rules.

Theorem 2. *If $B \triangleq \sup_{y,y',y''\in\mathcal{Y}}[\ell(y,y') - \ell(y,y'')] < +\infty$, then any empirical risk minimizer, which produces a prediction function $\hat{g}_{erm}$ in argmin$_{g\in\mathcal{G}}\, \Sigma_n$, satisfies:*

$$\mathbb{E}R(\hat{g}_{erm}) \leq \min_{g\in\mathcal{G}} R(g) + B\sqrt{\tfrac{2\log|\mathcal{G}|}{n}}.$$

Let $y_1, \tilde{y_1} \in \mathcal{Y} \cap]a; +\infty[$ and d be a positive integer. There exists a set $\mathcal{G}$ of d prediction functions taking their values in $\{2a - \tilde{y_1}, \tilde{y_1}\}$ such that: for any learning algorithm producing a prediction function in $\mathcal{G}$ (e.g. $\hat{g}_{erm}$) there exists a probability distribution generating the data for which

- *the output marginal is supported by $2a - y_1$ and y_1: $P(Y \in \{2a - y_1; y_1\}) = 1$,*
- $\mathbb{E}R(\hat{g}) \geq \min_{g \in \mathcal{G}} R(g) + \frac{\delta}{8}\left(\sqrt{\frac{\lfloor \log_2 |\mathcal{G}| \rfloor}{n}} \wedge 2\right)$, *with* $\delta \triangleq \ell(y_1, 2a - \tilde{y_1}) - \ell(y_1, \tilde{y_1}) > 0$.

Proof. See Appendix B.

3 Deviation convergence rate

The efficiency of an algorithm $\hat{g}$ can be summarized by its expected risk $\mathbb{E}\, R(\hat{g})$, but this does not precise the fluctuations of $R(\hat{g})$. In several application fields of learning algorithms, these fluctuations play a key role: in finance for instance, the bigger the losses can be, the more money the bank needs to freeze in order to alleviate these possible losses. In this case, a “good” algorithm is an algorithm having not only low expected risk but also small deviations.

The deviation convergence rate we define now is concerned with exponential deviation inequalities (such as Hoeffding’s inequality or more generally such as standard statistical learning inequalities on the supremum of empirical processes).

Definition 2. *Let $0 < \gamma \leq 1$. For a given reference set $\mathcal{G}$ of prediction functions and a set $\mathcal{P}$ of probability distributions on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, a positive sequence $(\Delta'_n)_{n \in \mathbb{N}}$ is said to be a* deviation convergence rate *of order γ of a learning algorithm iff there exist $C > c > 0$ such that*

1. *for any distribution $P \in \mathcal{P}$, integer $n \geq 2$, and $\epsilon > 0$, with probability at least $1 - \epsilon$ w.r.t. the training set distribution, we have*

$$R(\hat{g}) - \inf_{g \in \mathcal{G}} R(g) \leq C\big[\log^{\gamma}(e\epsilon^{-1})\big]\Delta'_n, \qquad (4)$$

2. *for large enough n, there exist $\epsilon > 0$ and a distribution $P \in \mathcal{P}$ such that with probability at least ϵ w.r.t. the training set distribution, we have*

$$R(\hat{g}) - \inf_{g \in \mathcal{G}} R(g) \geq c\big[\log^{\gamma}(e\epsilon^{-1})\big]\Delta'_n.$$

The following lemma shows that the expectation convergence rate of a learning algorithm is at least of order of its deviation convergence rate. The expectation convergence rate can also be strictly faster as the comparison between Theorems 1 and 3 shows.

Lemma 1. *Let $\hat{g}$ satisfy: for any $\epsilon > 0$, with probability at least $1 - \epsilon$, (4) holds. Then we have*

$$\mathbb{E}R(\hat{g}) - \inf_{g \in \mathcal{G}} R(g) \leq 2^{\gamma} C \Delta'_n.$$

Proof. It suffices to integrate the deviations. Let $R^* = \inf_{g \in \mathcal{G}} R(g)$. By Jensen's inequality, we have

$$\begin{aligned}
&\Big(\tfrac{\mathbb{E}R(\hat{g})-R^*}{C\varDelta'_n}\Big)^{1/\gamma} - 1 \\
&\quad \leq \mathbb{E}\Big(\tfrac{R(\hat{g})-R^*}{C\varDelta'_n}\Big)^{1/\gamma} - 1 \\
&\quad \leq \mathbb{E}\Big\{\Big[\big(\tfrac{R(\hat{g})-R^*}{C\varDelta'_n}\big)^{1/\gamma} - 1\Big] \vee 0\Big\} \\
&\quad = \textstyle\int_0^{+\infty} \mathbb{P}\big\{\big(\tfrac{R(\hat{g})-R^*}{C\varDelta'_n}\big)^{1/\gamma} - 1 > u\big\} du \\
&\quad = \textstyle\int_0^{1} \mathbb{P}\big\{R(\hat{g}) - R^* > C\varDelta'_n \log^\gamma(e\epsilon^{-1})\big\} \tfrac{d\epsilon}{\epsilon} \qquad [\text{setting } u = \log(\epsilon^{-1})] \\
&\quad \leq 1.
\end{aligned}$$

The following theorem shows that the deviation convergence rate of order $1/2$ of any progressive indirect mixture rule is at least $1/\sqrt{n}$ and that there exists $\mathcal{G}$ such that the deviation convergence rate of order $1/2$ of any progressive indirect mixture rule is $1/\sqrt{n}$ whether we take $\mathcal{P}$ as the set of all probability distributions on $\mathcal{Z}$ or the set of all probability distributions on $\mathcal{Z}$ for which the output has almost surely two symmetrical values (e.g. {-1;+1}-classication with exponential or logit losses).

Theorem 3. *If $B \triangleq \sup_{y,y',y'' \in \mathcal{Y}}[\ell(y,y') - \ell(y,y'')] < +\infty$, then any progressive indirect mixture rule satisfies: for any $\epsilon > 0$, with probability at least $1 - \epsilon$ w.r.t. the training set distribution, we have*

$$R(\hat{g}_{pim}) \leq \min_{g \in \mathcal{G}} R(g) + B\sqrt{\tfrac{2\log(2\epsilon^{-1})}{n+1}} + \tfrac{\log|\mathcal{G}|}{\lambda(n+1)}$$

Let y_1 and $\tilde{y_1}$ in $\mathcal{Y} \cap]a; +\infty[$ such that ℓ_{y_1} is twice continuously differentiable on $[a; \tilde{y_1}]$ and $\ell'_{y_1}(\tilde{y_1}) \leq 0$ and $\ell''_{y_1}(\tilde{y_1}) > 0$. Consider the prediction functions $g_1 \equiv \tilde{y_1}$ and $g_2 \equiv 2a - \tilde{y_1}$. For any training set size n large enough, there exist $\epsilon > 0$ and a distribution generating the data such that

- *the output marginal is supported by y_1 and $2a - y_1$*
- *with probability larger than ϵ, we have*

$$R(\hat{g}_{pim}) - \min_{g \in \{g_1,g_2\}} R(g) \geq c\sqrt{\tfrac{\log(e\epsilon^{-1})}{n}}$$

where c is a positive constant depending only on the loss function, the symmetry parameter a and the output values y_1 and $\tilde{y_1}$.

Proof. See Section 4.

This result is quite surprising since it gives an example of an algorithm which is optimal in terms of expectation convergence rate and for which the deviation convergence rate is (significantly) worse that the expectation convergence rate.

4 Proof of Theorem 3

4.1 Proof of the upper bound

We would like to thank an anonymous reviewer for suggesting the following proof, which leads to better constants than the original one based on PAC-Bayesian inequalities.

Let $Z_{n+1} = (X_{n+1}, Y_{n+1})$ be an input-output pair independent from the training set $Z_1, \dots, Z_n$ and with the same distribution P. From the convexity of $y' \mapsto \ell(y, y')$, we have

$$R(\hat{g}_{\text{pim}}) \le \tfrac{1}{n+1} \textstyle\sum_{i=0}^n R(\hat{h}_i). \tag{5}$$

Now from [16, Theorem 1] (see also [9, Proposition 1]), for any $\epsilon > 0$, with probability at least $1 - \epsilon$, we have

$$\tfrac{1}{n+1} \textstyle\sum_{i=0}^n R(\hat{h}_i) \le \tfrac{1}{n+1} \textstyle\sum_{i=0}^n \ell\big(Y_{i+1}, \hat{h}(X_{i+1})\big) + B\sqrt{\tfrac{\log(\epsilon^{-1})}{2(n+1)}} \tag{6}$$

Using [11, Theorem 3.8] and the exp-concavity assumption, we have

$$\textstyle\sum_{i=0}^n \ell\big(Y_{i+1}, \hat{h}(X_{i+1})\big) \le \min\limits_{g \in \mathcal{G}} \textstyle\sum_{i=0}^n \ell\big(Y_{i+1}, g(X_{i+1})\big) + \tfrac{\log|\mathcal{G}|}{\lambda} \tag{7}$$

Let $\tilde{g} \in \operatorname{argmin}_{\mathcal{G}} R$. By Hoeffding's inequality, with probability at least $1 - \epsilon$, we have

$$\tfrac{1}{n+1} \textstyle\sum_{i=0}^n \ell\big(Y_{i+1}, \tilde{g}(X_{i+1})\big) \le R(\tilde{g}) + B\sqrt{\tfrac{\log(\epsilon^{-1})}{2(n+1)}} \tag{8}$$

Merging (5), (6), (7) and (8), with probability at least $1 - 2\epsilon$, we get

$$\begin{aligned} R(\hat{g}_{\text{pim}}) &\le \tfrac{1}{n+1} \textstyle\sum_{i=0}^n \ell\big(Y_{i+1}, \tilde{g}(X_{i+1})\big) + \tfrac{\log|\mathcal{G}|}{\lambda(n+1)} + B\sqrt{\tfrac{\log(\epsilon^{-1})}{2(n+1)}} \\ &\le R(\tilde{g}) + B\sqrt{\tfrac{2\log(\epsilon^{-1})}{n+1}} + \tfrac{\log|\mathcal{G}|}{\lambda(n+1)}. \end{aligned}$$

4.2 Proof of the lower bound

We cannot use standard tools like Assouad's argument (see e.g. [10, Theorem 14.6]) because if it were possible, it would mean that the lower bound would hold for any algorithm and this is (non trivially) false.

To prove that any progressive indirect mixture rule have no fast exponential deviation inequalities, we will show that on some event with not too small probability, for most of the i in $\{0, \dots, n\}$, $\pi_{-\lambda\Sigma_i}$ concentrates on the wrong function.

The proof is organized as follows. First we define the probability distribution for which we will prove that the progressive indirect mixture rules cannot have fast deviation convergence rates. Then we define the event on which the progressive indirect mixture rules do not perform well. We lower bound the probability of this excursion event. Finally we conclude by lower bounding $R(\hat{g}_{\text{pim}})$ on the excursion event.

Before starting the proof, note that from the "well behaved at center" and exp-concavity assumptions, for any $y \in \mathcal{Y} \cap]a; +\infty[$, on a neighborhood of a, we have: $\ell_y'' \ge \lambda(\ell_y')^2$ and since $\ell_y'(a) < 0$, y_1 and $\tilde{y}_1$ exist.

Probability distribution generating the data and first consequences. Let $\gamma \in]0;1]$ be a parameter to be tuned later. We consider a distribution generating the data such that the output distribution satisfies for any $x \in \mathcal{X}$

$$P(Y = y_1 | X = x) = (1+\gamma)/2 = 1 - P(Y = y_2 | X = x),$$

where $y_2 = 2a - y_1$. Let $\tilde{y_2} = 2a - \tilde{y_1}$. From the symmetry and admissibility assumptions, we have $\ell(y_2, \tilde{y_2}) = \ell(y_1, \tilde{y_1}) < \ell(y_1, \tilde{y_2}) = \ell(y_2, \tilde{y_1})$. Introduce

$$\delta \triangleq \ell(y_1, \tilde{y_2}) - \ell(y_1, \tilde{y_1}) > 0. \tag{9}$$

We have

$$R(g_2) - R(g_1) = \tfrac{1+\gamma}{2}[\ell(y_1, \tilde{y_2}) - \ell(y_1, \tilde{y_1})] + \tfrac{1-\gamma}{2}[\ell(y_2, \tilde{y_2}) - \ell(y_2, \tilde{y_1})] = \gamma\delta. \tag{10}$$

Therefore g_1 is the best prediction function in $\{g_1, g_2\}$ for the distribution we have chosen. Introduce $W_j \triangleq \mathbf{1}_{Y_j = y_1} - \mathbf{1}_{Y_j = y_2}$ and $S_i \triangleq \sum_{j=1}^{i} W_j$. For any $i \in \{1, \dots, n\}$, we have

$$\varSigma_i(g_2) - \varSigma_i(g_1) = \textstyle\sum_{j=1}^{i} [\ell(Y_j, \tilde{y_2}) - \ell(Y_j, \tilde{y_1})] = \sum_{j=1}^{i} W_j \delta = \delta\, S_i$$

The weight given by the Gibbs distribution $\pi_{-\lambda\varSigma_i}$ to the function g_1 is

$$\pi_{-\lambda\varSigma_i}(g_1) = \tfrac{e^{-\lambda\varSigma_i(g_1)}}{e^{-\lambda\varSigma_i(g_1)} + e^{-\lambda\varSigma_i(g_2)}} = \tfrac{1}{1 + e^{\lambda[\varSigma_i(g_1) - \varSigma_i(g_2)]}} = \tfrac{1}{1 + e^{-\lambda\delta S_i}}. \tag{11}$$

An excursion event on which the progressive indirect mixture rules will not perform well. (11) leads us to consider the event:

$$E_\tau = \big\{\forall i \in \{\tau, \dots, n\},\ S_i \le -\tau\big\},$$

with τ the smallest integer larger than $(\log n)/(\lambda\delta)$ such that $n - \tau$ is even. (We could have just as well chosen $n - \tau$ odd; see (17) below.) We have

$$\tfrac{\log n}{\lambda\delta} \le \tau \le \tfrac{\log n}{\lambda\delta} + 2. \tag{12}$$

The event E_τ can be seen as an excursion event of the random walk defined through the random variables $W_j = \mathbf{1}_{Y_j = y_1} - \mathbf{1}_{Y_j = y_2}$, $j \in \{1, \dots, n\}$, which are equal to $+1$ with probability $(1+\gamma)/2$ and -1 with probability $(1-\gamma)/2$.

From (11), on the event E_τ, for any $i \in \{\tau, \dots, n\}$, we have

$$\pi_{-\lambda\varSigma_i}(g_1) \le \tfrac{1}{n+1}. \tag{13}$$

This means that $\pi_{-\lambda\varSigma_i}$ concentrates on the wrong function, i.e. the function g_2 having larger risk (see (10)).

Lower bound of the probability of the excursion event. This requires to look at the probability that a slightly shifted random walk in the integer space has a very long excursion above a certain threshold. To lower bound this probability, we will first look at the non-shifted random walk. Then we will see that for small enough shift parameter, probabilities of shifted random walk events are close to the ones associated to the non-shifted random walk.

Let N be a positive integer. Let $\sigma_1, \dots, \sigma_N$ be N independent Rademacher variables: $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = 1/2$. Let $s_i \triangleq \sum_{j=1}^{i} \sigma_i$ be the sum of the first i Rademacher variables. We start with the following lemma for sums of Rademacher variables.

Lemma 2. *Let m and t be positive integers. We have*

$$\mathbb{P}\big(\max_{1\le k\le N} s_k \ge t; s_N \neq t; \big|s_N - t\big| \le m\big) = 2\mathbb{P}\big(t < s_N \le t + m\big) \tag{14}$$

Proof (of Lemma 2). The result comes from the well known mirror trick used to compute the law of $\big(\sup_{s\le t} W_s, W_t\big)$ where W denotes a Brownian motion. Consider a sequence $\sigma_1, \dots, \sigma_N$ which belongs to the event $\mathcal{E}$ of the l.h.s. probability. Let J be the first integer j such that $s_j = t$. Since

- the sequences $\sigma_1, \dots, \sigma_N$ and $\sigma_1, \dots, \sigma_J, -\sigma_{J+1}, \dots, -\sigma_N$ have the same probabilities,
- both sequences belong to $\mathcal{E}$ and are different since $J < N$,
- exactly one of the sequences satisfy $s_N > t$,

we have

$$\mathbb{P}\big(\max_{1\le k\le N} s_k \ge t; s_N \neq t; \big|s_N - t\big| \le m\big) = 2\mathbb{P}\big(s_N > t; \big|s_N - t\big| \le m\big),$$

which is the desired result.

Let $\sigma'_1, \dots, \sigma'_N$ be N independent shifted Rademacher variables to the extent that $\mathbb{P}(\sigma'_i = +1) = (1+\gamma)/2 = 1 - \mathbb{P}(\sigma'_i = -1)$. These random variables satisfy the following key lemma

Lemma 3. *For any set $A \subset \big\{(\epsilon_1, \dots, \epsilon_N) \in \{-1, 1\}^n : \big|\sum_{i=1}^N \epsilon_i\big| \le M\big\}$ where M is a positive integer, we have*

$$\mathbb{P}\big\{(\sigma'_1, \dots, \sigma'_N) \in A\big\} \ge \Big(\tfrac{1-\gamma}{1+\gamma}\Big)^{M/2} \big(1-\gamma^2\big)^{N/2} \mathbb{P}\big\{(\sigma_1, \dots, \sigma_N) \in A\big\} \tag{15}$$

Proof (of Lemma 3). Let s be an integer such that $N - s$ is even and $|s| \le M$ Consider a sequence $\epsilon_1, \dots, \epsilon_N$ such that $\sum_{i=1}^N \epsilon_i = s$. Then the numbers of -1 and $+1$ in the sequence are respectively $(N-s)/2$ and $(N+s)/2$. Consequently, we have

$$\tfrac{\mathbb{P}[(\sigma'_1,\dots,\sigma'_N)=(\epsilon_1,\dots,\epsilon_N)]}{\mathbb{P}[(\sigma_1,\dots,\sigma_N)=(\epsilon_1,\dots,\epsilon_N)]} = (1+\gamma)^{(N-s)/2}(1-\gamma)^{(N+s)/2},$$

hence

$$\begin{aligned}\mathbb{P}\{(\sigma'_1, \dots, \sigma'_N) = (\epsilon_1, \dots, \epsilon_N)\}&\\ &\ge (1-\gamma^2)^{N/2}\big(\tfrac{1-\gamma}{1+\gamma}\big)^{M/2}\mathbb{P}\big\{(\sigma_1, \dots, \sigma_N) = (\epsilon_1, \dots, \epsilon_N)\big\}.\end{aligned}$$

By summing over the sequences $\epsilon_1, \dots, \epsilon_N$ in A, we obtain the desired result.

We may now lower bound the probability of the excursion event E_τ. Let M be an integer larger than τ. We still use $W_j \triangleq \mathbf{1}_{Y_j=y_1} - \mathbf{1}_{Y_j=y_2}$ for $j \in \{1, \ldots, n\}$. By using Lemma 3 with $N = n - 2\tau$, we obtain

$$\begin{aligned}
\mathbb{P}(E_\tau) &\geq \mathbb{P}\bigl(W_1 = -1, \ldots, W_{2\tau} = -1;\ \forall\, 2\tau < i \leq n,\ \textstyle\sum_{j=2\tau+1}^{i} W_j \leq \tau\bigr)\\
&= \bigl(\tfrac{1-\gamma}{2}\bigr)^{2\tau}\mathbb{P}\bigl(\forall\, i > 2\tau \quad \textstyle\sum_{j=2\tau+1}^{i} W_j \leq \tau\bigr)\\
&= \bigl(\tfrac{1-\gamma}{2}\bigr)^{2\tau}\mathbb{P}\bigl(\forall\, i \in \{1, \ldots, N\} \quad \textstyle\sum_{j=1}^{i} \sigma'_j \leq \tau\bigr)\\
&\geq \bigl(\tfrac{1-\gamma}{2}\bigr)^{2\tau}\mathbb{P}\bigl(\bigl|\textstyle\sum_{i=1}^{N} \sigma'_i\bigr| < M; \forall\, i \in \{1, \ldots, N\} \quad \textstyle\sum_{j=1}^{i} \sigma'_j \leq \tau\bigr)\\
&\geq \bigl(\tfrac{1-\gamma}{2}\bigr)^{2\tau}\bigl(\tfrac{1-\gamma}{1+\gamma}\bigr)^{M/2}\bigl(1-\gamma^2\bigr)^{\frac{N}{2}}\mathbb{P}\bigl(|s_N| \leq M; \forall\, i \in \{1, \ldots, N\} \quad s_i \leq \tau\bigr)
\end{aligned} \tag{16}$$

By using Lemma 2, since $\tau \leq M$, the r.h.s. probability can be lower bounded:

$$\begin{aligned}
&\mathbb{P}\bigl(|s_N| \leq M; \max_{1\leq i\leq N} s_i \leq \tau\bigr)\\
&\qquad = \mathbb{P}\Bigl\{\max_{1\leq i\leq N} s_i \leq \tau; s_N \geq -M\Bigr\}\\
&\qquad \geq \mathbb{P}\Bigl\{\max_{1\leq i\leq N} s_i < \tau; |s_N - \tau| \leq M + \tau; s_N \neq \tau\Bigr\}\\
&\qquad = \mathbb{P}\Bigl\{|s_N - \tau| \leq M + \tau; s_N \neq \tau\Bigr\}\\
&\qquad\qquad - \mathbb{P}\Bigl\{\max_{1\leq i\leq N} s_i \geq \tau; |s_N - \tau| \leq M + \tau; s_N \neq \tau\Bigr\}\\
&\qquad = \mathbb{P}\bigl\{|s_N - \tau| \leq M + \tau; s_N \neq \tau\bigr\} - 2\mathbb{P}\bigl\{\tau < s_N \leq M + 2\tau\bigr\}\\
&\qquad = \mathbb{P}\bigl\{-M \leq s_N < \tau\bigr\} - \mathbb{P}\bigl\{\tau < s_N \leq M + 2\tau\bigr\}\\
&\qquad = \mathbb{P}\bigl\{-\tau < s_N \leq M\bigr\} - \mathbb{P}\bigl\{\tau < s_N \leq M + 2\tau\bigr\}\\
&\qquad = \mathbb{P}\bigl\{-\tau < s_N \leq \tau\bigr\} - \mathbb{P}\bigl\{M < s_N \leq M + 2\tau\bigr\}
\end{aligned}$$

Let us consider only the integer $M > \tau$ such that $n - M$ is even, or equivalently $N - M$ is even. Since $N - \tau = n - 3\tau$ is also even, we have

$$\begin{aligned}
&\mathbb{P}\bigl(|s_N| \leq M; \max_{1\leq i\leq N} s_i \leq \tau\bigr)\\
&\qquad\qquad \geq \textstyle\sum_{k=0}^{\tau-1}\mathbb{P}(s_N = 2 - \tau + 2k) - \textstyle\sum_{k=1}^{\tau}\mathbb{P}(s_N = M + 2k)\\
&\qquad\qquad \geq \tau[\mathbb{P}(s_N = \tau) - \mathbb{P}(s_N = M)],
\end{aligned} \tag{17}$$

where the last inequality comes from properties of the binomial coefficients.

Combining (16) and (17), we obtain

$$\mathbb{P}(E_\tau) \geq \tau\bigl(\tfrac{1-\gamma}{2}\bigr)^{2\tau}\bigl(\tfrac{1-\gamma}{1+\gamma}\bigr)^{M/2}\bigl(1-\gamma^2\bigr)^{\frac{N}{2}}[\mathbb{P}(s_N = \tau) - \mathbb{P}(s_N = M)] \tag{18}$$

where we recall that τ have the order of $\log n$, $N = n - 2\tau$ has the order of n and that $\gamma > 0$ and $M \geq \tau$ have to be appropriately chosen.

To control the probabilities of the r.h.s., we use Stirling's formula

$$n^n e^{-n}\sqrt{2\pi n}\, e^{1/(12n+1)} < n! < n^n e^{-n}\sqrt{2\pi n}\, e^{1/(12n)}, \tag{19}$$

and get for any $s \in [0; N]$ such that $N - s$ even,

$$\begin{aligned}
\mathbb{P}(s_N = s) &= \big(\tfrac{1}{2}\big)^N \big(\tfrac{N}{\frac{N+s}{2}}\big) \\
&\geq \big(\tfrac{1}{2}\big)^N \frac{(\frac{N}{e})^N \sqrt{2\pi N} e^{\frac{1}{12N+1}}}{(\frac{N+s}{2e})^{\frac{N+s}{2}} (\frac{N-s}{2e})^{\frac{N-s}{2}} \sqrt{\pi(N+s)} \sqrt{\pi(N-s)} e^{\frac{1}{6(N+s)}} e^{\frac{1}{6(N-s)}}} \\
&= \frac{1}{(1+\frac{s}{N})^{\frac{N+s}{2}} (1-\frac{s}{N})^{\frac{N-s}{2}}} \sqrt{\frac{2N}{\pi(N^2-s^2)}} e^{\frac{1}{12N+1} - \frac{1}{6(N+s)} - \frac{1}{6(N-s)}} \\
&\geq \sqrt{\frac{2}{\pi N}} \Big(1 - \frac{s^2}{N^2}\Big)^{-\frac{N}{2}} \Big(\frac{1-\frac{s}{N}}{1+\frac{s}{N}}\Big)^{\frac{s}{2}} e^{-\frac{1}{6(N+s)} - \frac{1}{6(N-s)}}
\end{aligned} \tag{20}$$

and similarly

$$\mathbb{P}(s_N = s) \leq \sqrt{\frac{2}{\pi N}} \Big(1 - \frac{s^2}{N^2}\Big)^{-\frac{N}{2}} \Big(\frac{1-\frac{s}{N}}{1+\frac{s}{N}}\Big)^{\frac{s}{2}} e^{\frac{1}{12N+1}} \tag{21}$$

These computations and (18) leads us to take M as the smallest integer larger than $\sqrt{n}$ such that $n - M$ is even. Indeed, from (12), (20) and (21), we obtain $\lim_{n \to +\infty} \sqrt{n}[\mathbb{P}(s_N = \tau) - \mathbb{P}(s_N = M)] = c$, where $c = \sqrt{2/\pi}\big(1 - e^{-1/2}\big) > 0$. Therefore for n large enough we have

$$\mathbb{P}(E_\tau) \geq \tfrac{c\tau}{2\sqrt{n}} \big(\tfrac{1-\gamma}{2}\big)^{2\tau} \big(\tfrac{1-\gamma}{1+\gamma}\big)^{M/2} \big(1 - \gamma^2\big)^{\frac{N}{2}} \tag{22}$$

The last two terms of the r.h.s. of (22) leads us to take γ of order $1/\sqrt{n}$ up to possibly a logarithmic term. We obtain the following lower bound on the excursion probability

Lemma 4. *If $\gamma = \sqrt{C_0 (\log n)/n}$ with C_0 a positive constant, then for any large enough n,*

$$\mathbb{P}(E_\tau) \geq \tfrac{1}{n^{C_0}}.$$

Behavior of the progressive indirect mixture rule on the excursion event. From now on, we work on the event E_τ. We have $\hat{g}_{\text{pim}} = (\sum_{i=0}^n \hat{h}_i)/(n+1)$. We still use $\delta \triangleq \ell(y_1, \tilde{y_2}) - \ell(y_1, \tilde{y_1}) = \ell(y_2, \tilde{y_1}) - \ell(y_2, \tilde{y_2})$. On the event E_τ, for any $x \in \mathcal{X}$ and any $i \in \{\tau, \dots, n\}$, by definition of $\hat{h}_i$, we have

$$\begin{aligned}
\ell[y_2, \hat{h}_i(x)] - \ell(y_2, \tilde{y_2}) &\leq -\tfrac{1}{\lambda} \log \mathbb{E}_{g \sim \pi_{-\lambda \Sigma_i}} \, e^{-\lambda\{\ell[y_2, g(x)] - \ell(y_2, \tilde{y_2})\}} \\
&= -\tfrac{1}{\lambda} \log \big\{\pi_{-\lambda\Sigma_i}(g_1) e^{-\lambda\delta} + \pi_{-\lambda\Sigma_i}(g_2)\big\} \\
&= -\tfrac{1}{\lambda} \log \big\{e^{-\lambda\delta} + (1 - e^{-\lambda\delta}) \pi_{-\lambda\Sigma_i}(g_2)\big\} \\
&\leq -\tfrac{1}{\lambda} \log \big\{1 - (1 - e^{-\lambda\delta}) \tfrac{1}{n+1}\big\}
\end{aligned}$$

In particular, for any n large enough, we have $\ell[y_2, \hat{h}_i(x)] - \ell(y_2, \tilde{y_2}) \leq C n^{-1}$, with $C > 0$ *independent from* γ. From the convexity of the function $y \mapsto \ell(y_2, y)$ and by Jensen's inequality, we obtain

$$\begin{aligned}
\ell[y_2, \hat{g}_{\text{pim}}(x)] - \ell(y_2, \tilde{y_2}) &= \ell[y_2, \tfrac{1}{n+1} \textstyle\sum_{i=0}^n \hat{h}_i(x)] - \ell(y_2, \tilde{y_2}) \\
&\leq \tfrac{1}{n+1} \textstyle\sum_{i=0}^n \ell[y_2, \hat{h}_i(x)] - \ell(y_2, \tilde{y_2}) \\
&\leq \tfrac{\tau\delta}{n+1} + C n^{-1} \\
&< C_1 \tfrac{\log n}{n}
\end{aligned} \tag{23}$$

for some constant $C_1 > 0$ *independent from* γ. Let us now prove that for n large enough, we have

$$\tilde{y_2} \le \hat{g}_{\mathrm{pim}}(x) \le \tilde{y_2} + C\sqrt{\tfrac{\log n}{n}} \le \tilde{y_1}, \tag{24}$$

with $C > 0$ *independent from* γ.

Proof. For any $y \in \mathcal{Y}$, let $t = 2a - y$. We have $\ell(y_2, y) - \ell(y_2, \tilde{y_2}) = \ell_{y_1}(t) - \ell_{y_1}(\tilde{y_1})$. Since $\ell'_{y_1}(\tilde{y_1}) \le 0$, $\ell''_{y_1}(\tilde{y_1}) > 0$, $\ell''_{y_1} \ge \lambda(\ell'_{y_1})^2$ and ℓ''_{y_1} is continuous on $[a; \tilde{y_1}]$, there exists $m > 0$ such that $\ell''_{y_1} > m$ on $[a; \tilde{y_1}]$. For any $\tilde{y_2} < y \le a$, from Taylor's expansion, we have

$$\begin{aligned}\ell(y_2, y) - \ell(y_2, \tilde{y_2}) &> (t - \tilde{y_1})\ell'_{y_1}(\tilde{y_1}) + \tfrac{(t-\tilde{y_1})^2}{2} m \\ &\ge \tfrac{(t-\tilde{y_1})^2}{2} m \\ &= \tfrac{(y-\tilde{y_2})^2}{2} m\end{aligned} \tag{25}$$

Let $y_0 \triangleq \tilde{y_2} + \sqrt{\frac{2C_1 \log n}{mn}}$ where C_1 is the constant appearing in (23). For n large enough, we have $y_0 \le a$ and we may apply (25) to $y = y_0$. We get

$$\ell(y_2, y_0) - \ell(y_2, \tilde{y_2}) > C_1 \tfrac{\log n}{n}. \tag{26}$$

Since ℓ_{y_1} is convex, $\ell'_{y_1}(\tilde{y_1}) \le 0$ and $\ell''_{y_1}(\tilde{y_1}) > 0$, the function ℓ_{y_1} decreases on $] - \infty; \tilde{y_1}] \cap \mathcal{Y}$. By symmetry, the function $y \mapsto \ell(y_2, y)$ is non-decreasing on $[\tilde{y_2}; +\infty[\cap \mathcal{Y}$. From (23) and (26), we get $\hat{g}_{\mathrm{pim}}(x) \notin [y_0; +\infty[$, which ends the proof of the upper bound of $\hat{g}_{\mathrm{pim}}(x)$.

For the lower bound, for any $x \in \mathcal{X}$, by definition of $\hat{h}_i$, we have

$$\begin{aligned}\ell[y_1, \hat{h}_i(x)] - \ell(y_1, \tilde{y_1}) &\le -\tfrac{1}{\lambda} \log \mathbb{E}_{g \sim \pi_{-\lambda \Sigma_i}}\, e^{-\lambda\{\ell[y_1, g(x)] - \ell(y_1, \tilde{y_1})\}} \\ &= -\tfrac{1}{\lambda} \log \left\{\pi_{-\lambda\Sigma_i}(g_1) + \pi_{-\lambda\Sigma_i}(g_2) e^{-\lambda\delta}\right\} \\ &\le \delta.\end{aligned}$$

By Jensen's inequality, we obtain

$$\begin{aligned}\ell_{y_1}[\hat{g}_{\mathrm{pim}}(x)] - \ell_{y_1}(\tilde{y_1}) &= \ell[y_1, \tfrac{1}{n+1} \textstyle\sum_{i=0}^n \hat{h}_i(x)] - \ell(y_1, \tilde{y_1}) \\ &\le \tfrac{1}{n+1} \textstyle\sum_{i=0}^n \ell[y_1, \hat{h}_i(x)] - \ell(y_1, \tilde{y_1}) \\ &\le \delta \\ &= \ell_{y_1}(\tilde{y_2}) - \ell_{y_1}(\tilde{y_1}).\end{aligned}$$

Since the function ℓ_{y_1} decreases on $] - \infty; \tilde{y_2}] \cap \mathcal{Y}$, we get that $\hat{g}_{\mathrm{pim}}(x) \ge \tilde{y_2}$, which ends the proof of (24).

From (24), we obtain

$$\begin{aligned}R(\hat{g}_{\mathrm{pim}}) - R(g_1) &= \tfrac{1+\gamma}{2}\left[\ell(y_1, \hat{g}_{\mathrm{pim}}) - \ell(y_1, \tilde{y_1})\right] + \tfrac{1-\gamma}{2}\left[\ell(y_2, \hat{g}_{\mathrm{pim}}) - \ell(y_2, \tilde{y_1})\right] \\ &= \tfrac{1+\gamma}{2}\left[\ell_{y_1}(\hat{g}_{\mathrm{pim}}) - \ell_{y_1}(\tilde{y_1})\right] + \tfrac{1-\gamma}{2}\left[\ell_{y_1}(2a - \hat{g}_{\mathrm{pim}}) - \ell_{y_1}(\tilde{y_2})\right] \\ &= \tfrac{1+\gamma}{2}\left[\delta + \ell_{y_1}(\hat{g}_{\mathrm{pim}}) - \ell_{y_1}(\tilde{y_2})\right] \\ &\qquad + \tfrac{1-\gamma}{2}\left[-\delta + \ell_{y_1}(2a - \hat{g}_{\mathrm{pim}}) - \ell_{y_1}(\tilde{y_1})\right] \\ &\ge \gamma\delta - (\hat{g}_{\mathrm{pim}} - \tilde{y_2})|\ell'_{y_1}(\tilde{y_2})| \\ &\ge \gamma\delta - C_2\sqrt{\tfrac{\log n}{n}},\end{aligned} \tag{27}$$

with C_2 *independent from* γ. We may take $\gamma = \frac{2C_2}{\delta}\sqrt{(\log n)/n}$ and obtain: for n large enough, on the event E_τ, we have $R(\hat{g}_{\mathrm{pim}}) - R(g_1) \geq C\sqrt{\log n/n}$. From Lemma 4, this inequality holds with probability at least $1/n^{C_4}$ for some $C_4 > 0$. To conclude, for any n large enough, there exists $\epsilon > 0$ s.t. with probability at least ϵ,

$$R(\hat{g}_{\mathrm{pim}}) - R(g_1) \geq c\sqrt{\tfrac{\log(e\epsilon^{-1})}{n}}.$$

where c is a positive constant depending only on the loss function, the symmetry parameter a and the output values y_1 and $\tilde{y_1}$.

Remark 1. Had we consider the progressive mixture rule, this last part of the proof would have been much simpler. Indeed, for n large enough, on the event E_τ, from (13), we have

$$p \triangleq \tfrac{1}{n+1}\textstyle\sum_{i=0}^{n} \pi_{-\lambda\Sigma_i}(g_1) \leq \tfrac{\tau}{n+1} + \sup_{\tau \leq i \leq n} \pi_{-\lambda\Sigma_i}(g_1) \leq C\tfrac{\log n}{n}$$

and $\hat{g}_{\mathrm{pm}} = \frac{1}{n+1}\sum_{i=0}^{n} \mathbb{E}_{g\sim\pi_{-\lambda\Sigma_i}}\, g = pg_1 + (1-p)g_2 \equiv \tilde{y_2} + p(\tilde{y_1} - \tilde{y_2})$. So we have

$$\tilde{y_2} \leq \hat{g}_{\mathrm{pm}} \leq \tilde{y_2} + C\tfrac{\log n}{n} \leq \tilde{y_1},$$

which is much stronger than (24) (and much simpler to prove).

A Proof of Theorem 1

The first assertion is a direct consequence of Lemma 3.3 and Corollary 4.1 of [2]. The second assertion is based on an Assouad's type lower bound ([1, Inequality (8.19)]. Let $y_2 = 2a - y_1$ and $\tilde{m} = \lfloor \log_2 |\mathcal{G}| \rfloor$. We use the notation introduced in [1, Section 8.1]. We consider a $\big(\tilde{m}, \frac{1}{n+1} \wedge \frac{1}{\tilde{m}}, 1\big)$-hypercube of probability distributions with $h_1 \equiv \mathrm{argmin}_{y\in\mathcal{Y}} \ell_{y_1}(y)$ and $h_2 \equiv \mathrm{argmin}_{y\in\mathcal{Y}} \ell_{y_2}(y)$. We obtain

$$\begin{aligned}\mathbb{E}R(\hat{g}) - \min_{g\in\mathcal{G}} R(g) &\geq \big(\tfrac{\lfloor \log_2 |\mathcal{G}| \rfloor}{n+1} \wedge 1\big) d_{\mathrm{I}} \big(1 - \tfrac{1}{n+1} \wedge \tfrac{1}{\lfloor \log_2 |\mathcal{G}| \rfloor}\big)^n \\ &\geq \big(\tfrac{\lfloor \log_2 |\mathcal{G}| \rfloor}{n+1} \wedge 1\big) d_{\mathrm{I}} e^{-1},\end{aligned}$$

where the last inequality comes from $[1 - 1/(n+1)]^n \searrow e^{-1}$. Now the edge discrepancy d_{I} can be computed:

$$\begin{aligned}d_{\mathrm{I}} &= \psi_{1,0,y_1,y_2}(1/2) \\ &= \inf_{y\in\mathcal{Y}} \tfrac{\ell(y_1,y)+\ell(y_2,y)}{2} - \tfrac{1}{2}\inf_{y\in\mathcal{Y}} \ell(y_1,y) - \tfrac{1}{2}\inf_{y\in\mathcal{Y}} \ell(y_2,y) \\ &= \inf_{y\in\mathcal{Y}} \tfrac{\ell(y_1,y)+\ell(y_1,2a-y)}{2} - \inf_{y\in\mathcal{Y}} \ell(y_1,y) \\ &= \sup_{y\in\mathcal{Y}} [\ell(y_1,a) - \ell(y_1,y)],\end{aligned}$$

where the last equality uses that the function $y \mapsto \frac{\ell(y_1,y)+\ell(y_1,2a-y)}{2}$ is convex. Finally, from the “well behaved at center” assumption, the supremum is positive.

B Proof of Theorem 2

Let $\tilde{g} \in \operatorname{argmin}_{\mathcal{G}} R$ and $\eta > 0$. Hoeffding's inequality applied to the random variable $W = \ell[Y, \tilde{g}(X)] - \ell[Y, g(X)] \in [-B; B]$ for a fixed $g \in \mathcal{G}$ gives

$$\mathbb{E} e^{\eta[W - \mathbb{E}W]} \leq e^{\eta^2 B^2/2}$$

for any $\eta > 0$. Since the random variable $Z_1, \ldots, Z_n$ are independent, we obtain

$$\mathbb{E} e^{\eta[nR(g) - nR(\tilde{g}) + \Sigma_n(\tilde{g}) - \Sigma_n(g)]} \leq e^{\eta^2 n B^2/2}.$$

Consequently we have

$$\begin{aligned} n\big\{\mathbb{E}R(\hat{g}_{\mathrm{erm}}) - R(\tilde{g})\big\} &\leq \mathbb{E}\big\{nR(\hat{g}_{\mathrm{erm}}) - nR(\tilde{g}) + \Sigma_n(\tilde{g}) - \Sigma_n(\hat{g}_{\mathrm{erm}})\big\} \\ &\leq \tfrac{1}{\eta} \log \mathbb{E} e^{\eta[nR(\hat{g}_{\mathrm{erm}}) - nR(\tilde{g}) + \Sigma_n(\tilde{g}) - \Sigma_n(\hat{g}_{\mathrm{erm}})]} \\ &\leq \tfrac{1}{\eta} \log \mathbb{E} \sum_{g \in \mathcal{G}} e^{\eta[nR(g) - nR(\tilde{g}) + \Sigma_n(\tilde{g}) - \Sigma_n(g)]} \\ &\leq \tfrac{1}{\eta} \log \big(|\mathcal{G}| e^{\eta^2 n B^2/2}\big). \end{aligned}$$

The first assertion follows from the (optimal) choice $\eta = \sqrt{(2 \log |\mathcal{G}|)/(nB^2)}$.

The second assertion is based on an Assouad's type lower bound. Let $y_2 = 2a - y_1$ and $\tilde{m} = \lfloor \log_2 |\mathcal{G}| \rfloor$. We use the notation introduced in [1, Section 8.1]. We consider a $\big(\tilde{m}, \frac{1}{\tilde{m}}, \tilde{d_{\mathrm{II}}}\big)$-hypercube of probability distributions with $h_1 \equiv \tilde{y_1}$ and $h_2 \equiv \tilde{y_2} \triangleq 2a - \tilde{y_1}$ and $\tilde{d_{\mathrm{II}}}$ has to be optimized in $[0; 1]$. In the proof of Theorem 1, we take the set $\mathcal{G}$ such that $\min_{g \in \mathcal{G}} R(g) = \min_g R(g)$, where the second minimum is w.r.t. all possible prediction functions. Here the trick is to realize that $\min_{g \in \mathcal{G}} R(g)$ for our learning setting equals to $\min_g R(g)$ for the learning task in which the output space is only $\{\tilde{y_1}, \tilde{y_2}\}$. Therefore we apply ([1, Inequality (8.17)] with the function ϕ appearing in the edge discrepancy d_{I} defined as $\phi_{y_1,y_2}(p) = \min\limits_{y \in \{\tilde{y_1}, \tilde{y_2}\}} \big\{p\ell(y_1, y) + (1-p)\ell(y_2, y)\big\}$. We get

$$\begin{aligned} \mathbb{E}R(\hat{g}) &\geq \min_{g \in \mathcal{G}} R(g) + m w d_{\mathrm{I}} \big(1 - \sqrt{n w d_{\mathrm{II}}}\big) \\ &= \min_{g \in \mathcal{G}} R(g) + d_{\mathrm{I}} \Big(1 - \sqrt{\tfrac{n}{\tilde{m}} \tilde{d_{\mathrm{II}}}}\Big). \end{aligned}$$

From the symmetry and admissibility assumptions of the loss function, we have $\ell(y_2, \tilde{y_2}) = \ell(y_1, \tilde{y_1}) > \ell(y_2, \tilde{y_1}) = \ell(y_1, \tilde{y_2})$, hence $\delta \triangleq \ell(y_1, \tilde{y_2}) - \ell(y_1, \tilde{y_1}) > 0$. We obtain

$$\begin{aligned} d_{\mathrm{I}} &= \psi_{\frac{1+\sqrt{\tilde{d_{\mathrm{II}}}}}{2}, \frac{1-\sqrt{\tilde{d_{\mathrm{II}}}}}{2}, y_1, y_2}(1/2) \\ &= \phi_{y_1,y_2}(1/2) - \tfrac{1}{2}\phi_{y_1,y_2}\Big(\tfrac{1+\sqrt{\tilde{d_{\mathrm{II}}}}}{2}\Big) - \tfrac{1}{2}\phi_{y_1,y_2}\Big(\tfrac{1-\sqrt{\tilde{d_{\mathrm{II}}}}}{2}\Big) \\ &= \phi_{y_1,y_2}(1/2) - \phi_{y_1,y_2}\Big(\tfrac{1+\sqrt{\tilde{d_{\mathrm{II}}}}}{2}\Big) \\ &= \tfrac{1}{2}\ell(y_1, \tilde{y_1}) + \tfrac{1}{2}\ell(y_2, \tilde{y_1}) - \Big(\tfrac{1+\sqrt{\tilde{d_{\mathrm{II}}}}}{2}\ell(y_1, \tilde{y_1}) + \tfrac{1-\sqrt{\tilde{d_{\mathrm{II}}}}}{2}\ell(y_2, \tilde{y_1})\Big) \\ &= \tfrac{\sqrt{\tilde{d_{\mathrm{II}}}}}{2}\delta. \end{aligned}$$

The optimization of the lower bound leads us to choose $\tilde{d_{\mathrm{II}}} = \frac{\tilde{m}}{4n} \wedge 1$ and we get the desired result.

References

1. J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. Research report 06-20, Certis - Ecole des Ponts, http://cermics.enpc.fr/~audibert/RR0620d.pdf, 2006.
2. J.-Y. Audibert. A randomized online learning algorithm for better variance control. In *Proceedings of the 19th annual conference on Computational Learning Theory (COLT), Lecture Notes in Computer Science*, volume 4005, pages 392–407, 2006.
3. 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.
4. A. Barron and Y. Yang. Information-theoretic determination of minimax rates of convergence. *Ann. Stat.*, 27(5):1564–1599, 1999.
5. G. Blanchard. The progressive mixture estimator for regression trees. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 35(6):793–820, 1999.
6. F. Bunea and A. Nobel. Sequential procedures for aggregating arbitrary estimators of a conditional mean, 2005. Technical report, Available from http://stat.fsu.edu/~flori/ps/bnapril2005IEEE.pdf.
7. 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.
8. O. Catoni. Universal aggregation rules with exact bias bound. Preprint n.510, http://www.proba.jussieu.fr/mathdoc/preprints/index.html#1999, 1999.
9. N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.
10. L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, 1996.
11. 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.
12. A. Juditsky, P. Rigollet, and A.B. Tsybakov. Learning by mirror averaging. Preprint n.1034, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, http://arxiv.org/abs/math/0511468, 2006.
13. 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.
14. V.G. Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, pages 153–173, 1998.
15. Y. Yang. Combining different procedures for adaptive regression. *Journal of multivariate analysis*, 74:135–161, 2000.
16. T. Zhang. Data dependent concentration bounds for sequential prediction algorithms. In *Proceedings of the 18th annual conference on Computational Learning Theory (COLT), Lecture Notes in Computer Science*, pages 173–187, 2005.