1 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 $h_1 \equiv \operatorname{argmin}_{y \in \mathcal{Y}} \ell_{y_1}(y)$ and $h_2 \equiv \operatorname{argmin}_{y \in \mathcal{Y}} \ell_{y_2}(y)$. 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.

2 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. It can be proved by using [3, Theorem 14.5], but this would lead to much worse constants. Here we will rather use [1, Inequality (8.17)]. 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 $\phi_{y_1,y_2}(p) = \min_{y \in \{\tilde{y_1}, \tilde{y_2}\}} \big\{p\ell(y_1, y) + (1-p)\ell(y_2, y)\big\}$. We get

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

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.

3 Full proof of the lower bound of Theorem 3

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 $\{0, \ldots, n\}$, $\pi_{-\lambda\Sigma_i}$ 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 $R(\hat{g}_{\mathrm{pim}})$ 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: $\ell''_y \geq \lambda(\ell'_y)^2$ and since $\ell'_y(a) < 0$, y_1 and $\tilde{y_1}$ exist.

3.1 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

$$\delta \triangleq \ell(y_1, \tilde{y_2}) - \ell(y_1, \tilde{y_1}) > 0. \tag{1}$$

We have

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

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 $i \in \{1, \ldots, n\}$, we have

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

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

$$\pi_{-\lambda\Sigma_i}(g_1) = \tfrac{e^{-\lambda\Sigma_i(g_1)}}{e^{-\lambda\Sigma_i(g_1)} + e^{-\lambda\Sigma_i(g_2)}} = \tfrac{1}{1 + e^{\lambda[\Sigma_i(g_1) - \Sigma_i(g_2)]}} = \tfrac{1}{1 + e^{-\lambda\delta S_i}}. \tag{3}$$

3.2 An excursion event on which the progressive indirect mixture rules will not perform well.

Equality (3) leads us to consider the event:

$$E_\tau = \big\{\forall i \in \{\tau, \ldots, n\},\ S_i \leq -\tau\big\},$$

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 (9) below.) We have

$$\frac{\log n}{\lambda\delta} \leq \tau \leq \frac{\log n}{\lambda\delta} + 2. \tag{4}$$

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

From (3), on the event E_τ, for any $i \in \{\tau,\ldots,n\}$, we have

$$\pi_{-\lambda\Sigma_i}(g_1) \leq \tfrac{1}{n+1}. \tag{5}$$

This means that $\pi_{-\lambda\Sigma_i}$ concentrates on the wrong function, i.e. the function g_2 having larger risk (see (2)).

3.3 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 $\sigma_1,\ldots,\sigma_N$ 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 1 *Let m and t be positive integers. We have*

$$\mathbb{P}\big(\max_{1\leq k\leq N} s_k \geq t; s_N \neq t; \big|s_N - t\big| \leq m\big) = 2\mathbb{P}\big(t < s_N \leq t+m\big) \tag{6}$$

Proof 1 (of Lemma 1) *The result comes from the well known mirror trick used to compute the law of $\big(\sup_{s\leq t} W_s, W_t\big)$ where W denotes a Brownian motion. Consider a sequence $\sigma_1,\ldots,\sigma_N$ 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 $\sigma_1,\ldots,\sigma_N$ and $\sigma_1,\ldots,\sigma_J,-\sigma_{J+1},\ldots,-\sigma_N$ 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

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

which is the desired result.

Let $\sigma'_1,\ldots,\sigma'_N$ 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 2 *For any set $A \subset \big\{(\epsilon_1,\ldots,\epsilon_N) \in \{-1,1\}^n : \big|\sum_{i=1}^N \epsilon_i\big| \leq M\big\}$ where M is a positive integer, we have*

$$\mathbb{P}\big\{(\sigma'_1,\ldots,\sigma'_N) \in A\big\} \geq \Big(\tfrac{1-\gamma}{1+\gamma}\Big)^{M/2}\big(1-\gamma^2\big)^{N/2}\mathbb{P}\big\{(\sigma_1,\ldots,\sigma_N) \in A\big\} \tag{7}$$

Proof 2 (of Lemma 2) *Let s be an integer such that $N-s$ is even and $|s| \leq M$ Consider a sequence $\epsilon_1,\ldots,\epsilon_N$ 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*

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

hence

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

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 $j \in \{1,\dots,n\}$. By using Lemma 2 with $N = n-2\tau$, we obtain

$$\begin{aligned}
\mathbb{P}(E_\tau) &\geq \mathbb{P}\big(W_1=-1,\dots,W_{2\tau}=-1;\ \forall\, 2\tau<i\leq n,\ \textstyle\sum_{j=2\tau+1}^{i} W_j \leq \tau\big)\\
&= \big(\tfrac{1-\gamma}{2}\big)^{2\tau}\mathbb{P}\big(\forall\, i>2\tau \quad \textstyle\sum_{j=2\tau+1}^{i} W_j \leq \tau\big)\\
&= \big(\tfrac{1-\gamma}{2}\big)^{2\tau}\mathbb{P}\big(\forall\, i\in\{1,\dots,N\} \quad \textstyle\sum_{j=1}^{i} \sigma'_j \leq \tau\big) \qquad (8)\\
&\geq \big(\tfrac{1-\gamma}{2}\big)^{2\tau}\mathbb{P}\big(\big|\textstyle\sum_{i=1}^{N}\sigma'_i\big|<M;\forall\, i\in\{1,\dots,N\} \quad \textstyle\sum_{j=1}^{i}\sigma'_j\leq\tau\big)\\
&\geq \big(\tfrac{1-\gamma}{2}\big)^{2\tau}\big(\tfrac{1-\gamma}{1+\gamma}\big)^{M/2}\big(1-\gamma^2\big)^{\frac{N}{2}}\mathbb{P}\big(|s_N|\leq M;\forall\, i\in\{1,\dots,N\} \quad s_i\leq\tau\big)
\end{aligned}$$

By using Lemma 1, since $\tau \leq M$, the r.h.s. probability can be lower bounded:

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

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

$$\begin{aligned}
&\mathbb{P}\big(|s_N|\leq M;\max_{1\leq i\leq N} s_i\leq\tau\big)\\
&\qquad\geq\textstyle\sum_{k=0}^{\tau-1}\mathbb{P}(s_N=2-\tau+2k)-\sum_{k=1}^{\tau}\mathbb{P}(s_N=M+2k) \qquad (9)\\
&\qquad\geq\tau[\mathbb{P}(s_N=\tau)-\mathbb{P}(s_N=M)],
\end{aligned}$$

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

Combining (8) and (9), we obtain

$$\mathbb{P}(E_\tau) \quad\geq\quad \tau\big(\tfrac{1-\gamma}{2}\big)^{2\tau}\big(\tfrac{1-\gamma}{1+\gamma}\big)^{M/2}\big(1-\gamma^2\big)^{\frac{N}{2}}\left[\mathbb{P}(s_N=\tau)-\mathbb{P}(s_N=M)\right] \qquad (10)$$

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

$$n^n e^{-n}\sqrt{2\pi n}\,e^{1/(12n+1)} < n! < n^n e^{-n}\sqrt{2\pi n}\,e^{1/(12n)}, \qquad (11)$$

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

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

and similarly

$$\mathbb{P}(s_N = s) \leq \sqrt{\tfrac{2}{\pi N}}\Big(1 - \tfrac{s^2}{N^2}\Big)^{-\frac{N}{2}} \Big(\tfrac{1-\frac{s}{N}}{1+\frac{s}{N}}\Big)^{\frac{s}{2}} e^{\frac{1}{12N+1}} \tag{13}$$

These computations and (10) leads us to take M as the smallest integer larger than $\sqrt{n}$ such that $n - M$ is even. Indeed, from (4), (12) and (13), 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

$$\mathbb{P}(E_\tau) \quad \geq \quad \tfrac{c\tau}{2\sqrt{n}}\big(\tfrac{1-\gamma}{2}\big)^{2\tau}\big(\tfrac{1-\gamma}{1+\gamma}\big)^{M/2}\big(1-\gamma^2\big)^{\frac{N}{2}} \tag{14}$$

The last two terms of the r.h.s. of (14) 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 3 *If $\gamma = \sqrt{C_0(\log n)/n}$ with C_0 a positive constant, then for any large enough n,*

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

3.4 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 $i \in \{\tau, \ldots, n\}$, by definition of $\hat{h}_i$, we have

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

In particular, for any n large enough, we have $\ell[y_2, \hat{h}_i(x)] - \ell(y_2, \tilde{y_2}) \leq Cn^{-1}$, 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

$$\begin{array}{rcl} \ell[y_2, \hat{g}_{\text{pim}}(x)] - \ell(y_2, \tilde{y_2}) & = & \ell[y_2, \frac{1}{n+1}\sum_{i=0}^n \hat{h}_i(x)] - \ell(y_2, \tilde{y_2}) \\ & \leq & \frac{1}{n+1}\sum_{i=0}^n \ell[y_2, \hat{h}_i(x)] - \ell(y_2, \tilde{y_2}) \\ & \leq & \frac{\tau\delta}{n+1} + Cn^{-1} \\ & < & C_1\frac{\log n}{n} \end{array} \tag{15}$$

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

$$\tilde{y_2} \leq \hat{g}_{\text{pim}}(x) \leq \tilde{y_2} + C\sqrt{\tfrac{\log n}{n}} \leq \tilde{y_1}, \tag{16}$$

with $C > 0$ *independent from* γ.

Proof 3 *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 $\ell'_{y_1}(\tilde{y_1}) \leq 0$, $\ell''_{y_1}(\tilde{y_1}) > 0$, $\ell''_{y_1} \geq \lambda(\ell'_{y_1})^2$ 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 $\tilde{y_2} < y \leq a$, from Taylor's expansion, we have*

$$\begin{array}{rcl} \ell(y_2, y) - \ell(y_2, \tilde{y_2}) & > & (t - \tilde{y_1})\ell'_{y_1}(\tilde{y_1}) + \frac{(t-\tilde{y_1})^2}{2}m \\ & \geq & \frac{(t-\tilde{y_1})^2}{2}m \\ & = & \frac{(y-\tilde{y_2})^2}{2}m \end{array} \tag{17}$$

Let $y_0 \triangleq \tilde{y_2} + \sqrt{\frac{2C_1\log n}{mn}}$ where C_1 is the constant appearing in (15)*. For n large enough, we have $y_0 \leq a$ and we may apply* (17) *to $y = y_0$. We get*

$$\ell(y_2, y_0) - \ell(y_2, \tilde{y_2}) > C_1\tfrac{\log n}{n}. \tag{18}$$

Since ℓ_{y_1} is convex, $\ell'_{y_1}(\tilde{y_1}) \leq 0$ 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 (15) *and* (18)*, we get $\hat{g}_{pim}(x) \notin [y_0; +\infty[$, which ends the proof of the upper bound of $\hat{g}_{pim}(x)$.*

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

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

By Jensen's inequality, we obtain

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

Since the function ℓ_{y_1} decreases on $]-\infty; \tilde{y_2}] \cap \mathcal{Y}$, we get that $\hat{g}_{pim}(x) \geq \tilde{y_2}$, which ends the proof of (16)*.*

From (16), we obtain

$$\begin{aligned} R(\hat{g}_{\text{pim}}) - R(g_1) &= \tfrac{1+\gamma}{2} \big[\ell(y_1, \hat{g}_{\text{pim}}) - \ell(y_1, \tilde{y_1})\big] + \tfrac{1-\gamma}{2} \big[\ell(y_2, \hat{g}_{\text{pim}}) - \ell(y_2, \tilde{y_1})\big] \\ &= \tfrac{1+\gamma}{2} \big[\ell_{y_1}(\hat{g}_{\text{pim}}) - \ell_{y_1}(\tilde{y_1})\big] + \tfrac{1-\gamma}{2} \big[\ell_{y_1}(2a - \hat{g}_{\text{pim}}) - \ell_{y_1}(\tilde{y_2})\big] \\ &= \tfrac{1+\gamma}{2} \big[\delta + \ell_{y_1}(\hat{g}_{\text{pim}}) - \ell_{y_1}(\tilde{y_2})\big] \\ &\qquad + \tfrac{1-\gamma}{2} \big[-\delta + \ell_{y_1}(2a - \hat{g}_{\text{pim}}) - \ell_{y_1}(\tilde{y_1})\big] \\ &\geq \gamma\delta - (\hat{g}_{\text{pim}} - \tilde{y_2}) |\ell'_{y_1}(\tilde{y_2})| \\ &\geq \gamma\delta - C_2 \sqrt{\tfrac{\log n}{n}}, \end{aligned} \tag{19}$$

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 $R(\hat{g}_{\text{pim}}) - R(g_1) \geq C \sqrt{\log n / n}$. From Lemma 3, 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,

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

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 only the progressive mixture rule (instead of any progressive indirect mixture rule), this last part of the proof would have been much simpler. Indeed, for n large enough, on the event E_τ, from* (5)*, 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 $\hat{g}_{pm} = \frac{1}{n+1} \sum_{i=0}^n \mathbb{E}_{g \sim \pi_{-\lambda \Sigma_i}} \, g = p g_1 + (1-p) g_2 \equiv \tilde{y_2} + p(\tilde{y_1} - \tilde{y_2})$. So we have

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

which is much stronger than (16) *(and much simpler to prove).*

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 Gábor Lugosi and Hans-Ulrich Simon, editors, *Proceedings of the 19th annual conference on Computational Learning Theory (COLT), Lecture Notes in Computer Science*, volume 4005 of *Lecture Notes in Computer Science*, pages 392–407. Springer, 2006.

[3] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.