ADAPTING TO UNKNOWN SMOOTHNESS BY AGGREGATION OF THRESHOLDED WAVELET ESTIMATORS

Christophe Chesneau and Guillaume Lecué

University of Caen, and University Pierre et Marie Curie, Paris VI

Supplementary Material

This note contains the proof of Theorem 2.1.

S1. Notations.

Aggregation Procedures. Let us recall that we consider the following aggregation procedure:

$$\tilde{f}_n = \sum_{f \in \mathcal{F}_0} w^{(n)}(f) f, \tag{2.1}$$

where the exponential weights $w^{(n)}(f)$ are defined by

$$w^{(n)}(f) = \exp\left(-nA_n(f)\right) / \sum_{g \in \mathcal{F}_0} \exp\left(-nA_n(g)\right). \tag{2.2}$$

Oracle Inequalities. Let us recall that the quantity $\gamma = \gamma(n, M, \kappa, \mathcal{F}_0, \pi, Q)$ is defined by

$$\gamma = \begin{cases} \left(\mathcal{B}^{1/\kappa} \log M/(\beta_1 n)\right)^{1/2} & \text{if } \mathcal{B} \geq (\log M/\beta_1 n)^{\kappa/(2\kappa-1)}, \\ \\ (\log M/(\beta_2 n))^{\kappa/(2\kappa-1)} & \text{otherwise,} \end{cases} \tag{2.3}$$

where $\mathcal{B} = \mathcal{B}(\mathcal{F}_0, \pi, Q) = \min_{f \in \mathcal{F}_0} (A(f) - A^*)$, $\kappa \geq 1$ is the margin parameter, π is the underlying probability measure, Q is the loss function,

$$\beta_1 = \min\left(\log 2/(96cK), 3\log 2/(16K\sqrt{2}), (8(4c + K/3))^{-1}, (576c)^{-1}\right) \tag{2.4}$$

and

$$\beta_2 = \min\left(8^{-1}, 3\log 2/(32K), (2(16c + K/3))^{-1}, \beta_1/2\right), \tag{2.5}$$

where the constant $c > 0$ appears in the margin assumption MA($\kappa, c, \mathcal{F}_0$) and K is considered in the following theorem.

S2. Main result.

Theorem 2.1 *Let us consider the general framework introduced in the beginning of Section 2. Let $M \geq 2$ be an integer. Let $\mathcal{F}_0$ denote a finite subset of M elements $f_1, \dots, f_M$ in $\mathcal{F}$. Assume that the underlying probability measure π satisfies the margin assumption MA($\kappa, c, \mathcal{F}_0$) for some $\kappa \geq 1, c > 0$. Assume that $f \longmapsto Q(z, f)$ is convex for π-almost $z \in \mathcal{Z}$ and, for any $f \in \mathcal{F}_0$, there exists a constant $K \geq 1$ such that $|Q(Z, f) - Q(Z, f^*)| \leq K$. Then, the AEW procedure $\tilde{f}_n$ defined by (2.1) satisfies*

$$\mathbb{E}\left[A(\tilde{f}_n) - A^*\right] \leq \min_{j=1,\dots,M} \{A(f_j) - A^*\} + 4\gamma,$$

where $\gamma = \gamma(n, M, \kappa, \mathcal{F}_0, \pi, Q)$ is defined by (2.3).

Proof of Theorem 2.1: preliminaries. First of all, let us recall the notations of the general framework introduced in the beginning of Section 2. Consider a loss function $Q : \mathcal{Z} \times \mathcal{F} \longmapsto \mathbb{R}$, the risk $A(f) = \mathbb{E}[Q(Z, f)]$, the minimum risk $A^* = \min_{f \in \mathcal{F}} A(f)$, where we assume, w.l.o.g., that it is achieved by an element f^* in $\mathcal{F}$ and, for any $f \in \mathcal{F}$, the empirical risk $A_n(f) = (1/n) \sum_{i=1}^n Q(Z_i, f)$. Now, let us consider the convex set $\mathcal{C}$ defined by

$$\mathcal{C} = \Big\{(\theta_1, \dots, \theta_M) : \theta_j \geq 0, \forall j = 1, \dots, M, \text{ and } \sum_{j=1}^M \theta_j = 1\Big\}. \tag{2.6}$$

For any $\theta \in \mathcal{C}$, we define the functions $\tilde{A}(\theta)$ and $\tilde{A}_n(\theta)$ by

$$\tilde{A}(\theta) = \sum_{j=1}^M \theta_j A(f_j) \qquad \text{and} \qquad \tilde{A}_n(\theta) = \sum_{j=1}^M \theta_j A_n(f_j).$$

The first function is the linear version of the risk A. The second is the empirical version of this risk.

We are now in position to explain the form of the exponential weights described by (2.2). By virtue of the Lagrange method of optimization, we find that the exponential weights $w = (w^{(n)}(f_j))_{1\le j\le M}$ are the unique solution of the minimization problem

$$\min_{(\theta_1,\dots,\theta_M)\in\mathcal{C}} \Big\{ \tilde{A}_n(\theta) + (1/n)\sum_{j=1}^{M} \theta_j \log \theta_j \Big\}, \tag{2.7}$$

where we use the convention $0\log 0 = 0$. Take $\hat{\jmath} \in \{1,\dots,M\}$ such that $A_n(f_{\hat{\jmath}}) = \min_{j=1,\dots,M} A_n(f_j)$. If e_j denotes the vector in $\mathcal{C}$ with 1 for j-th coordinate and 0 elsewhere, then, by (2.7), the vector of exponential weights w satisfies

$$\tilde{A}_n(w) + (1/n)\sum_{j=1}^{M} w^{(n)}(f_j)\log w^{(n)}(f_j) \le \tilde{A}_n(e_{\hat{\jmath}}).$$

Using the fact that $\sum_{j=1}^{M} w^{(n)}(f_j)\log(Mw^{(n)}(f_j)) \ge 0$ (because this is the Kullback-Leibler divergence between the weights w and the uniform weights), we obtain

$$\tilde{A}_n(w) \le \tilde{A}_n(e_{\hat{\jmath}}) + \log M/n. \tag{2.8}$$

Now, observe that a linear function achieves its maximum over a convex polygon at one of the vertices of the polygon. Thus, for $j_0 \in \{1,\dots,M\}$ such that $\tilde{A}(e_{j_0}) = \min_{j=1,\dots,M} \tilde{A}(e_j)$ $(= \min_{j=1,\dots,M} A(f_j))$, we have $\tilde{A}(e_{j_0}) = \min_{\theta\in\mathcal{C}} \tilde{A}(\theta)$. We obtain the last inequality by linearity of $\tilde{A}$ and the convexity of $\mathcal{C}$. We define $\hat{w}$ by either:

$$\hat{w} = w \qquad \text{or} \qquad \hat{w} = e_{\hat{\jmath}}. \tag{2.9}$$

According to (2.8), we have

$$\tilde{A}_n(\hat{w}) \le \min_{j=1,\dots,M} \tilde{A}_n(e_j) + \log M/n \le \tilde{A}_n(e_{j_0}) + \log M/n. \tag{2.10}$$

This inequality, justified by the form of our weights, will be at the heart of the proof. Now, let us set two auxiliary lemmas.

Lemma 2.2 *Consider the framework introduced in the beginning of Section 2. Let $\mathcal{F}_0 = \{f_1,\dots,f_M\}$ be a finite subset of $\mathcal{F}$. We assume that π satisfies*

$MA(\kappa, c, \mathcal{F}_0)$, for some $\kappa \geq 1, c > 0$ and, for any $f \in \mathcal{F}_0$, there exists a constant $K \geq 1$ such that $|Q(Z,f) - Q(Z,f^*)| \leq K$. Then, for any positive numbers t, x and any integer n, we have:

$$\mathbb{P}\left[\max_{f\in\mathcal{F}} \frac{A(f) - A_n(f) - (A(f^*) - A_n(f^*))}{A(f) - A^* + x} > t\right]$$
$$\leq M\left[\left(1 + \frac{4cx^{1/\kappa}}{n(tx)^2}\right)\exp\left(-\frac{n(tx)^2}{4cx^{1/\kappa}}\right) + \left(1 + \frac{4K}{3ntx}\right)\exp\left(-\frac{3ntx}{4K}\right)\right].$$

The proof of Lemma 2.2 is postponed at the end of the proof of Theorem 2.1.

Lemma 2.3 *Let $\alpha \geq 1$ and $x, y > 0$. An integration by part yields*

$$\int_x^{+\infty} \exp\left(-yt^\alpha\right) dt \leq \exp(-yx^\alpha)/(\alpha y x^{\alpha-1}).$$

Proof of Theorem 2.1: technical details. Denote by $\tilde{A}_{\mathcal{C}}$ the minimum $\min_{\theta\in\mathcal{C}} \tilde{A}(\theta)$ where $\mathcal{C}$ is the set defined by (2.6). Using the following elementary inequality: for any $u \in \mathbb{R}$ and random variable $W \in]-\infty, K]$, we have $\mathbb{E}(W) = \mathbb{E}(W\mathbb{1}_{\{W<u\}} + W\mathbb{1}_{\{W\geq u\}}) \leq u + \int_0^K \mathbb{P}(W\mathbb{1}_{\{W\geq u\}} \geq \epsilon)d\epsilon = 2u + 2\int_{u/2}^{K/2}\mathbb{P}(W \geq 2\epsilon)d\epsilon$, we obtain:

$$\mathbb{E}[A(\tilde{f}_n) - \tilde{A}_{\mathcal{C}}] \leq \mathbb{E}\left[\tilde{A}(\hat{w}) - \tilde{A}_{\mathcal{C}}\right] \leq 2u + 2\int_{u/2}^{K/2} \mathbb{P}\left[\tilde{A}(\hat{w}) > \tilde{A}_{\mathcal{C}} + 2\epsilon\right] d\epsilon, \quad (2.11)$$

where $\hat{w}$ is defined by (2.9).

Now, let us investigate the upper bound of the term $\mathbb{P}\left[\tilde{A}(\hat{w}) > \tilde{A}_{\mathcal{C}} + 2\epsilon\right]$.

Let us consider $\mathcal{D}$, the subset of $\mathcal{C}$ defined by

$$\mathcal{D} = \left\{\theta \in \mathcal{C} : \tilde{A}(\theta) > \tilde{A}_{\mathcal{C}} + 2\epsilon\right\}.$$

If $\hat{w} \in \mathcal{D}$ then the inequality (2.10) implies the existence of $\theta \in \mathcal{D}$ such that $\tilde{A}_n(\theta) - \tilde{A}_n(f^*) \leq \tilde{A}_n(e_{j_0}) - \tilde{A}_n(f^*) + \log M/n$. Hence, for any $\epsilon > 0$, we have

$$\begin{aligned}\mathbb{P}\left[\tilde{A}(\hat{w}) > \tilde{A}_{\mathcal{C}} + 2\epsilon\right] &\leq \mathbb{P}\left[\inf_{\theta\in\mathcal{D}} \tilde{A}_n(\theta) - A_n(f^*) \leq \tilde{A}_n(e_{j_0}) - A_n(f^*) + \log M/n\right] \\ &\leq V_1 + V_2,\end{aligned}$$

where

$$V_1 = \mathbb{P}\left[\inf_{\theta\in\mathcal{D}} \tilde{A}_n(\theta) - A_n(f^*) < \tilde{A}_{\mathcal{C}} - A^* + \epsilon\right]$$

and

$$V_2 = \mathbb{P}\left[\tilde{A}_n(e_{j_0}) - A_n(f^*) \geq \tilde{A}_{\mathcal{C}} - A^* + \epsilon - \log M/n\right].$$

Let us investigate the upper bounds for V_1 and V_2, in turn.

The upper bound for V_1. We recall that $\tilde{A}_{\mathcal{C}}$ denotes the minimum $\min_{\theta\in\mathcal{C}} \tilde{A}(\theta)$. Assume that, for any $x > 0$, we have

$$\sup_{\theta\in\mathcal{D}} \frac{\tilde{A}(\theta) - A^* - (\tilde{A}_n(\theta) - A_n(f^*))}{\tilde{A}(\theta) - A^* + x} \leq \frac{\epsilon}{\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x}.$$

Since, for any $\theta \in \mathcal{D}$, $\tilde{A}(\theta) - A^* \geq \tilde{A}_{\mathcal{C}} - A^* + 2\epsilon$, we obtain

$$\tilde{A}_n(\theta) - A_n(f^*) \geq \tilde{A}(\theta) - A^* - \frac{\epsilon(\tilde{A}(\theta) - A^* + x)}{(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x)} \geq \tilde{A}_{\mathcal{C}} - A^* + \epsilon.$$

Hence, for any $x > 0$, we can bound V_1 by

$$V_1 \leq \mathbb{P}\left[\sup_{\theta\in\mathcal{D}} \frac{\tilde{A}(\theta) - A^* - [\tilde{A}_n(\theta) - A_n(f^*)]}{\tilde{A}(\theta) - A^* + x} > \frac{\epsilon}{\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x}\right]. \quad (2.12)$$

If, for any $x > 0$, we assume that

$$\sup_{\theta\in\mathcal{C}} \frac{\tilde{A}(\theta) - A^* - [\tilde{A}_n(\theta) - A_n(f^*)]}{\tilde{A}(\theta) - A^* + x} > \frac{\epsilon}{\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x},$$

then, there exists $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_M^{(0)}) \in \mathcal{C}$, such that

$$\frac{\tilde{A}(\theta^{(0)}) - A^* - [\tilde{A}_n(\theta^{(0)}) - A_n(f^*)]}{\tilde{A}(\theta^{(0)}) - A^* + x} > \frac{\epsilon}{\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x}.$$

The linearity of $\tilde{A}$ yields

$$\frac{\tilde{A}(\theta^{(0)}) - A^* - (\tilde{A}_n(\theta^{(0)}) - A_n(f^*))}{\tilde{A}(\theta^{(0)}) - A^* + x} = \frac{\sum_{j=1}^M \theta_j^{(0)}[A(f_j) - A^* - (A_n(f_j) - A_n(f^*))]}{\sum_{j=1}^M \theta_j^{(0)}[A(f_j) - A^* + x]}.$$

Let us notice that, for any numbers $a_1, \ldots, a_M$ and positive numbers $b_1, \ldots, b_M$, we have $\sum_{j=1}^M a_j / \sum_{j=1}^M b_j \leq \max_{j=1,\ldots,M}(a_j/b_j)$. It follows that

$$\max_{j=1,\ldots,M} \frac{A(f_j) - A^* - (A_n(f_j) - A_n(f^*))}{A(f_j) - A^* + x} > \frac{\epsilon}{\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x},$$

where $\tilde{A}_{\mathcal{C}} = \min_{j=1,\dots,M} A(f_j)$ (which is equal to the $\tilde{A}_{\mathcal{C}}$ previously defined).

Now, we use the relative concentration inequality of Lemma 2.2 to obtain

$$\begin{aligned}
&\mathbb{P}\left[\max_{j=1,\dots,M} \frac{A(f_j) - A^* - (A_n(f_j) - A_n(f^*))}{A(f_j) - A^* + x} > \frac{\epsilon}{\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x}\right] \\
&\leq \quad M\left(1 + \frac{4c(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x)^2 x^{1/\kappa}}{n(\epsilon x)^2}\right) \exp\left(-\frac{n(\epsilon x)^2}{4c(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x)^2 x^{1/\kappa}}\right) \\
&\qquad + M\left(1 + \frac{4K(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x)}{3n\epsilon x}\right) \exp\left(-\frac{3n\epsilon x}{4K(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x)}\right). \quad (2.13)
\end{aligned}$$

Putting (2.12) and (2.13) together, for any $x > 0$, we obtain:

$$\begin{aligned}
V_1 \quad \leq \quad & M\left(1 + \frac{4c(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x)^2 x^{1/\kappa}}{n(\epsilon x)^2}\right) \exp\left(-\frac{n(\epsilon x)^2}{4c(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x)^2 x^{1/\kappa}}\right) \\
& + M\left(1 + \frac{4K(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x)}{3n\epsilon x}\right) \exp\left(-\frac{3n\epsilon x}{4K(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x)}\right). \quad (2.14)
\end{aligned}$$

The upper bound for V_2. Using the margin assumption $\mathrm{MA}(\kappa, c, \mathcal{F}_0)$ to upper bound the variance term and applying Bernstein's inequality (cf. Massart (2006)), for any $\epsilon > \log M/n$, we get

$$V_2 \leq \exp\left(-\frac{n(\epsilon - (\log M)/n)^2}{2c(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa} + (2K/3)(\epsilon - \log M/n)}\right), \quad (2.15)$$

Combining the obtained upper bounds of V_1 with $x = \tilde{A}_{\mathcal{C}} - A^* + 2\epsilon$ and V_2, then, for any $\log M/n < \epsilon < K/2$, we have

$$\begin{aligned}
&\mathbb{P}\left(\tilde{A}(\hat{w}) > \tilde{A}_{\mathcal{C}} + 2\epsilon\right) \leq V_1 + V_2 \\
&\leq \exp\left(-\frac{n(\epsilon - \log M/n)^2}{2c(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa} + (2K/3)(\epsilon - \log M/n)}\right) \\
&\quad + M\left(1 + \frac{16c(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon)^{1/\kappa}}{n\epsilon^2}\right) \exp\left(-\frac{n\epsilon^2}{16c(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon)^{1/\kappa}}\right) \\
&\quad + M\left(1 + \frac{8K}{3n\epsilon}\right) \exp\left(-\frac{3n\epsilon}{8K}\right).
\end{aligned}$$

It follows from (2.11) that, for any $2\log M/n < u < K/2$, we have

$$\mathbb{E}[A(\tilde{f}_n) - \tilde{A}_{\mathcal{C}}] \leq 2u + 2\int_{u/2}^{K/2} [T_1(\epsilon) + M(T_2(\epsilon) + T_3(\epsilon))]\, d\epsilon, \quad (2.16)$$

where the quantities $T_1(\epsilon)$, $T_2(\epsilon)$ and $T_3(\epsilon)$ are defined by

$$T_1(\epsilon) = \exp\left(-\frac{n(\epsilon - (\log M)/n)^2}{2c(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa} + (2K/3)(\epsilon - \log M/n)}\right),$$

$$T_2(\epsilon) = \left(1 + \frac{16c(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon)^{1/\kappa}}{n\epsilon^2}\right)\exp\left(-\frac{n\epsilon^2}{16c(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon)^{1/\kappa}}\right)$$

and

$$T_3(\epsilon) = \left(1 + \frac{8K}{3n\epsilon}\right)\exp\left(-\frac{3n\epsilon}{8K}\right).$$

Now, let us investigate the upper bounds of $\int_{u/2}^1 T_1(\epsilon)d\epsilon$, $\int_{u/2}^1 T_2(\epsilon)d\epsilon$ and $\int_{u/2}^1 T_3(\epsilon)d\epsilon$, in turn. We distinguish two cases: the case where $\tilde{A}_{\mathcal{C}} - A^* \geq (\log M/(\beta_1 n))^{\kappa/(2\kappa-1)}$ and the case where $\tilde{A}_{\mathcal{C}} - A^* < (\log M/(\beta_1 n))^{\kappa/(2\kappa-1)}$. Let us recall that β_1 is defined in (2.4).

- *The case* $\tilde{A}_{\mathcal{C}} - A^* \geq (\log M/(\beta_1 n))^{\kappa/(2\kappa-1)}$. Denote by $\mu(M)$ the unique solution of the equation $\mu_0 - 3M\exp(-\mu_0) = 0$. Then, clearly $(\log M)/2 \leq \mu(M) \leq \log M$. Take u such that $(n\beta_1 u^2)/(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa} = \mu(M)$. Using the fact that $\tilde{A}_{\mathcal{C}} - A^* \geq (\log M/(\beta_1 n))^{\kappa/(2\kappa-1)}$ and the definition $\mu(M)$, we get $u \leq \tilde{A}_{\mathcal{C}} - A^*$. Moreover, since $u \geq 4\log M/n$, we have

$$\begin{aligned}\int_{u/2}^{K/2} T_1(\epsilon)d\epsilon &\leq \int_{u/2}^{(\tilde{A}_{\mathcal{C}}-A^*)/2} \exp\left(-\frac{n(\epsilon/2)^2}{(2c+K/6)(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa}}\right)d\epsilon \\ &\quad + \int_{(\tilde{A}_{\mathcal{C}}-A^*)/2}^{K/2} \exp\left(-\frac{n(\epsilon/2)^2}{(4c+K/3)\epsilon^{1/\kappa}}\right)d\epsilon.\end{aligned}$$

Using Lemma 2.3 and the inequality $u \leq \tilde{A}_{\mathcal{C}} - A^*$, we obtain

$$\int_{u/2}^{K/2} T_1(\epsilon)d\epsilon \leq \frac{8(4c+K/3)(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa}}{nu}\exp\left(-\frac{nu^2}{8(4c+K/3)(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa}}\right). \tag{2.17}$$

Since $16c(\tilde{A}_{\mathcal{C}} - A^* + 2u) \leq nu^2$, Lemma 2.3 yields

$$\begin{aligned}\int_{u/2}^{K/2} T_2(\epsilon)d\epsilon &\leq 2\int_{u/2}^{(\tilde{A}_{\mathcal{C}}-A^*)/2} \exp\left(-\frac{n\epsilon^2}{64c(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa}}\right)d\epsilon \\ &\quad +2\int_{(\tilde{A}_{\mathcal{C}}-A^*)/2}^{K/2} \exp\left(-\frac{n\epsilon^{2-1/\kappa}}{128c}\right)d\epsilon \\ &\leq \frac{2148c(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa}}{nu}\exp\left(-\frac{nu^2}{2148c(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa}}\right). \quad (2.18)\end{aligned}$$

Since $16(3n)^{-1} \leq u \leq \tilde{A}_{\mathcal{C}} - A^*$, we have

$$\int_{u/2}^{K/2} T_3(\epsilon)d\epsilon \leq \frac{16K(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa}}{3nu} \exp\left(-\frac{3nu^2}{16K(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa}}\right). \tag{2.19}$$

From (2.16), (2.17), (2.18), (2.19) and the definition of u (and, a fortiori, $\mu(M)$), we obtain

$$\begin{aligned}\mathbb{E}\left[A(\tilde{f}_n) - \tilde{A}_{\mathcal{C}}\right] &\leq 2u + 6M\frac{(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa}}{n\beta_1 u} \exp\left(-\frac{n\beta_1 u^2}{(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa}}\right) \\ &= 4u \leq 4\sqrt{(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa} \log M/(n\beta_1)}.\end{aligned}$$

- *The case* $\tilde{A}_{\mathcal{C}} - A^* < (\log M/(\beta_1 n))^{\kappa/(2\kappa-1)}$. We now choose u such that $n\beta_2 u^{(2\kappa-1)/\kappa} = \mu(M)$, where $\mu(M)$ denotes the unique solution of the equation $\mu_0 - 3M\exp(-\mu_0) = 0$ and β_2 is defined in (2.5). Using the fact that $\tilde{A}_{\mathcal{C}} - A^* < (\log M/(\beta_1 n))^{\kappa/(2\kappa-1)}$ and the definition of $\mu(M)$, we get $u \geq \tilde{A}_{\mathcal{C}} - A^*$ (since $\beta_1 \geq 2\beta_2$). Using the fact that $u > 4\log M/n$ and Lemma 2.3, we find

$$\int_{u/2}^{K/2} T_1(\epsilon)d\epsilon \leq \frac{2(16c + K/3)}{nu^{1-1/\kappa}} \exp\left(-\frac{3nu^{2-1/\kappa}}{2(16c + K/3)}\right). \tag{2.20}$$

Since $u \geq (128c/n)^{\kappa/(2\kappa-1)}$, Lemma 2.3 yields

$$\int_{u/2}^{K/2} T_2(\epsilon)d\epsilon \leq \frac{256c}{nu^{1-1/\kappa}} \exp\left(-\frac{nu^{2-1/\kappa}}{256c}\right). \tag{2.21}$$

Since $u > 16K/(3n)$, we have

$$\int_{u/2}^{K/2} T_3(\epsilon)d\epsilon \leq \frac{16K}{3nu^{1-1/\kappa}} \exp\left(-\frac{3nu^{2-1/\kappa}}{16K}\right). \tag{2.22}$$

Putting (2.16), (2.20), (2.21) and (2.22) together and using the definition of u (and, a fortiori, $\mu(M)$), we obtain

$$\mathbb{E}\left[A(\tilde{f}_n) - \tilde{A}_{\mathcal{C}}\right] \leq 2u + 6M\frac{\exp\left(-n\beta_2 u^{(2\kappa-1)/\kappa}\right)}{n\beta_2 u^{1-1/\kappa}} = 4u \leq 4(\log M/(n\beta_2))^{\kappa/(2\kappa-1)}.$$

This completes the proof of Theorem 2.1.

Proof of Lemma 2.2. We use a "peeling device". Let $x > 0$. For any integer j, we consider $\mathcal{F}_j = \{f \in \mathcal{F} : jx \leq A(f) - A^* < (j+1)x\}$. Define the empirical process $Z_x(f)$ by

$$Z_x(f) = \frac{A(f) - A_n(f) - (A(f^*) - A_n(f^*))}{A(f) - A^* + x}.$$

Using Bernstein's inequality and margin assumption $\mathrm{MA}(\kappa, c, \mathcal{F}_0)$ to upper bound the variance term, we have

$$\begin{aligned}
\mathbb{P}\left[\max_{f\in\mathcal{F}} Z_x(f) > t\right] &\leq \sum_{j=0}^{+\infty} \mathbb{P}\left[\max_{f\in\mathcal{F}_j} Z_x(f) > t\right] \\
&\leq \sum_{j=0}^{+\infty} \mathbb{P}\Big[\max_{f\in\mathcal{F}_j} A(f) - A_n(f) - (A(f^*) - A_n(f^*)) > t(j+1)x\Big] \\
&\leq M \sum_{j=0}^{+\infty} \exp\Big(- \frac{n[t(j+1)x]^2}{2c((j+1)x)^{1/\kappa} + (2K/3)t(j+1)x}\Big) \\
&\leq M \left[\sum_{j=0}^{+\infty} \exp\Big(- \frac{n(tx)^2 (j+1)^{2-1/\kappa}}{4cx^{1/\kappa}}\Big) + \exp\Big(- (j+1)\frac{3ntx}{4K}\Big)\right] \\
&\leq M \left[\exp\left(-\frac{nt^2 x^{2-1/\kappa}}{4c}\right) + \exp\left(-\frac{3ntx}{4K}\right)\right] \\
&\quad + M \int_1^{+\infty} \left[\exp\left(-\frac{nt^2x^{2-1/\kappa}}{4c} u^{2-1/\kappa}\right) + \exp\left(-\frac{3ntx}{4K}u\right)\right] du.
\end{aligned}$$

Lemma 2.3 completes the proof.