Supplementary Material for: Optimization by Gradient Boosting

Gérard Biau and Benoît Cadre

Proof (Theorem 1) Assume that, for some $t_0 \geq 0$, $\sup_{f \in \mathscr{F}} -\mathbb{E}\xi(F_{t_0}(X), Y) f(X) = 0$. Then, by the symmetry of the class $\mathscr{F}$, for all $f \in \mathscr{F}$, $\mathbb{E}\xi(F_{t_0}(X), Y) f(X) = 0$. We conclude by technical Lemma 2 that

$$C(F_t) = \inf_{F \in \operatorname{lin}(\mathscr{F})} C(F) \quad \text{for all } t \geq t_0,$$

and the result is proved. Thus, in the following, it is assumed that

$$\sup_{f \in \mathscr{F}} -\mathbb{E}\xi(F_t(X), Y) f(X) > 0 \quad \text{for all } t \geq 0.$$

Consequently, $-\mathbb{E}\xi(F_t(X), Y) f_{t+1}(X) > 0$ and $w_t > 0$ for all t. Since $w_t \to 0$ (by Lemma 1 of the Main Document), there exists a subsequence $(w_{t'})_{t'}$ such that

$$\begin{aligned} w_{t'+1} &= -(2L)^{-1} \mathbb{E}\xi(F_{t'}(X), Y) f_{t'+1}(X) \\ &= (2L)^{-1} \sup_{f \in \mathscr{F}} -\mathbb{E}\xi(F_{t'}(X), Y) f(X). \end{aligned} \tag{1}$$

Let $\varepsilon > 0$. For all t' large enough and all $f \in \mathscr{F}$, by the symmetry of $\mathscr{F}$,

$$-\mathbb{E}\xi(F_{t'}(X), Y) f(X) \leq \varepsilon \quad \text{and} \quad \mathbb{E}\xi(F_{t'}(X), Y) f(X) \leq \varepsilon,$$

and thus $\lim_{t' \to \infty} \mathbb{E}\xi(F_{t'}(X), Y) f(X) = 0$ for all $f \in \mathscr{F}$. We conclude that, for all $G \in \operatorname{lin}(\mathscr{F})$,

$$\lim_{t' \to \infty} \mathbb{E}\xi(F_{t'}(X), Y) G(X) = 0. \tag{2}$$

Gérard Biau
Sorbonne Université, CNRS, LPSM, 4 place Jussieu, 75005 Paris, France e-mail: gerard.biau@sorbonne-universite.fr

Benoît Cadre
Univ Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France e-mail: benoit.cadre@univ-rennes2.fr

Assume, without loss of generality, that $F_0 = 0$, and observe that $F_t = \sum_{k=1}^{t} w_k f_k$. Thus, we may write

$$\begin{aligned}
\mathbb{E}\xi(F_{t'}(X),Y)F_{t'}(X) &= \sum_{k=1}^{t'} w_k \mathbb{E}\xi(F_{t'}(X),Y)f_k(X) \\
&\leq \sup_{f\in\mathscr{F}} \mathbb{E}\xi(F_{t'}(X),Y)f(X)\sum_{k=1}^{t'} w_k \\
&= \sup_{f\in\mathscr{F}} -\mathbb{E}\xi(F_{t'}(X),Y)f(X)\sum_{k=1}^{t'} w_k \\
&\quad \text{(by the symmetry of } \mathscr{F}\text{)} \\
&= 2Lw_{t'+1}\sum_{k=1}^{t'} w_k,
\end{aligned}$$

by definition of $w_{t'+1}$—see (1). So,

$$\begin{aligned}
\mathbb{E}\xi(F_{t'}(X),Y)F_{t'}(X) \leq 2Lw_{t'}\sum_{k=1}^{t'} w_k &= 2Lw_{t'}\sum_{k=1}^{t'} w_k^{-1}w_k^2 \\
&\text{(because } w_{t'+1} \leq w_{t'}\text{)}.
\end{aligned}$$

Since $\sum_{k\geq 1} w_k^2 < \infty$, and since the sequence $(w_t)_t$ is nonincreasing, positive, and tends to 0 as $t \to \infty$, Kronecker's lemma reveals that $w_{t'}\sum_{k=1}^{t'} w_k^{-1}w_k^2 \to 0$ as $t' \to \infty$. Therefore,

$$\limsup_{t'\to\infty} \mathbb{E}\xi(F_{t'}(X),Y)F_{t'}(X) \leq 0. \tag{3}$$

Let $\varepsilon > 0$ and let $F_\varepsilon^\star \in \mathrm{lin}(\mathscr{F})$ be such that

$$\inf_{F\in\mathrm{lin}(\mathscr{F})} C(F) \geq C(F_\varepsilon^\star) - \varepsilon.$$

By the convexity of C, we have, for all t',

$$\begin{aligned}
\inf_{F\in\mathrm{lin}(\mathscr{F})} C(F) &\geq C(F_\varepsilon^\star) - \varepsilon \\
&\geq C(F_{t'}) + \mathbb{E}\xi(F_{t'}(X),Y)(F_\varepsilon^\star(X) - F_{t'}(X)) - \varepsilon \\
&\geq \inf_k C(F_k) + \mathbb{E}\xi(F_{t'}(X),Y)F_\varepsilon^\star(X) - \mathbb{E}\xi(F_{t'}(X),Y)F_{t'}(X) - \varepsilon.
\end{aligned}$$

Combining (2) and (3), we conclude that $\inf_{F\in\mathrm{lin}(\mathscr{F})} C(F) \geq \inf_k C(F_k) - \varepsilon$ for all $\varepsilon > 0$, so that

$$\lim_{t\to\infty} C(F_t) = \inf_k C(F_k) = \inf_{F\in\mathrm{lin}(\mathscr{F})} C(F),$$

which is the desired result. □

Proof (Theorem 2) The first step is to establish that there exists a subsequence $(F_{t''})_{t''}$ such that $\lim_{t''\to\infty} \mathbb{E}\xi(F_{t''}(X),Y)G(X) \to 0$ for all $G \in \mathrm{lin}(\mathscr{P})$. We start by observing that, by Lemma 2 of the Main Document, $C(F_t) \leq C(F_0)$. Thus, by technical Lemma 3, $\sup_t \|F_t\|_{\mu_X} \leq B$ for some positive constant B. Now,

$$\begin{aligned}
&|\mathbb{E}\xi(F_t(X),Y)f_{t+1}(X)| \\
&\quad = |\mathbb{E}\mathbb{E}(\xi(F_t(X),Y) \mid X)f_{t+1}(X)| \\
&\quad \leq \mathbb{E}\big|\mathbb{E}(\xi(F_t(X),Y) - \xi(0,Y) \mid X)\big| \cdot |f_{t+1}(X)| + \mathbb{E}|\xi(0,Y)f_{t+1}(X)| \\
&\quad \leq L\mathbb{E}|F_t(X)f_{t+1}(X)| + \mathbb{E}|\xi(0,Y)f_{t+1}(X)| \\
&\qquad \text{(by Assumption } \mathbf{A_3}\text{)}.
\end{aligned}$$

So,

$$\begin{aligned}
|\mathbb{E}\xi(F_t(X),Y)f_{t+1}(X)| &\leq L\|F_t\|_{\mu_X}\|f_{t+1}\|_{\mu_X} + \big(\mathbb{E}\xi(0,Y)^2\big)^{1/2}\|f_{t+1}\|_{\mu_X} \\
&\qquad \text{(by the Cauchy-Schwarz inequality)} \\
&\leq \big(LB + \big(\mathbb{E}\xi(0,Y)^2\big)^{1/2}\big)\|f_{t+1}\|_{\mu_X}.
\end{aligned}$$

Consequently, since $\lim_{t\to\infty} \|f_{t+1}\|_{\mu_X} = 0$ (by Lemma 2 of the Main Document),

$$\begin{aligned}
\inf_{f\in\mathscr{P}} \big(2\mathbb{E}\xi(F_t(X),Y)f(X) + \|f\|^2_{\mu_X}\big) &= 2\mathbb{E}\xi(F_t(X),Y)f_{t+1}(X) + \|f_{t+1}\|^2_{\mu_X} \\
&\to 0 \text{ as } t \to \infty.
\end{aligned}$$

Accordingly, by the symmetry of $\mathscr{P}$, for all $\varepsilon > 0$ and all t large enough, we have, for all $f \in \mathscr{P}$,

$$2\mathbb{E}\xi(F_t(X),Y)f(X) + \|f\|^2_{\mu_X} \geq -\varepsilon \quad \text{and} - 2\mathbb{E}\xi(F_t(X),Y)f(X) + \|f\|^2_{\mu_X} \geq -\varepsilon.$$

So, for all t large enough and all $f \in \mathscr{P}$,

$$|2\mathbb{E}\xi(F_t(X),Y)f(X)| \leq \varepsilon + \|f\|^2_{\mu_X}.$$

Since ε was arbitrary, we conclude that, for all $f \in \mathscr{P}$,

$$2\limsup_{t\to\infty}|\mathbb{E}\xi(F_t(X),Y)f(X)| \leq \|f\|^2_{\mu_X}. \tag{4}$$

On the other hand, by Assumption $\mathbf{A_3}$,

$$|\mathbb{E}(\xi(F_t(X),Y) \mid X)| \leq \mathbb{E}(|\xi(0,Y)| \,\big|\, X) + L|F_t(X)|.$$

Since $\sup_t \|F_t\|_{\mu_X} < \infty$, we deduce that

$$\sup_t \|\mathbb{E}(\xi(F_t(X),Y) \mid X = \cdot)\|_{\mu_X} < \infty.$$

Next, since $\sum_{k\geq 1}\|f_k\|^2_{\mu_X} < \infty$, there exists a subsequence $(f_{t'})_{t'}$ satisfying $t'\|f_{t'+1}\|^2_{\mu_X} \to 0$. Besides, recalling that the unit ball of $L^2(\mu_X)$ is weakly compact, there exists a subsequence $(F_{t''})_{t''}$ of $(F_{t'})_{t'}$ and $\tilde{F} \in L^2(\mu_X)$ such that, for all $G \in \mathrm{lin}(\mathscr{P})$,

$$\mathbb{E}\xi(F_{t''}(X), Y)G(X) = \mathbb{E}\mathbb{E}(\xi(F_{t''}(X), Y) \mid X)G(X) \to \mathbb{E}\tilde{F}(X)G(X).$$

Combining this identity with (4) reveals that $2|\mathbb{E}\tilde{F}(X)f(X)| \leq \|f\|^2_{\mu_X}$ for all $f \in \mathscr{P}$. In particular, for all $\varepsilon > 0$ and all $f \in \mathscr{P}$, $2|\mathbb{E}\tilde{F}(X)\varepsilon f(X)| \leq \varepsilon^2\|f\|^2_{\mu_X}$, and thus, letting $\varepsilon \downarrow 0$, we find that $\mathbb{E}\tilde{F}(X)f(X) = 0$ for all $f \in \mathscr{P}$. By a linearity argument, we conclude that $\mathbb{E}\tilde{F}(X)G(X) = 0$ for all $G \in \mathrm{lin}(\mathscr{P})$. Therefore, for all $G \in \mathrm{lin}(\mathscr{P})$,

$$\lim_{t''\to\infty} \mathbb{E}\xi(F_{t''}(X), Y)G(X) = 0, \tag{5}$$

which was our first objective.

The next step is to prove that $\limsup_{t''\to\infty}\mathbb{E}\xi(F_{t''}(X), Y)F_{t''}(X) \leq 0$. To simplify the notation, we assume, without loss of generality, that $F_0 = 0$. Fix $\varepsilon > 0$. Since $\sum_{k\geq 1}\|f_k\|^2_{\mu_X} < \infty$, there exists $T \geq 0$ such that $\sum_{k\geq T+1}\|f_k\|^2_{\mu_X} \leq \varepsilon$. In addition, for all $t > T$, $F_t = F_T + \nu\sum_{k=T+1}^t f_k$, so that

$$\mathbb{E}\xi(F_t(X), Y)F_t(X) = \mathbb{E}\xi(F_t(X), Y)F_T(X) + \nu\sum_{k=T+1}^{t}\mathbb{E}\xi(F_t(X), Y)f_k(X). \tag{6}$$

Also, by the very definition of f_{t+1} and the symmetry of $\mathscr{P}$, we have, for all $f \in \mathscr{P}$,

$$2\mathbb{E}\xi(F_t(X), Y)f_{t+1}(X) + \|f_{t+1}\|^2_{\mu_X} \leq -2\mathbb{E}\xi(F_t(X), Y)f(X) + \|f\|^2_{\mu_X}, \tag{7}$$

i.e., for all $f \in \mathscr{P}$,

$$2\mathbb{E}\xi(F_t(X), Y)f(X) \leq -2\mathbb{E}\xi(F_t(X), Y)f_{t+1}(X) - \|f_{t+1}\|^2_{\mu_X} + \|f\|^2_{\mu_X}.$$

Using (6), this leads to

$$\begin{aligned}
&\mathbb{E}\xi(F_t(X), Y)F_t(X)\\
&\quad\leq \mathbb{E}\xi(F_t(X), Y)F_T(X)\\
&\qquad + \frac{\nu}{2}\Big(t\big(-2\mathbb{E}\xi(F_t(X), Y)f_{t+1}(X) - \|f_{t+1}\|^2_{\mu_X}\big) + \sum_{k\geq T+1}\|f_k\|^2_{\mu_X}\Big)\\
&\quad\leq \frac{\varepsilon\nu}{2} + \mathbb{E}\xi(F_t(X), Y)F_T(X) + \frac{\nu t}{2}\big(-2\mathbb{E}\xi(F_t(X), Y)f_{t+1}(X) - \|f_{t+1}\|^2_{\mu_X}\big).
\end{aligned} \tag{8}$$

But, according to inequality (7) applied with $f = -2f_{t+1}$ (which belongs to $\mathscr{P}$ by assumption),

$$2\mathbb{E}\xi(F_t(X), Y)f_{t+1}(X) + \|f_{t+1}\|^2_{\mu_X} \leq 4\mathbb{E}\xi(F_t(X), Y)f_{t+1}(X) + 4\|f_{t+1}\|^2_{\mu_X},$$

i.e.,

$$-2\mathbb{E}\xi(F_t(X),Y)f_{t+1}(X) \leq 3\|f_{t+1}\|_{\mu_X}^2.$$

Combining this inequality with (8) shows that

$$\mathbb{E}\xi(F_t(X),Y)F_t(X) \leq \frac{\varepsilon\nu}{2} + \mathbb{E}\xi(F_t(X),Y)F_T(X) + \nu t\|f_{t+1}\|_{\mu_X}^2.$$

Since $F_T \in \mathrm{lin}(\mathscr{P})$, we know from (5) that $\mathbb{E}\xi(F_{t''}(X),Y)F_T(X) \to 0$. Therefore, recalling that $t''\|f_{t''+1}\|_{\mu_X}^2 \to 0$, for all $\varepsilon > 0$,

$$\limsup_{t''\to\infty}\mathbb{E}\xi(F_{t''}(X),Y)F_{t''}(X) \leq \frac{\varepsilon\nu}{2}.$$

Since ε is arbitrary, we have just shown that

$$\limsup_{t''\to\infty}\mathbb{E}\xi(F_{t''}(X),Y)F_{t''}(X) \leq 0, \tag{9}$$

as desired.

Let $\varepsilon > 0$ and let $F_\varepsilon^\star \in \mathrm{lin}(\mathscr{P})$ be such that

$$\inf_{F\in\mathrm{lin}(\mathscr{P})} C(F) \geq C(F_\varepsilon^\star) - \varepsilon.$$

By the convexity of C, along t'',

$$\begin{aligned}\inf_{F\in\mathrm{lin}(\mathscr{P})} C(F) &\geq C(F_\varepsilon^\star) - \varepsilon\\ &\geq \inf_k C(F_k) + \mathbb{E}\xi(F_{t''}(X),Y)F_\varepsilon^\star(X) - \mathbb{E}\xi(F_{t''}(X),Y)F_{t''}(X) - \varepsilon.\end{aligned}$$

Putting (5) and (9) together, we conclude that

$$\lim_{t\to\infty} C(F_t) = \inf_k C(F_k) = \inf_{F\in\mathrm{lin}(\mathscr{P})} C(F).$$

$\square$

Proof (Theorem 3) For $\beta \in \mathbb{R}^N$, we let $F_\beta = \sum_{j=1}^N \beta_j \mathbb{1}_{A_j^n}$ and notice that $\bar{F}_n = F_\alpha$ for some (data-dependent) $\alpha \in \mathbb{R}^N$. Let the event S be defined by

$$S = \left\{\forall j = 1,\ldots,N : P_n(A_j^n) \geq P(A_j^n)/2\right\}.$$

Observe that

$$\|\bar{F}_n\|_{P_n}^2 \leq \frac{\frac{1}{n}\sum_{i=1}^n \phi(0,Y_i)}{\gamma_n} \leq \frac{\bar{\phi}}{\gamma_n},$$

and, similarly, that

$$\|\bar{F}_n\|_{P_n}^2 = \sum_{j=1}^N \alpha_j^2 P_n(A_j^n).$$

Therefore, on S,

$$\frac{1}{2}\sum_{j=1}^{N}\alpha_j^2 P(A_j^n) \leq \frac{\bar{\phi}}{\gamma_n},$$

and so

$$\frac{\inf_{\mathscr{X}} g}{2} \cdot \nu_n \sum_{j=1}^{N}\alpha_j^2 \leq \frac{\bar{\phi}}{\gamma_n}.$$

We have just shown that, on the event S, $\alpha \in T$, where

$$T = \Big\{\beta \in \mathbb{R}^N : \sum_{j=1}^{N}\beta_j^2 \leq \frac{2\bar{\phi}}{\inf_{\mathscr{X}} g}\cdot\frac{1}{\nu_n\gamma_n}\Big\}.$$

Now, observe that

$$\begin{aligned}
\mathbb{E}C_n(\bar{F}_n) &= \mathbb{E}\inf_{F\in\mathrm{lin}(\mathscr{F}_n)} C_n(F)\\
&= \mathbb{E}\inf_{F\in\mathrm{lin}(\mathscr{F}_n)} C_n(F)\mathbb{1}_S + \mathbb{E}\inf_{F\in\mathrm{lin}(\mathscr{F}_n)} C_n(F)\mathbb{1}_{S^c}\\
&\leq \mathbb{E}\inf_{F\in\mathrm{lin}(\mathscr{F}_n)} C_n(F)\mathbb{1}_S + \mathbb{E}C_n(0)\mathbb{1}_{S^c}\\
&= \mathbb{E}\inf_{\beta\in T} C_n(F_\beta)\mathbb{1}_S + \mathbb{E}A_n(0)\mathbb{1}_{S^c}\\
&\leq \mathbb{E}\inf_{\beta\in T} C_n(F_\beta) + \bar{\phi}\mathbb{P}(S^c).
\end{aligned}$$

Define

$$D_n(F) = A(F) + \gamma_n\|F\|_{P_n}^2.$$

Since $C_n(F) - D_n(F) = A_n(F) - A(F)$, we deduce from Lemma 4 and Lemma 6 that whenever

$$\frac{\log N}{n\nu_n} \to 0 \quad \text{and} \quad \frac{1}{\sqrt{n\nu_n\gamma_n}}\zeta\left(\sqrt{\frac{2\bar{\phi}}{\nu_n\gamma_n\inf_{\mathscr{X}} g}}\right) \to 0,$$

we have

$$\begin{aligned}
\limsup_{n\to\infty}\mathbb{E}C_n(\bar{F}_n) &\leq \limsup_{n\to\infty}\mathbb{E}\inf_{\beta\in T} C_n(F_\beta)\\
&\leq \limsup_{n\to\infty}\mathbb{E}\inf_{\beta\in T} D_n(F_\beta) + \limsup_{n\to\infty}\big(\mathbb{E}\sup_{\beta\in T}|A_n(F_\beta) - A(F_\beta)|\big)\\
&= \limsup_{n\to\infty}\mathbb{E}\inf_{\beta\in T} D_n(F_\beta). \qquad (10)
\end{aligned}$$

Let $\varepsilon > 0$. By Lemma 5, there exists $(\beta_1^\varepsilon, \ldots, \beta_N^\varepsilon) \in T$ such that

$$\Big\|F^\star - \sum_{j=1}^{N}\beta_j^\varepsilon\mathbb{1}_{A_j^n}\Big\|_P \leq \varepsilon.$$

Define $F_\varepsilon^\star = \sum_{j=1}^{N}\beta_j^\varepsilon\mathbb{1}_{A_j^n}$. Then, according to (10),

$$\begin{aligned}
\limsup_{n\to\infty} \mathbb{E}C_n(\bar{F}_n) &\le \limsup_{n\to\infty} \big(A(F_\varepsilon^\star) + \gamma_n \mathbb{E}\|F_\varepsilon^\star\|_{P_n}^2\big) \\
&= \limsup_{n\to\infty} \big(A(F_\varepsilon^\star) + \gamma_n \|F_\varepsilon^\star\|_{P}^2\big) \\
&\le A(F_\varepsilon^\star). \qquad (11)
\end{aligned}$$

Since A is continuous, we conclude that $\limsup_{n\to\infty} \mathbb{E}C_n(\bar{F}_n) \le A(F^\star)$.

On the other hand, $C_n(\bar{F}_n) \ge A_n(\bar{F}_n)$, and, by Lemma 4 and Lemma 6,

$$\begin{aligned}
\mathbb{E}|A_n(\bar{F}_n) - A(\bar{F}_n)| &\le \mathbb{E}\sup_{\beta\in T} |A_n(F_\beta) - A(F_\beta)| + \bar{\phi}\mathbb{P}(S^c) \\
&\to 0 \quad \text{as } n \to \infty.
\end{aligned}$$

Therefore,

$$\limsup_{n\to\infty} \mathbb{E}A(\bar{F}_n) \le \limsup_{n\to\infty} C_n(\bar{F}_n).$$

So, with (11),

$$\limsup_{n\to\infty} \mathbb{E}A(\bar{F}_n) \le A(F^\star),$$

which is the desired result. □

Some technical lemmas

Lemma 1 *Assume that Assumptions* $\mathbf{A_1}$ *and* $\mathbf{A_3}$ *are satisfied. Then, for all* $a > 0$ *and all* $F, G \in L^2(\mu_X)$,

$$C(F) - C(F + aG) \ge -a^2 L\|G\|_{\mu_X}^2 - a\mathbb{E}\xi(F(X), Y)G(X).$$

Proof By inequality (1) of the Main Document,

$$\begin{aligned}
C(F) &\ge C(F + aG) - a\mathbb{E}\xi(F(X) + aG(X), Y)G(X) \\
&= C(F + aG) - a\mathbb{E}(\xi(F(X) + aG(X), Y) - \xi(F(X), Y))G(X) \\
&\quad - a\mathbb{E}\xi(F(X), Y)G(X) \\
&= C(F + aG) - a\mathbb{E}\mathbb{E}(\xi(F(X) + aG(X), Y) - \xi(F(X), Y) \mid X)G(X) \\
&\quad - a\mathbb{E}\xi(F(X), Y)G(X) \\
&\ge C(F + aG) - a\big(\mathbb{E}\mathbb{E}^2(\xi(F(X) + aG(X), Y) - \xi(F(X), Y) \mid X)\big)^{1/2}\|G\|_{\mu_X} \\
&\quad - a\mathbb{E}\xi(F(X), Y)G(X) \\
&\quad \text{(by the Cauchy-Schwarz inequality).}
\end{aligned}$$

Thus, by Assumption $\mathbf{A_3}$,

$$C(F) \geq C(F + aG) - a^2 L\|G\|^2_{\mu_X} - a\mathbb{E}\xi(F(X),Y)G(X).$$

□

Lemma 2 *Assume that Assumption* $\mathbf{A_1}$ *is satisfied, and let* $(F_t)_t$ *be defined by Algorithm 1 with* $(w_t)_t$ *as in (8) of the Main Document. If, for some* $t_0 \geq 0$,

$$\mathbb{E}\xi(F_{t_0}(X),Y)f(X) = 0 \quad \textit{for all } f \in \mathscr{F},$$

then $C(F_{t_0}) = \inf_{F \in \mathrm{lin}(\mathscr{F})} C(F)$.

Proof Fix $t_0 \geq 0$ and assume that $\mathbb{E}\xi(F_{t_0}(X),Y)f(X) = 0$ for all $f \in \mathscr{F}$. By linearity, $\mathbb{E}\xi(F_{t_0}(X),Y)G(X) = 0$ for all $G \in \mathrm{lin}(\mathscr{F})$. Let $\varepsilon > 0$ and let $F^\star_\varepsilon \in \mathrm{lin}(\mathscr{F})$ be such that

$$\inf_{F \in \mathrm{lin}(\mathscr{F})} C(F) \geq C(F^\star_\varepsilon) - \varepsilon.$$

By the convexity inequality (1) of the Main Document,

$$C(F^\star_\varepsilon) \geq C(F_0) + \mathbb{E}\xi(F_{t_0}(X),Y)(F^\star_\varepsilon(X) - F_{t_0}(X)) = C(F_{t_0}).$$

Thus,

$$\inf_{F \in \mathrm{lin}(\mathscr{F})} C(F) \geq C(F_{t_0}) - \varepsilon.$$

Since ε is arbitrary, the result follows. □

Lemma 3 *Assume that Assumptions* $\mathbf{A_1}$ *and* $\mathbf{A_2}$ *are satisfied. Then, for all* $F \in L^2(\mu_X)$,

$$\|F\|_{\mu_X} \leq \frac{2}{\alpha}\big(\mathbb{E}\xi(0,Y)^2\big)^{1/2} + \sqrt{\frac{2C(F)}{\alpha}}.$$

Proof By inequality (2) of the Main Document and the Cauchy-Schwarz inequality,

$$\begin{aligned} C(F) &\geq C(0) + \mathbb{E}\xi(0,Y)F(X) + \frac{\alpha}{2}\|F\|^2_{\mu_X} \\ &\geq C(0) - \big(\mathbb{E}\xi(0,Y)^2\big)^{1/2}\|F\|_{\mu_X} + \frac{\alpha}{2}\|F\|^2_{\mu_X}. \end{aligned}$$

Let $\kappa = (\mathbb{E}\xi(0,Y)^2)^{1/2}$. Since $C(0) \geq 0$,

$$C(F) + \kappa\|F\|_{\mu_X} - \frac{\alpha}{2}\|F\|^2_{\mu_X} \geq 0.$$

Therefore,

$$\|F\|_{\mu_X} \leq \frac{\kappa + \sqrt{\kappa^2 + 2\alpha C(F)}}{\alpha} \leq \frac{2\kappa}{\alpha} + \sqrt{\frac{2C(F)}{\alpha}}.$$

□

Lemma 4 *Let the event* S *be defined by*

$$S = \big\{\forall j = 1, \ldots, N : P_n(A^n_j) \geq P(A^n_j)/2\big\}.$$

If $\frac{\log N}{n v_n} \to 0$*, then* $\lim_{n\to\infty} \mathbb{P}(S^c) = 0$*.*

Proof We have

$$\begin{aligned}
\mathbb{P}(S^c) &= \mathbb{P}\big(\exists j \leq N : P_n(A_j^n) < P(A_j^n)/2\big)\\
&= \mathbb{P}\big(\exists j \leq N : P_n(A_j^n) - P(A_j^n) < -P(A_j^n)/2\big)\\
&= \mathbb{P}\Big(\exists j \leq N : \frac{P(A_j^n) - P_n(A_j^n)}{\sqrt{P(A_j^n)}} > \sqrt{P(A_j^n)}/2\Big)\\
&\leq \mathbb{P}\Big(\max_{1\leq j\leq N} \frac{P(A_j^n) - P_n(A_j^n)}{\sqrt{P(A_j^n)}} > \sqrt{v_n \inf_{\mathscr{X}} g}/2\Big)\\
&\leq c_1 N e^{-n v_n \inf_{\mathscr{X}} g/c_2},
\end{aligned}$$

where c_1 and c_2 are positive constants. In the last inequality, we used a Vapnik-Chervonenkis inequality [Vapnik, 1988] for relative deviations. □

In the sequel, we let

$$T = \Big\{\beta \in \mathbb{R}^N : \sum_{j=1}^N \beta_j^2 \leq \frac{2\bar{\phi}}{\inf_{\mathscr{X}} g} \cdot \frac{1}{v_n \gamma_n}\Big\},$$

where $\bar{\phi} = \sup_{\mathscr{Y}} \phi(0, y) < \infty$. We recall that $A^n(x) := A_j^n$ whenever $x \in A_j^n$.

Lemma 5 *Assume that* $\mathrm{diam}(A^n(X)) \to 0$ *in probability and that* $\gamma_n \to 0$ *as* $n \to \infty$*. For all* $\varepsilon > 0$ *and all n large enough, there exists* $(\beta_1^\varepsilon, \ldots, \beta_N^\varepsilon) \in T$ *such that*

$$\Big\|F^\star - \sum_{j=1}^N \beta_j^\varepsilon \mathbb{1}_{A_j^n}\Big\|_P \leq \varepsilon.$$

Proof Let K be a bounded and uniformly continuous function on $\mathbb{R}^d$, with $\int K \mathrm{d}\lambda = 1$. Let, for $p > 0$,

$$K_p(x) = p^d K\Big(\frac{x}{p}\Big), \quad x \in \mathbb{R}^d.$$

With a slight abuse of notation, we consider $F^\star$ as a function defined on the whole space $\mathbb{R}^d$ (instead of $\mathscr{X}$) by implicitly assuming that $F^\star = 0$ on $\mathscr{X}^c$. We also define $F_p^\star = F^\star \star K_p$, i.e.,

$$F_p^\star(x) = \int_{\mathbb{R}^d} K_p(z) F^\star(x - z)\mathrm{d}z, \quad x \in \mathbb{R}^d.$$

Let $(L^2(\lambda), \|\cdot\|_\lambda)$ be the vector space of all real-valued square integrable functions on $\mathbb{R}^d$. For all p large enough, we have

$$\|F_p^\star - F^\star\|_\lambda \leq \frac{\varepsilon}{2\sqrt{\sup_{\mathscr{X}} g}}$$

[see, e.g., Wheeden and Zygmund, 1977, Theorem 9.6]. Therefore, for all p large enough,

$$\|F_p^\star - F^\star\|_P \leq \varepsilon/2. \tag{12}$$

In addition, $F_p^\star$ is uniformly continuous on $\mathscr{X}$ [Wheeden and Zygmund, 1977, Theorem 9.4]. Thus, there exists $\eta = \eta(\varepsilon, p) > 0$ such that, for all $(x, x') \in \mathscr{X}^2$ with $\|x - x'\| \leq \eta$,

$$|F_p^\star(x) - F_p^\star(x')| \leq \varepsilon/\sqrt{8}.$$

For each $j \in \{1, \ldots, N\}$, choose an arbitrary $a_j^n \in A_j^n$ and set $G_p^\star = \sum_{j=1}^N F_p^\star(a_j^n)\mathbb{1}_{A_j^n}$. Then

$$\begin{aligned}
\|G_p^\star - F_p^\star\|_P^2 &= \sum_{j=1}^N \mathbb{E}(G_p^\star(X) - F_p^\star(X))^2 \mathbb{1}_{[X \in A_j^n, \mathrm{diam}(A^n(X)) \leq \eta]} \\
&\quad + \sum_{j=1}^N \mathbb{E}(G_p^\star(X) - F_p^\star(X))^2 \mathbb{1}_{[X \in A_j^n, \mathrm{diam}(A^n(X)) > \eta]} \\
&= \sum_{j=1}^N \mathbb{E}(F_p^\star(a_j^n) - F_p^\star(X))^2 \mathbb{1}_{[X \in A_j^n, \mathrm{diam}(A^n(X)) \leq \eta]} \\
&\quad + \sum_{j=1}^N \mathbb{E}(G_p^\star(X) - F_p^\star(X))^2 \mathbb{1}_{[X \in A_j^n, \mathrm{diam}(A^n(X)) > \eta]} \\
&\leq \frac{\varepsilon^2}{8} \sum_{j=1}^N \mathbb{P}(X \in A_j^n) + 4 \sup_{\mathscr{X}} (F^\star)^2 \sum_{j=1}^N \mathbb{P}(X \in A_j^n, \mathrm{diam}(A^n(X)) > \eta) \\
&\qquad \text{(since } \sup_{\mathscr{X}} |F_p^\star| \leq \sup_{\mathscr{X}} |F^\star| < \infty \text{ and } \sup_{\mathscr{X}} |G_p^\star| \leq \sup_{\mathscr{X}} |G^\star| < \infty) \\
&\leq \frac{\varepsilon^2}{8} + 4 \sup_{\mathscr{X}} (F^\star)^2 \mathbb{P}(\mathrm{diam}(A^n(X)) > \eta),
\end{aligned}$$

because the $(A_j^n)_{1 \leq j \leq N}$ form a partition of $\mathscr{X}$. Since $\mathrm{diam}(A^n(X)) \to 0$ in probability, we see that for all n large enough (depending on ε and p),

$$\|G_p^\star - F_p^\star\|_P \leq \varepsilon/2.$$

Letting $\beta_j^\varepsilon = F_p^\star(a_j^n)$, $1 \leq j \leq N$, and combining this inequality and (12), we conclude that for every fixed $\varepsilon > 0$ and all n large enough, there exists $(\beta_1^\varepsilon, \ldots, \beta_N^\varepsilon) \in \mathbb{R}^N$ such that

$$\Big\|F^\star - \sum_{j=1}^N \beta_j^\varepsilon \mathbb{1}_{A_j^n}\Big\|_P \leq \varepsilon.$$

To complete the proof, it remains to show that $(\beta_1^\varepsilon, \ldots, \beta_N^\varepsilon) \in T$. Observe that

$$\sum_{j=1}^{N}(\beta_j^{\varepsilon})^2 \leq \sup_{\mathscr{X}}(F^\star)^2 N.$$

The right-hand side is bounded by $\frac{2\bar{\phi}}{\inf_{\mathscr{X}} g}\cdot\frac{1}{\nu_n\gamma_n}$ for all n large enough. To see this, just note that

$$N\nu_n \leq \sum_{j=1}^{N}\lambda(A_j^n) = \lambda(\mathscr{X}) < \infty.$$

Therefore, $N\nu_n\gamma_n \leq \lambda(\mathscr{X})\gamma_n \to 0$ as $n \to \infty$. This concludes the proof of the lemma. □

Lemma 6 *For* $\beta \in \mathbb{R}^N$, *let* $F_\beta = \sum_{j=1}^{N}\beta_j \mathbb{1}_{A_j^n}$. *Assume that Assumption* $\mathbf{A_4}$ *is satisfied. If*

$$\frac{1}{\sqrt{n\nu_n\gamma_n}}\zeta\left(\sqrt{\frac{2\bar{\phi}}{\nu_n\gamma_n\inf_{\mathscr{X}} g}}\right) \to 0,$$

then

$$\lim_{n\to\infty}\mathbb{E}\sup_{\beta\in T}|A_n(F_\beta) - A(F_\beta)| = 0.$$

Proof Let

$$s_n = \sqrt{\frac{2\bar{\phi}}{\nu_n\gamma_n\inf_{\mathscr{X}} g}},$$

and let $\|\beta\|_\infty = \max_{1\leq j\leq N}|\beta_j|$ be the supremum norm of $\beta = (\beta_1,\ldots,\beta_N) \in \mathbb{R}^N$. By definition of T, we have, for all $\beta \in T$,

$$\sup_{\mathscr{X}}|F_\beta| = \sup_{\mathscr{X}}\Big|\sum_{j=1}^{N}\beta_j\mathbb{1}_{A_j^n}\Big| \leq \|\beta\|_\infty \leq s_n.$$

In addition, according to Assumption $\mathbf{A_4}$, we may write, for β_1 and $\beta_2 \in T$,

$$|\phi(F_{\beta_1}(x),y) - \phi(F_{\beta_2}(x),y)| \leq \zeta(s_n)|F_{\beta_1}(x) - F_{\beta_2}(x)| \leq \zeta(s_n)\|\beta_1 - \beta_2\|_\infty.$$

This shows that the process

$$\Big(\frac{A_n(F_\beta) - A(F_\beta)}{\zeta(s_n)}\Big)_{\beta\in T}$$

is subgaussian [e.g., van Handel, 2016, Chapter 5] for the distance $d(\beta_1,\beta_2) = \frac{1}{\sqrt{n}}\|\beta_1 - \beta_2\|_\infty$. Now, let $N(T,d,\varepsilon)$ denote the ε-covering number of T for the distance d. Then, by Dudley's inequality [van Handel, 2016, Corollary 5.25], one has

$$\begin{aligned}\mathbb{E}\sup_{\beta\in T}(A_n(F_\beta)-A(F_\beta)) &\leq 12\zeta(s_n)\int_0^\infty \sqrt{\log\left(N(T,\frac{1}{\sqrt{n}}\|\cdot\|_\infty,\varepsilon)\right)}\mathrm{d}\varepsilon\\ &= 12\zeta(s_n)\cdot\frac{1}{\sqrt{n}}\int_0^\infty\sqrt{\log(N(T,\|\cdot\|_\infty,\varepsilon))}\mathrm{d}\varepsilon.\end{aligned}$$

Let $B_2(0,1)$ denote the unit Euclidean ball in $(\mathbb{R}^N,\|\cdot\|_2)$. Since $T=s_nB_2(0,1)$, we see that

$$\mathbb{E}\sup_{\beta\in T}(A_n(F_\beta)-A(F_\beta))\leq 12\zeta(s_n)\cdot\frac{s_n}{\sqrt{n}}\int_0^\infty\sqrt{\log(B_2(0,1),\|\cdot\|_\infty,\varepsilon)}\mathrm{d}\varepsilon.$$

But $\|\cdot\|_2\leq\sqrt{N}\|\cdot\|_\infty$, and so

$$\begin{aligned}\mathbb{E}\sup_{\beta\in T}(A_n(F_\beta)-A(F_\beta)) &\leq 12\zeta(s_n)\cdot\frac{s_n}{\sqrt{n}}\int_0^\infty\sqrt{\log\left(B_2(0,1),\frac{1}{\sqrt{N}}\|\cdot\|_2,\varepsilon\right)}\mathrm{d}\varepsilon\\ &= 12\zeta(s_n)\cdot\frac{s_n}{\sqrt{n}}\cdot\frac{1}{\sqrt{N}}\int_0^\infty\sqrt{\log(3/\varepsilon)^N}\mathrm{d}\varepsilon\\ &= 12\,\frac{s_n\zeta(s_n)}{\sqrt{n}}\int_0^\infty\sqrt{\log(3/\varepsilon)}\mathrm{d}\varepsilon.\end{aligned}$$

In the last equality, we used the fact that $N(B_2(0,1),\|\cdot\|_2,\varepsilon)$ equals 1 for $\varepsilon\geq 1$ and is not larger than $(3/\varepsilon)^N$ for $\varepsilon<1$ [e.g., van Handel, 2016, Chapter 5]. The same conclusion holds for $\mathbb{E}\sup_{\beta\in T}(A(F_\beta)-A_n(F_\beta))$, and this proves the result. □

References

R. van Handel. *Probability in High Dimension*. APC 550 Lecture Notes, Princeton University, 2016.

V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1988.

R.L. Wheeden and A. Zygmund. *Measure and Integral*. Marcel Dekker, New York, 1977.