Submitted to the Annals of Statistics

SUPPLEMENTARY MATERIALS FOR: CONSISTENCY OF RANDOM FORESTS

By Erwan Scornet

Sorbonne Universités, UPMC Univ Paris 06, F-75005, Paris, France

and

By Gérard Biau

Sorbonne Universités, UPMC Univ Paris 06, F-75005, Paris, France

and

By Jean-Philippe Vert

MINES ParisTech, PSL-Research University, CBIO-Centre for Computational Biology, F-77300, Fontainebleau, France & Institut Curie, Paris, F-75248, France & U900, INSERM, Paris, F-75248, France

1. Proof of Lemma 1.

Technical Lemma 1. *Assume that* **(H1)** *is satisfied and that* $L^\star \equiv 0$ *for all cuts in some given cell* A*. Then the regression function* m *is constant on* A.

Proof of Technical Lemma 1. We start by proving the result in dimension $p = 1$. Letting $A = [a, b]$ $(0 \le a < b \le 1)$, and recalling that $Y = m(\mathbf{X}) + \varepsilon$, one has

$$\begin{aligned} L^\star(1, z) &= \mathbb{V}\left[Y|\mathbf{X} \in A\right] - \mathbb{P}\left[a \le \mathbf{X} \le z \,|\, \mathbf{X} \in A\right] \mathbb{V}\left[Y|a \le \mathbf{X} \le z\right] \\ &\quad - \mathbb{P}\left[z \le \mathbf{X} \le b \,|\, \mathbf{X} \in A\right] \mathbb{V}\left[Y|z < \mathbf{X} \le b\right] \\ &= -\frac{1}{(b-a)^2}\left(\int_a^b m(t)\mathrm{d}t\right)^2 + \frac{1}{(b-a)(z-a)}\left(\int_a^z m(t)\mathrm{d}t\right)^2 \\ &\quad + \frac{1}{(b-a)(b-z)}\left(\int_z^b m(t)\mathrm{d}t\right)^2. \end{aligned}$$

Let $C = \int_a^b m(t)\mathrm{d}t$ and $M(z) = \int_a^z m(t)\mathrm{d}t$. Simple calculations show that

$$L^\star(1, z) = \frac{1}{(z-a)(b-z)}\left(M(z) - C\frac{z-a}{b-a}\right)^2.$$

Therefore, since $L^\star \equiv 0$ on $\mathcal{C}_A$ by assumption, we obtain

$$M(z) = C\,\frac{z-a}{b-a}\,.$$

This proves that $M(z)$ is linear in z, and that m is therefore constant on $[a,b]$.

Let us now examine the general multivariate case, where $A = \Pi_{j=1}^p [a_j, b_j] \subset [0,1]^p$. From the univariate analysis, we know that, for all $1 \leq j \leq p$, there exists a constant C_j such that

$$\int_{a_1}^{b_1} \ldots \int_{a_p}^{b_p} m(\mathbf{x}) \mathrm{d}x_1 \ldots \mathrm{d}x_{j-1} \mathrm{d}x_{j+1} \ldots \mathrm{d}x_p = C_j\,.$$

Since m is additive this implies that, for all j and all x_j,

$$m_j(x_j) = C_j - \int_{a_1}^{b_1} \ldots \int_{a_p}^{b_p} \sum_{\ell \neq j} m_\ell(x_\ell) \mathrm{d}x_1 \ldots \mathrm{d}x_{j-1} \mathrm{d}x_{j+1} \ldots \mathrm{d}x_p\,,$$

which does not depend upon x_i. This shows that m is constant on A. $\square$

Proof of Lemma 1. Take $\xi > 0$ and fix $\mathbf{x} \in [0,1]^p$. Let θ be a realization of the random variable Θ. Since m is uniformly continuous, the result is clear if $\mathrm{diam}(A_k^\star(\mathbf{x},\theta))$ tends to zero as k tends to infinity. Thus, in the sequel, it is assumed that $\mathrm{diam}(A_k^\star(\mathbf{x},\theta))$ does not tend to zero. In that case, since $(A_k^\star(\mathbf{x},\theta))_k$ is a decreasing sequence of compact sets, there exist $\mathbf{a}_\infty(\mathbf{x},\theta) = (\mathbf{a}_\infty^{(1)}(\mathbf{x},\theta), \ldots, \mathbf{a}_\infty^{(p)}(\mathbf{x},\theta)) \in [0,1]^p$ and $\mathbf{b}_\infty(\mathbf{x},\theta) = (\mathbf{b}_\infty^{(1)}(\mathbf{x},\theta), \ldots, \mathbf{b}_\infty^{(p)}(\mathbf{x},\theta)) \in [0,1]^p$ such that

$$\begin{aligned}\bigcap_{k=1}^\infty A_k^\star(\mathbf{x},\theta) &= \prod_{j=1}^p [\mathbf{a}_\infty^{(j)}(\mathbf{x},\theta), \mathbf{b}_\infty^{(j)}(\mathbf{x},\theta)] \\ &\stackrel{\mathrm{def}}{=} A_\infty^\star(\mathbf{x},\theta).\end{aligned}$$

Since $\mathrm{diam}(A_k^\star(\mathbf{x},\theta))$ does not tend to zero, there exists an index j' such that $\mathbf{a}_\infty^{(j')}(\mathbf{x},\theta) < \mathbf{b}_\infty^{(j')}(\mathbf{x},\theta)$ (i.e., the cell $A_\infty^\star(\mathbf{x},\theta)$ is not reduced to one point). Let $A_k^\star(\mathbf{x},\theta) \stackrel{\mathrm{def}}{=} \prod_{j=1}^p [\mathbf{a}_k^{(j)}(\mathbf{x},\theta), \mathbf{b}_k^{(j)}(\mathbf{x},\theta)]$ be the cell containing $\mathbf{x}$ at level k. If the criterion $L^\star$ is identically zero for all cuts in $A_\infty^\star(\mathbf{x},\theta)$ then m is constant on $A_\infty^\star(\mathbf{x},\theta)$ according to Lemma 1. This implies that $\Delta(m, A_\infty^\star(\mathbf{x},\theta)) = 0$. Thus, in that case, since m is uniformly continuous,

$$\lim_{k\to\infty} \Delta(m, A_k^\star(\mathbf{x},\theta)) = \Delta(m, A_\infty^\star(\mathbf{x},\theta)) = 0.$$

Let us now show by contradiction that $L^\star$ is almost surely necessarily null on the cuts of $A^\star_\infty(\mathbf{x},\theta)$. In the rest of the proof, for all $k \in \mathbb{N}^\star$, we let $L^\star_k$ be the criterion $L^\star$ used in the cell $A^\star_k(\mathbf{x},\theta)$, that is

$$\begin{aligned} L^\star_k(d) = \mathbb{V}[Y|\mathbf{X} \in A^\star_k(\mathbf{x},\theta)] & \\ &- \mathbb{P}[\mathbf{X}^{(j)} < z \,|\, \mathbf{X} \in A^\star_k(\mathbf{x},\theta)]\ \mathbb{V}[Y|\mathbf{X}^{(j)} < z, \mathbf{X} \in A^\star_k(\mathbf{x},\theta)] \\ &- \mathbb{P}[\mathbf{X}^{(j)} \geq z \,|\, \mathbf{X} \in A^\star_k(\mathbf{x},\theta)]\ \mathbb{V}[Y|\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A^\star_k(\mathbf{x},\theta)], \end{aligned}$$

for all $d = (j,z) \in \mathcal{C}_{A^\star_k(\mathbf{x},\theta)}$. If $L^\star_\infty$ is not identically zero, then there exists a cut $d_\infty(\mathbf{x},\theta)$ in $\mathcal{C}_{A^\star_\infty(\mathbf{x},\theta)}$ such that $L^\star(d_\infty(\mathbf{x},\theta)) = c > 0$. Fix $\xi > 0$. By the uniform continuity of m, there exists $\delta_1 > 0$ such that

$$\sup_{\|\mathbf{w}-\mathbf{w}'\|_\infty \leq \delta_1} |m(\mathbf{w}) - m(\mathbf{w}')| \leq \xi\,.$$

Since $A^\star_k(\mathbf{x},\theta) \downarrow A^\star_\infty(\mathbf{x},\theta)$, there exists k_0 such that, for all $k \geq k_0$,

$$\max\left(\|\mathbf{a}_k(\mathbf{x},\theta) - \mathbf{a}_\infty(\mathbf{x},\theta)\|_\infty, \|\mathbf{b}_k(\mathbf{x},\theta) - \mathbf{b}_\infty(\mathbf{x},\theta)\|_\infty\right) \leq \delta_1\,. \tag{1}$$

Observe that, for all $k \in \mathbb{N}^\star$, $\mathbb{V}[Y|\mathbf{X} \in A^\star_{k+1}(\mathbf{x},\theta)] < \mathbb{V}[Y|\mathbf{X} \in A^\star_k(\mathbf{x},\theta)]$. Thus,

$$\underline{L}^\star_k := \sup_{\substack{d \in \mathcal{C}_{A_k(\mathbf{x},\theta)} \\ d^{(1)} \in \mathcal{M}_{\text{try}}}} L^\star_k(d) \leq \xi\,. \tag{2}$$

From inequality (1), we deduce that

$$\Big|\mathbb{E}[m(\mathbf{X})|\mathbf{X} \in A^\star_k(\mathbf{x},\theta)] - \mathbb{E}[m(\mathbf{X})|\mathbf{X} \in A^\star_\infty(\mathbf{x},\theta)]\Big| \leq \xi.$$

Consequently, there exists a constant $C > 0$ such that, for all $k \geq k_0$ and all cuts $d \in \mathcal{C}_{A^\star_\infty(\mathbf{x},\theta)}$,

$$|L^\star_k(d) - L^\star_\infty(d)| \leq C\xi^2. \tag{3}$$

Let $k_1 \geq k_0$ be the first level after k_0 at which the direction $d^{(1)}_\infty(\mathbf{x},\theta)$ is amongst the m_{try} selected coordinates. Almost surely, $k_1 < \infty$. Thus, by the definition of $d_\infty(\mathbf{x},\theta)$ and inequality (3),

$$c - C\xi^2 \leq L^\star_\infty(d_\infty(\mathbf{x},\theta)) - C\xi^2 \leq L^\star_k(d_\infty(\mathbf{x},\theta)),$$

which implies that $c - C\xi^2 \leq \underline{L}^\star_k$. Hence, using inequality (2), we have

$$c - C\xi^2 \leq \underline{L}^\star_k \leq \xi,$$

which is absurd, since $c > 0$ is fixed and ξ is arbitrarily small. Thus, by Lemma 1, m is constant on $A^\star_\infty(\mathbf{x},\theta)$. This implies that $\Delta(m, A^\star_k(\mathbf{x},\Theta)) \to 0$ as $k \to \infty$.

2. Proof of Lemma 2. We start by proving Lemma 2 in the case $k = 1$, i.e., when the first cut is performed at the root of a tree. Since in that case $L_{n,1}(\mathbf{x}, \cdot)$ does not depend on $\mathbf{x}$, we simply write $L_{n,1}(\cdot)$ instead of $L_{n,1}(\mathbf{x}, \cdot)$.

PROOF OF LEMMA 2 IN THE CASE $k = 1$. Fix $\alpha, \rho > 0$. Observe that if two cuts d_1, d_2 satisfy $\|d_1 - d_2\|_\infty < 1$, then the cut directions are the same, i.e., $d_1^{(1)} = d_2^{(1)}$. Using this fact and symmetry arguments, we just need to prove Lemma 2 when the cuts are performed along the first dimension. In other words, we only need to prove that

$$\lim_{n\to\infty} \mathbb{P}\left[\sup_{|x_1 - x_2|\leq\delta} |L_{n,1}(1, x_1) - L_{n,1}(1, x_2)| > \alpha\right] \leq \rho/p. \tag{4}$$

Preliminary results. Letting $Z_i = \max_{1\leq i\leq n} |\varepsilon_i|$, simple calculations show that

$$\mathbb{P}\left[Z_i \geq t\right] = 1 - \exp\Big(n \ln\big(1 - 2\mathbb{P}\left[\varepsilon_1 \geq t\right]\big)\Big).$$

The last probability can be upper bounded by using the following standard inequality on Gaussian tail:

$$\mathbb{P}\left[\varepsilon_1 \geq t\right] \leq \frac{\sigma}{t\sqrt{2\pi}} \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Consequently, there exists a constant $C_\rho > 0$ and $N_1 \in \mathbb{N}^\star$ such that, with probability $1 - \rho$, for all $n > N_1$,

$$\max_{1\leq i\leq n} |\varepsilon_i| \leq C_\rho \sqrt{\log n}\,. \tag{5}$$

Besides, by simple calculations on Gaussian tail, for all $n \in \mathbb{N}^\star$, we have

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i\right| \geq \alpha\right] \leq \frac{\sigma}{\alpha\sqrt{n}} \exp\left(-\frac{\alpha^2 n}{2\sigma^2}\right).$$

Since there are, at most, n^2 sets of the form $\{i : X_i \in [a_n, b_n]\}$ for $0 \leq a_n < b_n \leq 1$, we deduce from the last inequality and the union bound, that there exists $N_2 \in \mathbb{N}^\star$ such that, with probability $1 - \rho$, for all $n > N_2$ and all $0 \leq a_n < b_n \leq 1$ satisfying $N_n([a_n, b_n] \times [0, 1]^{p-1}) > \sqrt{n}$,

$$\Big|\frac{1}{N_n([a_n, b_n] \times [0, 1]^{p-1})} \sum_{\substack{i: X_i \in [a_n, b_n] \\ \times [0,1]^{p-1}}} \varepsilon_i\Big| \leq \alpha. \tag{6}$$

By the Glivenko-Cantelli theorem, there exists $N_3 \in \mathbb{N}^\star$ such that, with probability $1-\rho$, for all $0 \leq a < b \leq 1$, and all $n > N_3$,

$$(b-a-\delta^2)n \leq N_n([a,b]\times[0,1]^{p-1}) \leq (b-a+\delta^2)n. \tag{7}$$

Throughout the proof, we assume to be on the event where assertions (5)-(7) hold, which occurs with probability $1-3\rho$, for all $n > N$, where $N = \max(N_1, N_2, N_3)$.

Take $x_1, x_2 \in [0,1]$ such that $|x_1 - x_2| \leq \delta$ and assume, without loss of generality, that $x_1 < x_2$. In the remainder of the proof, we will need the following quantities (see Figure 1 for an illustration in dimension two):

$$\begin{cases} A_{L,\sqrt{\delta}} = [0,\sqrt{\delta}]\times[0,1]^{p-1} \\ A_{R,\sqrt{\delta}} = [1-\sqrt{\delta},1]\times[0,1]^{p-1} \\ A_{C,\sqrt{\delta}} = [\sqrt{\delta},1-\sqrt{\delta}]\times[0,1]^{p-1}. \end{cases}$$

Similarly, we define

$$\begin{cases} A_{L,1} & = [0,x_1]\times[0,1]^{p-1} \\ A_{R,1} & = [x_1,1]\times[0,1]^{p-1} \\ A_{L,2} & = [0,x_2]\times[0,1]^{p-1} \\ A_{R,2} & = [x_2,1]\times[0,1]^{p-1} \\ A_C & = [x_1,x_2]\times[0,1]^{p-1}. \end{cases}$$

Recall that, for any cell A, $\bar{Y}_A$ is the mean of the Y_i's falling in A and $N_n(A)$ is the number of data points in A. To prove (4), five cases are to be considered, depending upon the positions of x_1 and x_2. We repeatedly use the decomposition

$$L_{n,1}(1,x_1) - L_{n,1}(1,x_2) = J_1 + J_2 + J_3,$$

where

$$\begin{aligned} J_1 &= \frac{1}{n}\sum_{i:\mathbf{X}_i^{(1)}<x_1}(Y_i-\bar{Y}_{A_{L,1}})^2 - \frac{1}{n}\sum_{i:\mathbf{X}_i^{(1)}<x_1}(Y_i-\bar{Y}_{A_{L,2}})^2, \\ J_2 &= \frac{1}{n}\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]}(Y_i-\bar{Y}_{A_{R,1}})^2 - \frac{1}{n}\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]}(Y_i-\bar{Y}_{A_{L,2}})^2, \\ \text{and}\quad J_3 &= \frac{1}{n}\sum_{i:\mathbf{X}_i^{(1)}\geq x_2}(Y_i-\bar{Y}_{A_{R,1}})^2 - \frac{1}{n}\sum_{i:\mathbf{X}_i^{(1)}\geq x_2}(Y_i-\bar{Y}_{A_{R,2}})^2. \end{aligned}$$

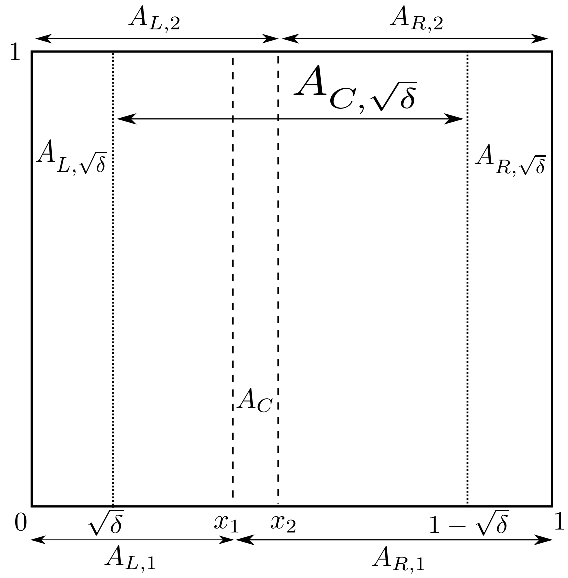

FIGURE 1. *Illustration of the notation in dimension $p = 2$.*

First case. Assume that $x_1, x_2 \in A_{C,\sqrt{\delta}}$. Since $N_n(A_{L,2}) > N_n(A_{L,\sqrt{\delta}}) > \sqrt{n}$ for all $n > N$, we have, according to inequalities (6),

$$|\bar{Y}_{A_{L,2}}| \leq \|m\|_\infty + \alpha \quad \text{and} \quad |\bar{Y}_{A_{R,1}}| \leq \|m\|_\infty + \alpha.$$

Therefore

$$\begin{aligned}
|J_2| &= 2\left|\bar{Y}_{A_{L,2}} - \bar{Y}_{A_{R,1}}\right| \times \frac{1}{n}\left|\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]}\left(Y_i - \frac{\bar{Y}_{A_{L,2}} + \bar{Y}_{A_{R,1}}}{2}\right)\right| \\
&\leq 4(\|m\|_\infty + \alpha)\left(\frac{(\|m\|_\infty + \alpha)N_n(A_C)}{n} + \frac{1}{n}\left|\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]} m(\mathbf{X}_i)\right|\right. \\
&\qquad\qquad \left. + \frac{1}{n}\left|\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]} \varepsilon_i\right|\right)
\end{aligned}$$

$$\leq 4(\|m\|_\infty + \alpha)\Bigg((\delta + \delta^2)(\|m\|_\infty + \alpha) + \|m\|_\infty(\delta+\delta^2) + \frac{1}{n}\Bigg| \sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]} \varepsilon_i \Bigg| \Bigg).$$

If $N_n(A_C) \geq \sqrt{n}$, we obtain

$$\frac{1}{n}\Bigg| \sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]} \varepsilon_i \Bigg| \leq \frac{1}{N_n(A_C)}\Bigg| \sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]} \varepsilon_i \Bigg| \leq \alpha \quad \big(\text{according to (6)}\big)$$

or, if $N_n(A_C) < \sqrt{n}$, we have

$$\frac{1}{n}\Bigg| \sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]} \varepsilon_i \Bigg| \leq \frac{C_\rho \sqrt{\log n}}{\sqrt{n}} \quad \big(\text{according to (5)}\big).$$

Thus, for all n large enough,

$$|J_2| \leq 4(\|m\|_\infty + \alpha)\Big((\delta+\delta^2)(2\|m\|_\infty + \alpha) + \alpha \Big). \tag{8}$$

With respect to J_1, observe that

$$\begin{aligned}
|\bar{Y}_{A_{L,1}} - \bar{Y}_{A_{L,2}}| &= \Bigg| \frac{1}{N_n(A_{L,1})} \sum_{i:\mathbf{X}_i^{(1)}<x_1} Y_i - \frac{1}{N_n(A_{L,2})} \sum_{i:\mathbf{X}_i^{(1)}<x_2} Y_i \Bigg| \\
&\leq \Bigg| \frac{1}{N_n(A_{L,1})} \sum_{i:\mathbf{X}_i^{(1)}<x_1} Y_i - \frac{1}{N_n(A_{L,2})} \sum_{i:\mathbf{X}_i^{(1)}<x_1} Y_i \Bigg| \\
&\quad + \Bigg| \frac{1}{N_n(A_{L,2})} \sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]} Y_i \Bigg| \\
&\leq \Bigg| 1 - \frac{N_n(A_{L,1})}{N_n(A_{L,2})} \Bigg| \times \frac{1}{N_n(A_{L,1})} \times \Bigg| \sum_{i:\mathbf{X}_i^{(1)}<x_1} Y_i \Bigg| \\
&\quad + \frac{1}{N_n(A_{L,2})} \Bigg| \sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]} Y_i \Bigg|.
\end{aligned}$$

Since $N_n(A_{L,2}) - N_n(A_{L,1}) \leq n(\delta + \delta^2)$, we obtain

$$1 - \frac{N_n(A_{L,1})}{N_n(A_{L,2})} \leq \frac{n(\delta+\delta^2)}{N_n(A_{L,2})} \leq \frac{\delta+\delta^2}{\sqrt{\delta}-\delta^2} \leq 4\sqrt{\delta},$$

for all δ small enough, which implies that

$$\begin{aligned}
|\bar{Y}_{A_{L,1}} - \bar{Y}_{A_{L,2}}| &\leq \frac{4\sqrt{\delta}}{N_n(A_{L,1})}\left|\sum_{i:\mathbf{X}_i^{(1)}<x_1} Y_i\right| \\
&\qquad + \frac{N_n(A_{L,1})}{N_n(A_{L,2})} \times \frac{1}{N_n(A_{L,1})}\left|\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]} Y_i\right| \\
&\leq 4\sqrt{\delta}(\|m\|_\infty + \alpha) + \frac{N_n(A_{L,1})}{N_n(A_{L,2})}(\|m\|_\infty\delta + \alpha) \\
&\leq 5(\|m\|_\infty\sqrt{\delta} + \alpha).
\end{aligned}$$

Thus,

$$\begin{aligned}
|J_1| &= \left|\frac{1}{n}\sum_{i:\mathbf{X}_i^{(1)}<x_1}(Y_i - \bar{Y}_{A_{L,1}})^2 - \frac{1}{n}\sum_{i:\mathbf{X}_i^{(1)}<x_1}(Y_i - \bar{Y}_{A_{L,2}})^2\right| \\
&= \left|(\bar{Y}_{A_{L,2}} - \bar{Y}_{A_{L,1}}) \times \frac{2}{n}\sum_{i:\mathbf{X}_i^{(1)}<x_1}\left(Y_i - \frac{\bar{Y}_{A_{L,1}} + \bar{Y}_{A_{L,2}}}{2}\right)\right| \\
&\leq |\bar{Y}_{A_{L,2}} - \bar{Y}_{A_{L,1}}|^2 \\
&\leq 25(\|m\|_\infty\sqrt{\delta} + \alpha)^2. \qquad (9)
\end{aligned}$$

The term J_3 can be bounded with the same arguments.

Finally, by (8) and (9), for all $n > N$, and all δ small enough, we conclude that

$$\begin{aligned}
|L_n(1,x_1) - L_n(1,x_2)| &\leq 4(\|m\|_\infty + \alpha)\Big((\delta + \delta^2)(2\|m\|_\infty + \alpha) + \alpha\Big) \\
&\qquad + 25(\|m\|_\infty\sqrt{\delta} + \alpha)^2 \\
&\leq \alpha.
\end{aligned}$$

Second case. Assume that $x_1, x_2 \in A_{L,\sqrt{\delta}}$. With the same arguments as above, one proves that

$$\begin{aligned}
|J_1| &\leq \max\left(4(\sqrt{\delta} + \delta^2)(\|m\|_\infty + \alpha)^2, \alpha\right), \\
|J_2| &\leq \max(4(\|m\|_\infty + \alpha)(2\delta\|m\|_\infty + 2\alpha), \alpha), \\
|J_3| &\leq 25(\|m\|_\infty\sqrt{\delta} + \alpha)^2.
\end{aligned}$$

Consequently, for all n large enough,

$$|L_n(1, x_1) - L_n(1, x_2)| = J_1 + J_2 + J_3 \leq 3\alpha.$$

The other cases $\{x_1, x_2 \in A_{R,\sqrt{\delta}}\}$, $\{x_1, x_2 \in A_{L,\sqrt{\delta}} \times A_{C,\sqrt{\delta}}\}$, and $\{x_1, x_2 \in A_{C,\sqrt{\delta}} \times A_{R,\sqrt{\delta}}\}$ can be treated in the same way. Details are omitted. $\square$

PROOF OF LEMMA 2. We proceed similarly as in the proof of the case $k = 1$. Here, we establish the result for $k = 2$ and $p = 2$ only. Extensions are easy and left to the reader.

Preliminary results. Fix $\rho > 0$. At first, it should be noted that there exists $N_1 \in \mathbb{N}^\star$ such that, with probability $1 - \rho$, for all $n > N_0$ and all $A_n = [a_n^{(1)}, b_n^{(1)}] \times [a_n^{(2)}, b_n^{(2)}] \subset [0, 1]^2$ satisfying $N_n(A_n) > \sqrt{n}$, we have

$$\Big|\frac{1}{N_n(A_n)} \sum_{i:X_i \in A_n} \varepsilon_i\Big| \leq \alpha, \tag{10}$$

and

$$\frac{1}{N_n(A_n)} \sum_{i:X_i \in A_n} \varepsilon_i^2 \leq \tilde{\sigma}^2, \tag{11}$$

where $\tilde{\sigma}^2$ is a positive constant, depending only on ρ. Inequality (11) is a straightforward consequence of the following inequality (see, e.g., Laurent and Massart, 2000), which is valid for all $n \in \mathbb{N}^\star$:

$$\mathbb{P}\left[\chi^2(n) \geq 5n\right] \leq \exp(-n).$$

Throughout the proof, we assume to be on the event where assertions (5), (7), (10)-(11) hold, which occurs with probability $1 - 3\rho$, for all n large enough. We also assume that $d_1 = (1, x_1)$ and $d_2 = (2, x_2)$ (see Figure 2). The other cases can be treated similarly.

Main argument. Let $d_1' = (1, x_1')$ and $d_2' = (2, x_2')$ be such that $|x_1 - x_1'| < \delta$ and $|x_2 - x_2'| < \delta$. Then the CART-split criterion $L_{n,2}$ writes

$$\begin{aligned}
L_n(d_1, d_2) = &\frac{1}{N_n(A_{R,1})} \sum_i (Y_i - \bar{Y}_{A_{R,1}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1} \\
&- \frac{1}{N_n(A_{R,1})} \sum_{i:\mathbf{X}_i^{(2)} > x_2} (Y_i - \bar{Y}_{A_{H,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1} \\
&- \frac{1}{N_n(A_{R,1})} \sum_{i:\mathbf{X}_i^{(2)} \leq x_2} (Y_i - \bar{Y}_{A_{B,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1}.
\end{aligned}$$

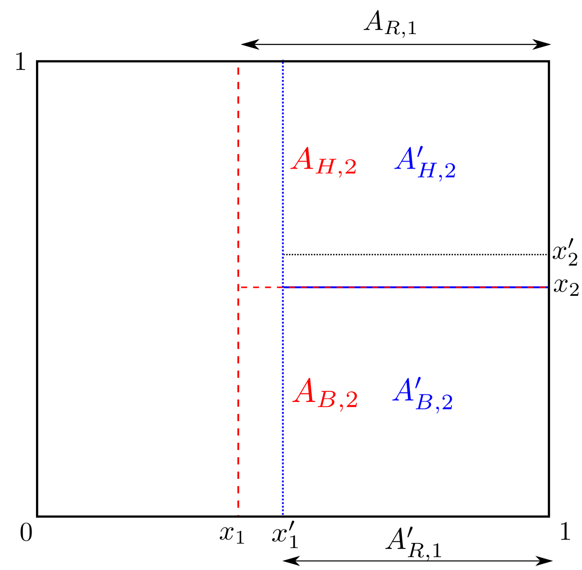

FIGURE 2. *An example of cells in dimension $p = 2$.*

Clearly,

$$L_n(d_1, d_2) - L_n(d'_1, d'_2) = L_n(d_1, d_2) - L_n(d'_1, d_2) + L_n(d'_1, d_2) - L_n(d'_1, d'_2).$$

We have (Figure 2):

$$\begin{aligned} L_n(d_1, d_2) - L_n(d'_1, d_2) &= \Bigg[\frac{1}{N_n(A_{R,1})} \sum_{i:\mathbf{X}_i^{(2)} > x_2} (Y_i - \bar{Y}_{A_{H,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1} \\ &\qquad - \frac{1}{N_n(A'_{R,1})} \sum_{i:\mathbf{X}_i^{(2)} > x_2} (Y_i - \bar{Y}_{A'_{H,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x'_1} \Bigg] \\ &\quad + \Bigg[\frac{1}{N_n(A_{R,1})} \sum_{i:\mathbf{X}_i^{(2)} \leq x_2} (Y_i - \bar{Y}_{A_{B,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1} \\ &\qquad - \frac{1}{N_n(A'_{R,1})} \sum_{i:\mathbf{X}_i^{(2)} \leq x_2} (Y_i - \bar{Y}_{A'_{B,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x'_1} \Bigg] \\ &\stackrel{\text{def}}{=} A_1 + B_1. \end{aligned}$$

The term A_1 can be rewritten as $A_1 = A_{1,1} + A_{1,2} + A_{1,3}$, where

$$\begin{aligned} A_{1,1} &= \frac{1}{N_n(A_{R,1})} \sum_{i:\mathbf{X}_i^{(2)} > x_2} (Y_i - \bar{Y}_{A_{H,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1'} \\ &\quad - \frac{1}{N_n(A_{R,1})} \sum_{i:\mathbf{X}_i^{(2)} > x_2} (Y_i - \bar{Y}_{A'_{H,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1'}, \\ A_{1,2} &= \frac{1}{N_n(A_{R,1})} \sum_{i:\mathbf{X}_i^{(2)} > x_2} (Y_i - \bar{Y}_{A'_{H,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1'} \\ &\quad - \frac{1}{N_n(A'_{R,1})} \sum_{i:\mathbf{X}_i^{(2)} > x_2} (Y_i - \bar{Y}_{A'_{H,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1'}, \\ \text{and} \quad A_{1,3} &= \frac{1}{N_n(A_{R,1})} \sum_{i:\mathbf{X}_i^{(2)} > x_2} (Y_i - \bar{Y}_{A_{H,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} \in [x_1, x_1']}. \end{aligned}$$

Easy calculations show that

$$A_{1,1} = \frac{N_n(A'_{H,2})}{N_n(A_{R,1})} (\bar{Y}_{A'_{H,2}} - \bar{Y}_{A_{H,2}})^2,$$

which implies, with the same arguments as in the proof for $k = 1$, that $A_{1,1} \to 0$ as $n \to \infty$. With respect to $A_{1,2}$ and $A_{1,3}$, we write

$$\max(A_{1,2}, A_{1,3}) \le \max(C_\rho \frac{\log n}{\sqrt{n}}, 2(\tilde{\sigma}^2 + 4\|m\|_\infty^2 + \alpha^2) \frac{\sqrt{\delta}}{\xi}).$$

Thus, $A_{1,2} \to 0$ and $A_{1,3} \to 0$ as $n \to \infty$. Collecting bounds, we conclude that $A_1 \to 0$. One proves with similar arguments that $B_1 \to 0$ and, consequently, that $L_n(d_1', d_2) - L_n(d_1', d_2') \to 0$. □

3. Proof of Lemma 3. We prove by induction that, for all k, with probability $1 - \rho$, for all $\xi > 0$ and all n large enough,

$$d_\infty(\hat{\mathbf{d}}_{k,n}(\mathbf{X}, \Theta), \mathcal{A}_k^\star(\mathbf{X}, \Theta)) \le \xi.$$

Call this property H_k. Fix $k > 1$ and assume that H_{k-1} is true. For all $\mathbf{d}_{k-1} \in \mathcal{A}_{k-1}(\mathbf{X})$, let

$$\hat{d}_{k,n}(\mathbf{d}_{k-1}) \in \arg\min_{d_k} L_n(\mathbf{X}, \mathbf{d}_{k-1}, d_k),$$

and

$$d_k^\star(\mathbf{d}_{k-1}) \in \arg\min_{d_k} L^\star(\mathbf{X}, \mathbf{d}_{k-1}, d_k),$$

where the minimum is evaluated, as usual, over $\{d_k \in \mathcal{C}_{A(\mathbf{X},\mathbf{d}_{k-1})} : d_k^{(1)} \in \mathcal{M}_{\mathrm{try}}\}$. Fix $\rho > 0$. In the rest of the proof, we assume Θ to be fixed and we omit the dependence on Θ.

Preliminary result. We momentarily consider $\mathbf{x} \in [0,1]^d$. Note that, for all $\mathbf{d}_{k-1}$,

$$\begin{aligned}
&L_n(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1}))\\
&\quad \leq L_n(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1}))\\
&\quad\quad \text{(by definition of } d_k^\star(\mathbf{d}_{k-1}))\\
&\quad \leq L_n(\mathbf{x}, \mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1})) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1}))\\
&\quad\quad \text{(by definition of } \hat{d}_{k,n}(\mathbf{d}_{k-1})).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\Big|L_n(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1}))\Big|\\
&\quad \leq \max\Big(\Big|L_n(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1}))\Big|,\\
&\qquad\qquad\qquad \Big|L_n(\mathbf{x}, \mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1})) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1}))\Big|\Big)\\
&\quad \leq \sup_{d_k} \left|L_n(\mathbf{x}, \mathbf{d}_{k-1}, d_k) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, d_k)\right|.
\end{aligned}$$

Moreover,

$$\begin{aligned}
&|L^\star(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1}))|\\
&\quad \leq |L^\star(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L_n(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1}))|\\
&\qquad + |L_n(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1}))|\\
&\quad \leq 2\sup_{d_k} |L_n(\mathbf{x}, \mathbf{d}_{k-1}, d_k) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, d_k)|\\
&\quad = 2\sup_{d_k} |L_n(\mathbf{x}, \mathbf{d}_k) - L^\star(\mathbf{x}, \mathbf{d}_k)|. \qquad (12)
\end{aligned}$$

Let $\bar{\mathcal{A}}_k^\xi(\mathbf{x}) = \{\mathbf{d}_k : \mathbf{d}_{k-1} \in \mathcal{A}_{k-1}^\xi(\mathbf{x})\}$. So, taking the supremum on both sides of (12) leads to

$$\begin{aligned}
&\sup_{\mathbf{d}_{k-1} \in \mathcal{A}_{k-1}^\xi(\mathbf{x})} |L^\star(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1}))|\\
&\quad \leq 2 \sup_{\mathbf{d}_k \in \bar{\mathcal{A}}_k^\xi(\mathbf{x})} |L_n(\mathbf{x}, \mathbf{d}_k) - L^\star(\mathbf{x}, \mathbf{d}_k)|. \qquad (13)
\end{aligned}$$

By Lemma 3, for all $\xi' > 0$, one can find $\delta > 0$ such that, for all n large enough,

$$\mathbb{P}\left[\sup_{\mathbf{x}\in[0,1]^d}\sup_{\substack{\|\mathbf{d}_k-\mathbf{d}'_k\|_\infty\le\delta\\ \mathbf{d}_k,\mathbf{d}'_k\in\bar{\mathcal{A}}_k^{\xi}(\mathbf{x})}}|L_n(\mathbf{x},\mathbf{d}_k)-L_n(\mathbf{x},\mathbf{d}'_k)|\le\xi'\right]\ge 1-\rho. \tag{14}$$

Now, let $\mathcal{G}$ be a regular grid of $[0,1]^d$ whose grid step equal to $\xi/2$. Note that, for all $\mathbf{x}\in\mathcal{G}$, $\bar{\mathcal{A}}_k^{\xi}(\mathbf{x})$ is compact. Thus, for all $\mathbf{x}\in\mathcal{G}$, there exists a finite subset $\mathcal{C}_{\delta,\mathbf{x}}=\{c_{j,\mathbf{x}}:1\le j\le p\}$ such that, for all $\mathbf{d}_k\in\bar{\mathcal{A}}_k^{\xi}(\mathbf{x})$, $d_\infty(\mathbf{d}_k,\mathcal{C}_{\delta,\mathbf{x}})\le\delta$. Set $\xi'>0$. Observe that, since the subset $\cup_{\mathbf{x}\in\mathcal{G}}\mathcal{C}_{\delta,\mathbf{x}}$ is finite, one has, for all n large enough,

$$\sup_{\mathbf{x}\in\mathcal{G}}\sup_{c_{j,\mathbf{x}}\in\mathcal{C}_{\delta,\mathbf{x}}}|L_n(\mathbf{x},c_{j,\mathbf{x}})-L^\star(\mathbf{x},c_{j,\mathbf{x}})|\le\xi'. \tag{15}$$

Hence, for all n large enough,

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{G}}\sup_{\mathbf{d}_k\in\bar{\mathcal{A}}_k^{\xi}(\mathbf{x})}|L_n(\mathbf{x},\mathbf{d}_k)-L^\star(\mathbf{x},\mathbf{d}_k)|\le\sup_{\mathbf{x}\in\mathcal{G}}\sup_{\mathbf{d}_k\in\bar{\mathcal{A}}_k^{\xi}(\mathbf{x})}\Big(&|L_n(\mathbf{x},\mathbf{d}_k)-L_n(\mathbf{x},c_{j,\mathbf{x}})|\\&+|L_n(\mathbf{x},c_{j,\mathbf{x}})-L^\star(\mathbf{x},c_{j,\mathbf{x}})|+|L^\star(\mathbf{x},c_{j,\mathbf{x}})-L^\star(\mathbf{x},\mathbf{d}_k)|\Big),\end{aligned}$$

where $c_{j,\mathbf{x}}$ satisfies $\|c_{j,\mathbf{x}}-\mathbf{d}_k\|_\infty\le\delta$. Using inequalities (14) and (15), with probability $1-\rho$, we obtain, for all n large enough,

$$\sup_{\mathbf{x}\in\mathcal{G}}\sup_{\mathbf{d}_k\in\bar{\mathcal{A}}_k^{\xi}(\mathbf{x})}|L_n(\mathbf{x},\mathbf{d}_k)-L^\star(\mathbf{x},\mathbf{d}_k)|\le 3\xi'.$$

Finally, by inequality (13), with probability $1-\rho$, for all n large enough,

$$\sup_{\mathbf{x}\in\mathcal{G}}\sup_{\mathbf{d}_{k-1}\in\mathcal{A}_{k-1}^{\xi}(\mathbf{x})}|L^\star(\mathbf{x},\mathbf{d}_{k-1},\hat{d}_{k,n}(\mathbf{d}_{k-1}))-L^\star(\mathbf{x},\mathbf{d}_{k-1},d_k^\star(\mathbf{d}_{k-1}))|\le 6\xi'. \tag{16}$$

Hereafter, to simplify, we assume that, for any given $(k-1)$-tuple of theoretical cuts, there is only one theoretical cut at level k, and leave the general case as an easy adaptation. Thus, we can define unambiguously

$$d_k^\star(\mathbf{d}_{k-1})=\arg\min_{d_k}L^\star(\mathbf{d}_{k-1},d_k).$$

Fix $\xi'' > 0$. From inequality (16), by evoking the equicontinuity of L_n and the compactness of $\mathcal{U} = \{(\mathbf{x}, \mathbf{d}_{k-1}) : \mathbf{x} \in \mathcal{G}, \mathbf{d}_{k-1} \in \mathcal{A}_{k-1}^{\xi}(\mathbf{x})\}$, we deduce that, with probability $1-\rho$, for all n large enough,

$$\sup_{(\mathbf{x},\mathbf{d}_{k-1})\in\mathcal{U}} d_{\infty}\Big(\hat{d}_{k,n}(\mathbf{d}_{k-1}), d_k^{\star}(\mathbf{d}_{k-1})\Big) \leq \xi''. \tag{17}$$

Besides,

$$\mathbb{P}\big[(\mathbf{X}, \hat{\mathbf{d}}_{k-1,n}(\mathbf{X})) \in \mathcal{U}\big] = \mathbb{E}\big[\mathbb{P}\big[(\mathbf{X}, \hat{\mathbf{d}}_{k-1,n}(\mathbf{X})) \in \mathcal{U}|\mathcal{D}_n\big]\big] \geq 1 - 2^{k-1}\xi. \tag{18}$$

In the rest of the proof, we consider $\xi \leq \rho/2^{k-1}$, which, by inequalities (17) and (18), leads to

$$\mathbb{P}\big[\sup_{(\mathbf{x},\mathbf{d}_{k-1})\in\mathcal{U}} d_{\infty}\Big(\hat{d}_{k,n}(\mathbf{d}_{k-1}), d_k^{\star}(\mathbf{d}_{k-1})\Big) \leq \xi'', (\mathbf{X}, \hat{\mathbf{d}}_{k-1,n}(\mathbf{X})) \in \mathcal{U}\big] \geq 1 - 2\rho.$$

This implies, with probability $1-2\rho$, for all n large enough,

$$d_{\infty}\Big(\hat{d}_{k,n}(\hat{\mathbf{d}}_{k-1,n}), d_k^{\star}(\hat{\mathbf{d}}_{k-1,n})\Big) \leq \xi''. \tag{19}$$

Main argument. Now, using triangle inequality,

$$\begin{aligned} d_{\infty}\Big(\hat{d}_{k,n}(\hat{\mathbf{d}}_{k-1,n}), \mathcal{A}_k^{\star}\Big) &\leq d_{\infty}\Big(\hat{d}_{k,n}(\hat{\mathbf{d}}_{k-1,n}), d_k^{\star}(\hat{\mathbf{d}}_{k-1,n})\Big) \\ &\quad + d_{\infty}\Big(d_k^{\star}(\hat{\mathbf{d}}_{k-1,n}), \mathcal{A}_k^{\star}\Big). \end{aligned} \tag{20}$$

Thus, we just have to show that $d_{\infty}(d_k^{\star}(\hat{\mathbf{d}}_{k-1,n}), \mathcal{A}_k^{\star}) \to 0$ in probability as $n \to \infty$, and the proof will be complete. To avoid confusion, we let $\{\mathbf{d}_{k-1}^{\star,i} : i \in \mathcal{I}\}$ be the set of best first $(k-1)$-th theoretical cuts (which can be either countable or not). With this notation, $d_k^{\star}(\mathbf{d}_{k-1}^{\star,i})$ is the k-th theoretical cuts given that the $(k-1)$ previous ones are $\mathbf{d}_{k-1}^{\star,i}$. For simplicity, let

$$L^{i,\star}(\mathbf{x}, d_k) = L_k^{\star}(\mathbf{x}, \mathbf{d}_{k-1}^{\star,i}, d_k) \quad \text{and} \quad \hat{L}^{\star}(\mathbf{x}, d_k) = L_k^{\star}(\mathbf{x}, \hat{\mathbf{d}}_{k-1,n}, d_k).$$

As before,

$$d_k^{\star}(\mathbf{d}_{k-1}^{\star,i}) \in \arg\min_{d_k} L^{i,\star}(\mathbf{x}, d_k) \quad \text{and} \quad d_k^{\star}(\hat{\mathbf{d}}_{k-1,n}) \in \arg\min_{d_k} \hat{L}^{\star}(\mathbf{x}, d_k).$$

Clearly, the result will be proved if we establish that,

$$\inf_{i\in\mathcal{I}} d_{\infty}(d_k^{\star}(\hat{\mathbf{d}}_{k-1,n}), d_k^{\star}(\mathbf{d}_{k-1}^{\star,i})) \to 0, \quad \text{in probability}, \quad \text{as } n \to \infty.$$

Note that, for all $\mathbf{x} \in \mathcal{G}$, $\bar{\mathcal{A}}_k^{\xi}(\mathbf{x})$ is compact. Thus, for all $\mathbf{x} \in \mathcal{G}$, there exists a finite subset $\mathcal{C}'_{\delta,\mathbf{x}} = \{c'_{j,\mathbf{x}} : 1 \le j \le p\}$ such that, for all d_k, $d_\infty(d_k, \mathcal{C}'_{\delta,\mathbf{x}}) \le \delta$. Hence, with probability $1-\rho$, for all n large enough,

$$\begin{aligned}
|\hat{L}^\star(\mathbf{x}, d_k) - L^{i,\star}(\mathbf{x}, d_k)| &\le |\hat{L}^\star(\mathbf{x}, d_k) - \hat{L}^\star(\mathbf{x}, c'_{j,\mathbf{x}})| \\
&\quad + |\hat{L}^\star(\mathbf{x}, c'_{j,\mathbf{x}}) - L^{i,\star}(\mathbf{x}, c'_{j,\mathbf{x}})| \\
&\quad + |L^{i,\star}(\mathbf{x}, c'_{j,\mathbf{x}}) - L^{i,\star}(\mathbf{x}, d_k)| \\
&\le 2\xi' + |\hat{L}^\star(\mathbf{x}, c'_{j,\mathbf{x}}) - L^{i,\star}(\mathbf{x}, c'_{j,\mathbf{x}})| \\
&\quad \text{(by the continuity of } L_k^\star\text{)}.
\end{aligned}$$

Therefore, as in inequality (13), with probability $1-\rho$, for all i and all n large enough,

$$\begin{aligned}
|L^{i,\star}(\mathbf{x}, d_k^\star(\hat{\mathbf{d}}_{k-1,n})) - L^{i,\star}(\mathbf{x}, d_k^\star(\mathbf{d}_{k-1}^{\star,i}))| &\le 2 \sup_{d_k} |\hat{L}^\star(\mathbf{x}, d_k) - L^{i,\star}(\mathbf{x}, d_k)| \\
&\le 4\xi' + 2 \max_j |\hat{L}^\star(\mathbf{x}, c'_{j,\mathbf{x}}) - L^{i,\star}(\mathbf{x}, c'_{j,\mathbf{x}})|.
\end{aligned}$$

Taking the infimum over all i, we obtain

$$\begin{aligned}
\inf_i |L^{i,\star}(\mathbf{x}, d_k^\star(\hat{\mathbf{d}}_{k-1,n})) - L^{i,\star}(\mathbf{x}, d_k^\star(\mathbf{d}_{k-1}^{\star,i}))| &\le 4\xi' \\
&\quad + 2 \inf_i \max_j |\hat{L}^\star(\mathbf{x}, c'_{j,\mathbf{x}}) - L^{i,\star}(\mathbf{x}, c'_{j,\mathbf{x}})|. \qquad (21)
\end{aligned}$$

Introduce ω, the modulus of continuity of $L_k^\star$:

$$\omega(\mathbf{x}, \delta) = \sup_{\|\mathbf{d}-\mathbf{d}'\|_\infty \le \delta} |L_k^\star(\mathbf{x}, \mathbf{d}) - L_k^\star(\mathbf{x}, \mathbf{d}')|.$$

Observe that, since $L_k^\star(\mathbf{x}, \cdot)$ is uniformly continuous, $\omega(\delta) \to 0$ as $\delta \to 0$. Hence, for all n large enough,

$$\begin{aligned}
&\inf_i \max_j |\hat{L}^\star(\mathbf{x}, c'_{j,\mathbf{x}}) - L^{i,\star}(\mathbf{x}, c'_{j,\mathbf{x}})| \\
&= \inf_i \max_j |L_k^\star(\mathbf{x}, \hat{\mathbf{d}}_{k-1,n}, c'_{j,\mathbf{x}}) - L_k^\star(\mathbf{x}, \mathbf{d}_{k-1}^{\star,i}, c'_{j,\mathbf{x}})| \\
&\le \inf_i \omega(\mathbf{x}, \|\hat{\mathbf{d}}_{k-1,n} - \mathbf{d}_{k-1}^{\star,i}\|_\infty) \\
&\le \xi', \qquad (22)
\end{aligned}$$

since, by assumption H_{k-1}, $\inf_i \|\hat{\mathbf{d}}_{k-1,n} - \mathbf{d}_{k-1}^{\star,i}\|_\infty \to 0$. Therefore, combining (21) and (22), with probability $1-\rho$, for all n large enough,

$$\inf_i |L^{i,\star}(\mathbf{X}, d_k^\star(\hat{\mathbf{d}}_{k-1,n})) - L^{i,\star}(\mathbf{X}, d_k^\star(\mathbf{d}_{k-1}^{\star,i}))| \le 6\xi.$$

Finally, by Technical Lemma 2 below, with probability $1-\rho$, for all n large enough,

$$\inf_i d_\infty(d_k^\star(\hat{\mathbf{d}}_{k-1,n}), d_k^\star(\mathbf{d}_{k-1}^{\star,i})) \le \xi''. \tag{23}$$

Plugging inequality (23) and (19) into (20), we conclude that, with probability $1-3\rho$, for all n large enough,

$$d_\infty\Big(\hat{d}_{k,n}(\hat{\mathbf{d}}_{k-1,n}), \mathcal{A}_k^\star\Big) \le 2\xi'',$$

which proves H_k. Property H_1 can be proved in the same way.

TECHNICAL LEMMA 2. *For all $\delta, \rho > 0$, there exists $\xi > 0$ such that, if, with probability $1-\rho$,*

$$\inf_i |L^{i,\star}(\mathbf{X}, d_k^\star(\hat{\mathbf{d}}_{k-1,n})) - L^{i,\star}(\mathbf{X}, d_k^\star(\mathbf{d}_{k-1}^{\star,i}))| \le \xi,$$

then, with probability $1-\rho$,

$$\inf_i d_\infty(d_k^\star(\hat{\mathbf{d}}_{k-1,n}), d_k^\star(\mathbf{d}_{k-1}^{\star,i})) \le \delta. \tag{24}$$

PROOF OF TECHNICAL LEMMA 2. Fix $\rho > 0$. Note that, for all $\delta > 0$, there exists $\xi > 0$ such that,

$$\inf_{\mathbf{x}\in[0,1]^d} \inf_i \inf_{y: d_\infty(y, d_k^\star(\mathbf{d}_{k-1}^{\star,i}))\ge\delta} |L_k^\star(\mathbf{x}, \mathbf{d}_{k-1}^{\star,i}, d_k^\star(\mathbf{d}_{k-1}^{\star,i})) - L_k^\star(\mathbf{x}, \mathbf{d}_{k-1}^{\star,i}, y)| \ge \xi.$$

To see this, assume that one can find $\delta > 0$ such that, for all $\xi > 0$, there exist $i_\xi, y_\xi, \mathbf{x}_\xi$ satisfying

$$|L_k^\star(\mathbf{x}_\xi, \mathbf{d}_{k-1}^{\star,i_\xi}, d_k^\star(\mathbf{d}_{k-1}^{\star,i_\xi})) - L_k^\star(\mathbf{x}_\xi, \mathbf{d}_{k-1}^{\star,i_\xi}, y_\xi)| \le \xi,$$

with $d_\infty(y_\xi, d_k^\star(\mathbf{d}_{k-1}^{\star,i_\xi})) \ge \delta$. Recall that $\{\mathbf{d}_{k-1}^{\star,i} : i \in \mathbb{N}\}$, $\{d_k^\star(\mathbf{d}_{k-1}^{\star,i}) : i \in \mathbb{N}\}$ are compact. Then, letting $\xi_p = 1/p$, we can extract three sequences $\mathbf{d}_{k-1}^{\star,i_p} \to \mathbf{d}_{k-1}$, $d_k^\star(\mathbf{d}_{k-1}^{\star,i_p}) \to d_k$ and $y_{\xi_{i_p}} \to y$ as $p \to \infty$ such that

$$L_k^\star(\mathbf{d}_{k-1}, d_k) = L_k^\star(\mathbf{d}_{k-1}, y), \tag{25}$$

and $d_\infty(y, d_k) \ge \delta$. Since we assume that given the $(k-1)$-th first cuts $\mathbf{d}_{k-1}$, there is only one best cut d_k, equation (25) implies that $y = d_k$, which is absurd.

Now, to conclude the proof, fix $\delta > 0$ and assume that, with probability $1-\rho$,

$$\inf_i d_\infty(d_k^\star(\mathbf{d}_{k-1}^{\star,i}), d_k^\star(\hat{\mathbf{d}}_{k-1,n})) \geq \delta.$$

Thus, with probability $1-\rho$,

$$\begin{aligned}
&\inf_i |L^{i,\star}(\mathbf{X}, d_k^\star(\hat{\mathbf{d}}_{k-1,n})) - L^{i,\star}(\mathbf{X}, d_k^\star(\mathbf{d}_{k-1}^{\star,i}))| \\
&\quad = \inf_i |L_k^\star(\mathbf{X}, \mathbf{d}_{k-1}^{\star,i}, d_k^\star(\hat{\mathbf{d}}_{k-1,n})) - L_k^\star(\mathbf{X}, \mathbf{d}_{k-1}^{\star,i}, d_k^\star)| \\
&\quad \geq \inf_{\mathbf{x}\in[0,1]^d} \inf_i \inf_{d_\infty(y, d_k^\star(\mathbf{d}_{k-1}^{\star,i}))\geq\delta} |L_k^\star(\mathbf{x}, \mathbf{d}_{k-1}^{\star,i}, y) - L_k^\star(\mathbf{x}, \mathbf{d}_{k-1}^{\star,i}, d_k^\star)| \\
&\quad \geq \xi,
\end{aligned}$$

which, by contraposition, concludes the proof. □

Proof of Proposition 1. Fix $k \in \mathbb{N}^\star$ and $\rho, \xi > 0$. According to Lemma 3, with probability $1-\rho$, for all n large enough, there exists a sequence of theoretical first k cuts $\mathbf{d}_k^\star(\mathbf{X}, \Theta)$ such that

$$d_\infty(\mathbf{d}_k^\star(\mathbf{X}, \Theta), \hat{\mathbf{d}}_{k,n}(\mathbf{X}, \Theta)) \leq \xi. \tag{26}$$

This implies that, with probability $1-\rho$, for all n large enough and all $1 \leq j \leq k$, the j-th empirical cut $\hat{d}_{j,n}(\mathbf{X}, \Theta)$ is performed along the same coordinate as $d_j^\star(\mathbf{X}, \Theta)$.

Now, for any cell A, since the regression function is not constant on A, one can find a theoretical cut $d_A^\star$ on A such that $L^\star(d_A^\star) > 0$. Thus, the cut $d_A^\star$ is made along an informative variable, in the sense that it is performed along one of the first S variables. Consequently, for all $\mathbf{X}, \Theta$ and for all $1 \leq j \leq k$, each theoretical cut $d_j^\star(\mathbf{X}, \Theta)$ is made along one of the first S coordinates. The proof is then a consequence of inequality (26). □

Acknowledgements. This work was supported by the European Research Council [SMAC-ERC-280032]. We greatly thank two referees for valuable comments and insightful suggestions.

References.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28:1302–1338, 2000.

E-mail: erwan.scornet@upmc.fr

E-mail: gerard.biau@upmc.fr

E-mail: jean-philippe.vert@mines-paristech.fr