Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

SUPPLEMENTARY MATERIAL FOR THE PAPER "MINIMAX OPTIMAL RATES FOR MONDRIAN TREES AND FORESTS"

By Jaouad Mourtada[*,†], Stéphane Gaïffas[*,‡] and Erwan Scornet[*,†]

École polytechnique[†] *and Université Paris Diderot*[‡]

1. Introduction. This supplementary material to the paper "Minimax optimal rates for Mondrian trees and forests" gathers several proofs and technical details and definitions that were omitted in the main paper. Namely, we start with a glossary of notations, then give extra definitions and notations for trees and nested trees partitions in Section 2. Then, we provide proofs that were omitted in the main paper by order of appearance, namely the proofs of Proposition 2, Theorem 1, Proposition 3, Proposition 4 and Lemma 1.

Sign	Description
$\mathscr{D}_n$	Data set
μ	Distribution of X on $[0,1]^d$
C, resp. $\lvert C\rvert$	A generic cell $C \subset [0,1]^d$, resp. half-perimeter of C
λ	Lifetime parameter of Mondrian process
$\mathsf{MP}(\lambda, C)$	Distribution of a Mondrian process defined on cell C with lifetime parameter λ.
Π_λ, resp. $\Pi_\lambda \lvert C$	Partition drawn from $\mathsf{MP}(\lambda, [0,1]^d)$, resp. from $\mathsf{MP}(\lambda, C)$
$C_\lambda(x)$	Cell of a Mondrian Tree with parameter λ containing x.
$D_\lambda(x)$	Diameter of $C_\lambda(x)$
K_λ	Number of cells in a Mondrian Tree partition Π_λ
$\widehat{f}^{(m)}_{\lambda,n}(x)$	Mondrian Tree estimate at query point x based on the Mondrian partition $\Pi^{(m)}_\lambda$
$\widehat{f}_{\lambda,n,M}(x)$	Mondrian Forest estimate at query point x based on the Mondrian partitions $\Pi_{\lambda,M} = (\Pi^{(1)}_\lambda, \ldots, \Pi^{(M)}_\lambda)$

[*]Data Science Initiative of École polytechnique

Sign	Description
$\bar{f}_\lambda^{(m)}(x)$	Expected value of the regression function f inside the cell $C_\lambda^{(m)}(x)$
$\widetilde{f}_\lambda(x)$	Expected value of $\bar{f}_\lambda^{(m)}(x)$ over $\Pi_\lambda^{(m)} \sim \mathsf{MP}(\lambda, [0,1]^d)$
$\mathcal{N}(T), \mathcal{N}^\circ(T), \mathcal{L}(T)$	Nodes, interior nodes and leaves of a tree
$\Sigma = (\sigma_\mathbf{v})_{\mathbf{v}\in\mathcal{N}^\circ(T)}$	Set of splits for all nodes in the tree
$\sigma_\mathbf{v} = (j_\mathbf{v}, s_\mathbf{v})$	A split at node $\mathbf{v}$ characterized by its split dimension $j_\mathbf{v} \in \{1, \dots, d\}$ and its threshold $s_\mathbf{v} \in [0,1]$
$\tau_\mathbf{v}$	Birth time of a node $\mathbf{v}$

2. Specific notations. Let us now introduce some specific notations to describe the decision tree structure and the Mondrian Process.

2.1. *Trees and nested trees partitions.* A decision tree (T, Σ) is composed of the following components:

- A finite rooted ordered binary tree T, with nodes $\mathcal{N}(T)$, interior nodes $\mathcal{N}^\circ(T)$ and leaves $\mathcal{L}(T)$ (so that $\mathcal{N}(T)$ is the disjoint union of $\mathcal{N}^\circ(T)$ and $\mathcal{L}(T)$). The nodes $\mathbf{v} \in \mathcal{N}(T)$ are finite words on the alphabet $\{0,1\}$, that is elements of the set $\{0,1\}^* = \bigcup_{n\geqslant 0}\{0,1\}^n$: the root ϵ of T is the empty word, and for every interior $\mathbf{v} \in \{0,1\}^*$, its left child is $\mathbf{v}0$ (obtained by adding a 0 at the end of $\mathbf{v}$) while its right child is $\mathbf{v}1$ (obtained by adding a 1 at the end of $\mathbf{v}$).
- A family of *splits* $\Sigma = (\sigma_\mathbf{v})_{\mathbf{v}\in\mathcal{N}^\circ(T)}$ at each interior node, where each split $\sigma_\mathbf{v} = (j_\mathbf{v}, s_\mathbf{v})$ is characterized by its split dimension $j_\mathbf{v} \in \{1, \dots, d\}$ and its threshold $s_\mathbf{v} \in [0,1]$.

We associate to $\Pi = (T, \Sigma)$ a partition $(C_\mathbf{v})_{\mathbf{v}\in\mathcal{L}(T)}$ of the unit cube $[0,1]^d$, called a *tree partition* (or *guillotine partition*). For each node $\mathbf{v} \in \mathcal{N}(T)$, we define a hyper-rectangular region $C_\mathbf{v}$ recursively:

- The cell associated to the root of T is $[0,1]^d$;
- For each $\mathbf{v} \in \mathcal{N}^\circ(T)$, we define

$$C_{\mathbf{v}0} := \{x \in C_\mathbf{v} : x_{j_\mathbf{v}} \leqslant s_{j_\mathbf{v}}\} \quad \text{and} \quad C_{\mathbf{v}1} := C_\mathbf{v} \setminus C_{\mathbf{v}0}.$$

The leaf cells $(C_\mathbf{v})_{\mathbf{v}\in\mathcal{L}(T)}$ form a partition of $[0,1]^d$ by construction. In what follows, we will identify a tree with splits (T, Σ) with its associated tree partition, and a node $\mathbf{v} \in \mathcal{N}(T)$ with the cell $C_\mathbf{v} \subset [0,1]^d$. The Mondrian process, described in the next Section, defines a distribution over nested tree partitions, defined below.

DEFINITION 1 (Nested tree partitions). A tree partition $\Pi' = (T', \Sigma')$ is a *refinement* of the tree partition $\Pi = (T, \Sigma)$ if T is a subtree of T' and, for every $\mathbf{v} \in \mathcal{N}(T) \subseteq \mathcal{N}(T')$, $\sigma_{\mathbf{v}} = \sigma'_{\mathbf{v}}$. A *nested tree partition* is a family $(\Pi_t)_{t \geqslant 0}$ of tree partitions such that, for every $t, t' \in \mathbf{R}^+$ with $t \leqslant t'$, $\Pi_{t'}$ is a refinement of Π_t. Such a family can be described as follows: let $\mathbf{T}$ be the (in general infinite, and possibly complete) rooted binary tree, such that $\mathcal{N}(\mathbf{T}) = \bigcup_{t \geqslant 0} \mathcal{N}(T_t) \subseteq \{0,1\}^*$. For each $\mathbf{v} \in \mathcal{N}(T)$, let $\tau_{\mathbf{v}} = \inf\{t \geqslant 0 \mid \mathbf{v} \in \mathcal{N}(T_t)\} < \infty$ denote the *birth time* of the node $\mathbf{v}$. Additionally, let $\sigma_{\mathbf{v}}$ be the value of the split $\sigma_{\mathbf{v},t}$ in Π_t for $t > \tau_{\mathbf{v}}$ (which does not depend on t by the refinement property). Then, Π is completely characterized by $\mathbf{T}$, $\Sigma = (\sigma_{\mathbf{v}})_{\mathbf{v} \in \mathcal{N}(\mathbf{T})}$ and $\mathfrak{T} = (\tau_{\mathbf{v}})_{\mathbf{v} \in \mathcal{N}(\mathbf{T})}$.

2.2. *Mondrian Process.* To define rigorously the Mondrian Process, we introduce the function Φ_C, which maps any family of couples $(e^j_{\mathbf{v}}, u^j_{\mathbf{v}}) \in \mathbf{R}^+ \times [0,1]$ indexed by the coordinates $j \in \{1, \dots, d\}$ and the nodes $\mathbf{v} \in \{0,1\}^*$ to a nested tree partition $\Pi = \Phi_C((e^j_{\mathbf{v}}, u^j_{\mathbf{v}})_{\mathbf{v},j})$ of C. The splits $\sigma_{\mathbf{v}} = (j_{\mathbf{v}}, s_{\mathbf{v}})$ and birth times $\tau_{\mathbf{v}}$ of the nodes $\mathbf{v} \in \{0,1\}^*$ are defined recursively, starting from the root ϵ:

- For the root node ϵ, we let $\tau_\epsilon = 0$ and $C_\epsilon = C$.
- At each node $\mathbf{v} \in \{0,1\}^*$, given the labels of all its ancestors $\mathbf{v}' \sqsubset \mathbf{v}$ (so that in particular $\tau_{\mathbf{v}}$ and $C_{\mathbf{v}}$ are determined), denote $C_{\mathbf{v}} = \prod_{j=1}^d [a^j_{\mathbf{v}}, b^j_{\mathbf{v}}]$. Then, select the split dimension $j_{\mathbf{v}} \in \{1, \dots, d\}$ and its location $s_{\mathbf{v}}$ as follows:

$$j_{\mathbf{v}} = \operatorname*{argmin}_{j=1,\dots,d} \frac{e^j_{\mathbf{v}}}{b^j_{\mathbf{v}} - a^j_{\mathbf{v}}}, \qquad s_{\mathbf{v}} = a^{j_{\mathbf{v}}}_{\mathbf{v}} + (b^{j_{\mathbf{v}}}_{\mathbf{v}} - a^{j_{\mathbf{v}}}_{\mathbf{v}}) \cdot u^{j_{\mathbf{v}}}_{\mathbf{v}}, \tag{2.1}$$

  where we break ties in the choice of $j_{\mathbf{v}}$ e.g., by choosing the smallest index j in the argmin. The node $\mathbf{v}$ is then split at time $\tau_{\mathbf{v}} + e^{j_{\mathbf{v}}}_{\mathbf{v}} / (b^{j_{\mathbf{v}}}_{\mathbf{v}} - a^{j_{\mathbf{v}}}_{\mathbf{v}}) = \tau_{\mathbf{v}0} = \tau_{\mathbf{v}1}$, we let $C_{\mathbf{v}0} = \{x \in C_{\mathbf{v}} : x_{j_{\mathbf{v}}} \leqslant s_{\mathbf{v}}\}$, $C_{\mathbf{v}1} = C_{\mathbf{v}} \setminus C_{\mathbf{v}0}$ and recursively apply the procedure to its children $\mathbf{v}0$ and $\mathbf{v}1$.

For each $\lambda \in \mathbf{R}^+$, the tree partition $\Pi_\lambda = \Phi_{\lambda,C}((e^j_{\mathbf{v}}, u^j_{\mathbf{v}})_{\mathbf{v},j})$ is the *pruning of* Π *at time* λ, obtained by removing all the splits in Π that occurred strictly after λ, so that the leaves of the tree are the maximal nodes (in the prefix order) $\mathbf{v}$ such that $\tau_{\mathbf{v}} \leqslant \lambda$.

DEFINITION 2 (Mondrian process). Let $(E^j_{\mathbf{v}}, U^j_{\mathbf{v}})_{\mathbf{v},j}$ be a family of independent random variables, with $E^j_{\mathbf{v}} \sim \mathsf{Exp}(1)$, $U^j_{\mathbf{v}} \sim \mathcal{U}([0,1])$. The *Mondrian process* $\mathsf{MP}(C)$ on C is the distribution of the random nested tree partition $\Phi_C((E^j_{\mathbf{v}}, U^j_{\mathbf{v}})_{\mathbf{v},j})$. In addition, we denote $\mathsf{MP}(\lambda, C)$ the distribution of $\Phi_{\lambda,C}((E^j_{\mathbf{v}}, U^j_{\mathbf{v}})_{\mathbf{v},j})$.

3. Proof of Proposition 2. At a high level, the idea of the proof is to modify the construction of the Mondrian partition (and hence, the distribution of the underlying process) without affecting the expected number of cells. More precisely, we show a recursive way to transform the Mondrian process that leaves $\mathbb{E}[K_\lambda]$ unchanged, and which eventually leads to a random partition $\widetilde{\Pi}_\lambda$ for which this quantity can be computed directly and equals $(1+\lambda)^d$. We will in fact show the result for a general box C (not just the unit cube). The proof proceeds in two steps:

1. Define a modified process $\widetilde{\Pi}$, and show that $\mathbb{E}[\widetilde{K}_\lambda] = \prod_{j=1}^d (1+\lambda|C^j|)$.
2. It remains to show that $\mathbb{E}[K_\lambda] = \mathbb{E}[\widetilde{K}_\lambda]$. For this, it is sufficient to show that the distribution of the birth times $\tau_{\mathbf{v}}$ and $\widetilde{\tau}_{\mathbf{v}}$ of the node $\mathbf{v}$ is the same for both processes. This is done by induction on $\mathbf{v}$, by showing that the splits at one node of both processes have the same conditional distribution given the splits at previous nodes.

Let $(E^j_{\mathbf{v}}, U^j_{\mathbf{v}})_{\mathbf{v}\in\{0,1\}^*, 1\leqslant j\leqslant d}$ be a family of independent random variables with $E^j_{\mathbf{v}} \sim \mathsf{Exp}(1)$ and $U^j_{\mathbf{v}} \sim \mathcal{U}([0,1])$. By definition, $\Pi = \Phi_C((E^j_{\mathbf{v}}, U^j_{\mathbf{v}})_{\mathbf{v},j})$ (Φ_C being defined in Section 3) follows a Mondrian process distribution $\mathsf{MP}(C)$. Denote for every node $\mathbf{v} \in \{0,1\}^*$ $C_{\mathbf{v}}$ the cell of $\mathbf{v}$, $\tau_{\mathbf{v}}$ its birth time, as well as its split time $T_{\mathbf{v}}$, dimension $J_{\mathbf{v}}$, and threshold $S_{\mathbf{v}}$ (note that $T_{\mathbf{v}} = \tau_{\mathbf{v}0} = \tau_{\mathbf{v}1}$). In addition, for $\lambda \in \mathbf{R}^+$, denote $\Pi_\lambda \sim \mathsf{MP}(\lambda, C)$ the tree partition restricted to time λ, and $K_\lambda \in \mathbf{N} \cup \{+\infty\}$ its number of nodes.

Construction of the modified process. Now, consider the following modified nested partition of C, denoted $\widetilde{\Pi}$, and defined through its split times, dimension and threshold $\widetilde{T}_{\mathbf{v}}, \widetilde{J}_{\mathbf{v}}, \widetilde{S}_{\mathbf{v}}$ (which determine the birth times $\tau_{\mathbf{v}}$ and cells $C_{\mathbf{v}}$), and *current j-dimensional node* $\mathbf{v}_j(\mathbf{v}) \in \{0,1\}^*$ $(1 \leqslant j \leqslant d)$ at each node $\mathbf{v}$. First, for every $j = 1, \ldots, d$, let $\Pi'^j = \Phi_{C^j}((E^j_{\mathbf{v}}, U^j_{\mathbf{v}})_{\mathbf{v}\in\{0,1\}^*}) \sim \mathsf{MP}(C^j)$ be the nested partition of the interval C^j determined by $(E^j_{\mathbf{v}}, U^j_{\mathbf{v}})_{\mathbf{v}}$; its split times and thresholds are denoted $(S'^j_{\mathbf{v}}, T'^j_{\mathbf{v}})$. Then, $\widetilde{\Pi}$ is defined recursively as follows:

- At the root node ϵ, let $\widetilde{\tau}_\epsilon = 0$, $\widetilde{C}_\epsilon = C$ and $\mathbf{v}_j(\epsilon) := \epsilon$ for $1 \leqslant j \leqslant d$.
- At node $\mathbf{v}$, given $(\tau_{\mathbf{v}'}, C_{\mathbf{v}'}, \mathbf{v}_j(\mathbf{v}'))_{\mathbf{v}' \sqsubseteq \mathbf{v}}$ (*i.e.*, given $(\widetilde{J}_{\mathbf{v}'}, \widetilde{S}_{\mathbf{v}'}, \widetilde{T}_{\mathbf{v}'})_{\mathbf{v}' \sqsubset \mathbf{v}}$) define:

$$\widetilde{T}_{\mathbf{v}} = \min_{1\leqslant j\leqslant d} T'^j_{\mathbf{v}_j(\mathbf{v})}, \qquad \widetilde{J}_{\mathbf{v}} := \operatorname*{argmin}_{1\leqslant j\leqslant d} T'^j_{\mathbf{v}_j(\mathbf{v})}, \qquad \widetilde{S}_{\mathbf{v}} = S'^j_{\mathbf{v}_j(\mathbf{v})}, \tag{3.1}$$

$$\mathbf{v}_j(\mathbf{v}a) = \begin{cases} \mathbf{v}_j(\mathbf{v})a & \text{if } j = \widetilde{J}_{\mathbf{v}} \\ \mathbf{v}_j(\mathbf{v}) & \text{else.} \end{cases} \tag{3.2}$$

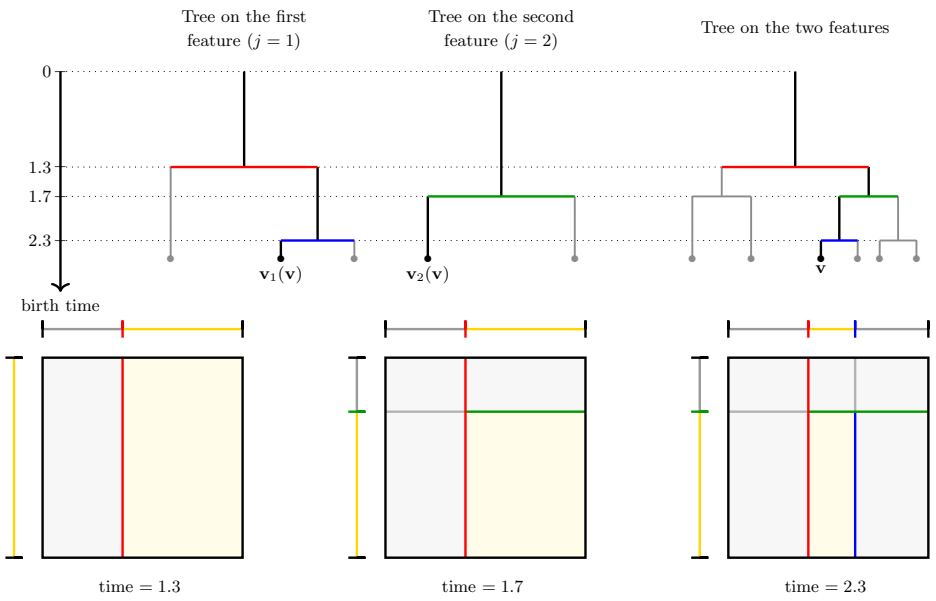

FIG 1. *Modified construction in dimension two. At the top, from left to right: trees associated to partitions* Π'^1, Π'^2 *and* $\widetilde{\Pi}$ *respectively. At the bottom, from left to right: successive splits in* $\widetilde{\Pi}$ *leading to the leaf* $\mathbf{v}$ *(depicted in yellow).*

Finally, for every $\lambda \in \mathbf{R}^+$, define $\widetilde{\Pi}_\lambda$ and $\widetilde{K}_\lambda$ as before from $\widetilde{\Pi}$. This construction is illustrated in Figure 1.

Computation of $\mathbb{E}[\widetilde{K}_\lambda]$. Now, it can be seen that the partition $\widetilde{\Pi}_\lambda$ is a rectangular grid which is the "product" of the partitions Π'^j of the intervals C^j, $1 \leqslant j \leqslant d$. Indeed, let $x \in [0,1]^d$, and let $\widetilde{C}_\lambda(x)$ be the cell in $\widetilde{\Pi}_\lambda$ that contains x; we need to show that $\widetilde{C}_\lambda(x) = \prod_{j=1}^d C'^j_\lambda(x)$, where $C'^j_\lambda(x)$ is the subinterval of C^j in the partition Π'^j that contains x_j. The proof proceeds in several steps:

- First, Equation (3.1) shows that, for every node $\mathbf{v}$, we have $\widetilde{C}_\mathbf{v} = \prod_{1 \leqslant j \leqslant d} C'^j_{\mathbf{v}_j(\mathbf{v})}$, since the successive splits on the j-th coordinate of $\widetilde{C}_\mathbf{v}$ are precisely the ones of $C'^j_{\mathbf{v}_j(\mathbf{v})}$.
- Second, it follows from (3.1) that $\widetilde{T}_\mathbf{v} = \min_{1 \leqslant j \leqslant d} T'^j_{\mathbf{v}_j(\mathbf{v})}$; also, since the cell $C_\mathbf{v}$ is formed when its last split is performed, $\widetilde{\tau}_\mathbf{v} = \max_{1 \leqslant j \leqslant d} \tau'^j_{\mathbf{v}_j(\mathbf{v})}$.
- Let $\widetilde{\mathbf{v}}$ be the node such that $\widetilde{C}_{\widetilde{\mathbf{v}}} = \widetilde{C}_\lambda(x)$, and $\mathbf{v}'^j$ be such that $C'^j_{\mathbf{v}'^j} = C'^j_\lambda(x_j)$. By the first point, it suffices to show that $\mathbf{v}_j(\widetilde{\mathbf{v}}) = \mathbf{v}'_j$ for $1 \leqslant j \leqslant d$.

- Observe that $\widetilde{\mathbf{v}}$ (resp. $\mathbf{v}'_j$) is characterized by the fact that $x \in \widetilde{C}_{\widetilde{\mathbf{v}}}$ and $\widetilde{\tau}_{\widetilde{\mathbf{v}}} \leqslant \lambda < \widetilde{T}_{\widetilde{\mathbf{v}}}$ (resp. $x_j \in C'^j_{\mathbf{v}'^j}$ and $\tau'^j_{\mathbf{v}'^j} \leqslant \lambda < T'^j_{\mathbf{v}'^j}$). But since $\widetilde{C}_{\widetilde{\mathbf{v}}} = \prod_{1 \leqslant j \leqslant d} C'^j_{\mathbf{v}_j(\widetilde{\mathbf{v}})}$ (first point), $x \in \widetilde{C}_{\widetilde{\mathbf{v}}}$ implies $x_j \in C'^j_{\mathbf{v}_j(\widetilde{\mathbf{v}})}$. Likewise, since $\widetilde{\tau}_{\widetilde{\mathbf{v}}} = \max_{1 \leqslant j \leqslant d} \tau'^j_{\mathbf{v}_j(\widetilde{\mathbf{v}})}$ and $\widetilde{T}_{\widetilde{\mathbf{v}}} = \min_{1 \leqslant j \leqslant d} T'^j_{\mathbf{v}_j(\widetilde{\mathbf{v}})}$ (second point), $\widetilde{\tau}_{\widetilde{\mathbf{v}}} \leqslant \lambda < \widetilde{T}_{\widetilde{\mathbf{v}}}$ implies $\tau'^j_{\mathbf{v}_j(\widetilde{\mathbf{v}})} \leqslant \lambda < T'^j_{\mathbf{v}_j(\widetilde{\mathbf{v}})}$. Since these properties characterize $\mathbf{v}'^j$, we have $\mathbf{v}_j(\widetilde{\mathbf{v}}) = \mathbf{v}'_j$, which concludes the proof.

Hence, the partition $\widetilde{\Pi}_\lambda$ is the product of the partitions $\Pi'^j = \Phi_{C^j}((E^j_{\mathbf{v}}, U^j_{\mathbf{v}})_{\mathbf{v}})_\lambda$ of the intervals C^j, $1 \leqslant j \leqslant d$, which are independent Mondrians distributed as $\mathsf{MP}(\lambda, C^j)$. By Fact 1, the splits of the Mondrian partition $\mathsf{MP}(\lambda, C^j)$ are distributed as a Poisson point process on C^j of intensity λ, so that the expected number of cells in such a partition is $1 + \lambda|C^j|$. Since $\widetilde{\Pi}_\lambda$ is a "product" of such independent partitions, we have:

$$\mathbb{E}[\widetilde{K}_\lambda] = \prod_{j=1}^{d} (1 + \lambda |C^j|)\,. \tag{3.3}$$

Equality of $\mathbb{E}[K_\lambda]$ *and* $\mathbb{E}[\widetilde{K}_\lambda]$. In order to establish Proposition 2, it is thus sufficient to prove that $\mathbb{E}[K_\lambda] = \mathbb{E}[\widetilde{K}_\lambda]$. First, note that, since the number of cells in a partition is one plus the number of splits (each split increases the number of cells by one)

$$K_\lambda = 1 + \sum_{\mathbf{v} \in \{0,1\}^*} \mathbf{1}(T_{\mathbf{v}} \leqslant \lambda)$$

so that we have, respectively,

$$\mathbb{E}[K_\lambda] = 1 + \sum_{\mathbf{v} \in \{0,1\}^*} \mathbb{P}(T_{\mathbf{v}} \leqslant \lambda) \tag{3.4}$$

$$\mathbb{E}[\widetilde{K}_\lambda] = 1 + \sum_{\mathbf{v} \in \{0,1\}^*} \mathbb{P}(\widetilde{T}_{\mathbf{v}} \leqslant \lambda)\,. \tag{3.5}$$

Hence, it suffices to show that $\mathbb{P}(T_{\mathbf{v}} \leqslant \lambda) = \mathbb{P}(\widetilde{T}_{\mathbf{v}} \leqslant \lambda)$ for every $\mathbf{v} \in \{0,1\}^*$ and $\lambda \geqslant 0$, *i.e.* that $T_{\mathbf{v}}$ and $\widetilde{T}_{\mathbf{v}}$ have the same distribution for every $\mathbf{v}$.

In order to establish this, we show that, for every $\mathbf{v} \in \{0,1\}^*$, the conditional distribution of $(\widetilde{T}_{\mathbf{v}}, \widetilde{J}_{\mathbf{v}}, \widetilde{S}_{\mathbf{v}})$ given $\widetilde{\mathscr{F}}_{\mathbf{v}} = \sigma((\widetilde{T}_{\mathbf{v}'}, \widetilde{J}_{\mathbf{v}'}, \widetilde{S}_{\mathbf{v}'}), \mathbf{v}' \sqsubset \mathbf{v})$ has the same form as the conditional distribution of $(T_{\mathbf{v}}, J_{\mathbf{v}}, S_{\mathbf{v}})$ given $\mathscr{F}_{\mathbf{v}} = \sigma((T_{\mathbf{v}'}, J_{\mathbf{v}'}, S_{\mathbf{v}'}), \mathbf{v}' \sqsubset \mathbf{v})$, in the sense that there exits a family of conditional distributions $(\Psi_{\mathbf{v}})_{\mathbf{v}}$ such that, for every $\mathbf{v}$, the conditional distribution of

$(T_{\mathbf{v}}, J_{\mathbf{v}}, S_{\mathbf{v}})$ given $\mathscr{F}_{\mathbf{v}}$ is $\Psi_{\mathbf{v}}(\cdot|(T_{\mathbf{v}'}, J_{\mathbf{v}'}, S_{\mathbf{v}'}), \mathbf{v}' \sqsubset \mathbf{v})$ and the conditional distribution of $(\widetilde{T}_{\mathbf{v}}, \widetilde{J}_{\mathbf{v}}, \widetilde{S}_{\mathbf{v}})$ given $\widetilde{\mathscr{F}}_{\mathbf{v}}$ is $\Psi_{\mathbf{v}}(\cdot|(\widetilde{T}_{\mathbf{v}'}, \widetilde{J}_{\mathbf{v}'}, \widetilde{S}_{\mathbf{v}'}), \mathbf{v}' \sqsubset \mathbf{v})$.

First, recall that the variables $(E^j_{\mathbf{v}'}, U^j_{\mathbf{v}'})_{\mathbf{v}' \in \{0,1\}^*, 1 \leqslant j \leqslant d}$ are independent, so $(E^j_{\mathbf{v}}, U^j_{\mathbf{v}})_{1 \leqslant j \leqslant d}$ is independent from $\mathscr{F}_{\mathbf{v}}$. Hence, conditionally on $\mathscr{F}_{\mathbf{v}}$, $E^j_{\mathbf{v}}, U^j_{\mathbf{v}}$, $1 \leqslant j \leqslant d$ are independent with $E^j_{\mathbf{v}} \sim \mathsf{Exp}(1)$ and $U^j_{\mathbf{v}} \sim \mathcal{U}([0,1])$. Also, recall that if $T_1, \dots, T_d$ are independent exponential random variables of intensities $\lambda_1, \dots, \lambda_d$, and if $T = \min_{1 \leqslant j \leqslant d} T_j$ and $J = \operatorname{argmin}_{1 \leqslant j \leqslant d} T_j$, then $\mathbb{P}(J = j) = \lambda_j / \sum_{j'=1}^d \lambda_{j'}$, $T \sim \mathsf{Exp}(\sum_{j=1}^d \lambda_j)$ and J and T are independent. Hence, conditionally on $\mathscr{F}_{\mathbf{v}}$, $T_{\mathbf{v}} - \tau_{\mathbf{v}} = \min_{1 \leqslant j \leqslant d} E^j_{\mathbf{v}} / |C^j_{\mathbf{v}}| \sim \mathsf{Exp}(\sum_{j=1}^d |C^j_{\mathbf{v}}|) = \mathsf{Exp}(|C_{\mathbf{v}}|)$, $J_{\mathbf{v}} := \operatorname{argmin}_{1 \leqslant j \leqslant d} E^j_{\mathbf{v}} / |C^j_{\mathbf{v}}|$ equals j with probability $|C^j_{\mathbf{v}}|/|C_{\mathbf{v}}|$, $T_{\mathbf{v}}, J_{\mathbf{v}}$ are independent and $(S_{\mathbf{v}}|T_{\mathbf{v}}, J_{\mathbf{v}}) \sim \mathcal{U}(C^{J_{\mathbf{v}}}_{\mathbf{v}})$.

Now consider the conditional distribution of $(\widetilde{T}_{\mathbf{v}}, \widetilde{J}_{\mathbf{v}}, \widetilde{S}_{\mathbf{v}})$ given $\widetilde{\mathscr{F}}_{\mathbf{v}}$. Let $(\mathbf{v}_v)_{v \in \mathbf{N}}$ be a path in $\{0,1\}^*$ from the root: $\mathbf{v}_0 := \epsilon$, $\mathbf{v}_{v+1}$ is a child of $\mathbf{v}_v$ for $v \in \mathbf{N}$, and $\mathbf{v}_v \sqsubseteq \mathbf{v}$ for $0 \leqslant v \leqslant \operatorname{depth}(\mathbf{v})$. Define for $v \in \mathbf{N}$, $E^j_v = E^j_{\mathbf{v}_v}$ and $U^j_v = U^j_{\mathbf{v}_v}$ if $\mathbf{v}_{v+1}$ is the left child of $\mathbf{v}_v$, and $1 - U^j_{\mathbf{v}_v}$ otherwise. Then, the variables $(E^j_v, U^j_v)_{v \in \mathbf{N}, 1 \leqslant j \leqslant d}$ are independent, with $E^j_v \sim \mathsf{Exp}(1)$, $U^j_v \sim \mathcal{U}([0,1])$, so that the following Lemma applies.

LEMMA 1. *Let $(E^j_v, U^j_v)_{v \in \mathbf{N}^\star, 1 \leqslant j \leqslant d}$ be a family of independent random variables, with $U^j_v \sim \mathcal{U}([0,1])$ and $E^j_v \sim \mathsf{Exp}(1)$. Let $a_1, \dots, a_d > 0$. For $1 \leqslant j \leqslant d$, define the sequence $(T^j_v, L^j_v)_{v \in \mathbf{N}}$ as follows*:

- $L^j_0 = a_j$, $T^j_0 = \frac{E^j_0}{a_j}$;
- *for $v \in \mathbf{N}$, $L^j_{v+1} = U^j_v L^j_v$, $T^j_{v+1} = T^j_v + \frac{E^j_{v+1}}{L^j_{v+1}}$.*

Define recursively the variables $\widetilde{V}^j_v$ $(v \in \mathbf{N}, 1 \leqslant j \leqslant d)$ as well as $\widetilde{J}_v, \widetilde{T}_v, \widetilde{U}_v$ $(v \in \mathbf{N})$ as follows:

- $\widetilde{V}^j_0 = 0$ *for $j = 1, \dots, d$.*
- *for $v \in \mathbf{N}$, given $\widetilde{V}^j_v$ $(1 \leqslant j \leqslant d)$, denoting $\widetilde{T}^j_v = T^j_{\widetilde{V}^j_v}$ and $\widetilde{U}^j_v = U^j_{\widetilde{V}^j_v}$, set*

$$\widetilde{J}_v = \operatorname*{argmin}_{1 \leqslant j \leqslant d} \widetilde{T}^j_v, \quad \widetilde{T}_v = \min_{1 \leqslant j \leqslant d} \widetilde{T}^j_v = \widetilde{T}^{\widetilde{J}_v}_v, \widetilde{U}_v = \widetilde{U}^{\widetilde{J}_v}_v, \tag{3.6}$$
$$\text{and} \quad \widetilde{V}^j_{v+1} = \widetilde{V}^j_v + \mathbf{1}(\widetilde{J}_v = j).$$

Then, the conditional distribution of $(\widetilde{J}_v, \widetilde{T}_v, \widetilde{U}_v)$ given $\mathscr{F}_v = \sigma((\widetilde{J}_{v'}, \widetilde{T}_{v'}, \widetilde{U}_{v'}), 0 \leqslant v' < v)$ is the following (denoting $\widetilde{L}^j_v = L^j_{\widetilde{V}^j_v}$):

- $\widetilde{J}_v, \widetilde{T}_v, \widetilde{U}_v$ *are independent,*

- $\mathbb{P}(\widetilde{J}_v = j \,|\, \widetilde{\mathscr{F}}_v) = \widetilde{L}_v^j / (\sum_{j'=1}^d \widetilde{L}_v^{j'})$,
- $\widetilde{T}_v - \widetilde{T}_{v-1} \sim \mathsf{Exp}(\sum_{j=1}^d \widetilde{L}_v^j)$ (*with the convention* $\widetilde{T}_{-1} = 0$) *and* $\widetilde{U}_v \sim \mathcal{U}([0,1])$.

In addition, note that, with the notations of Lemma 1, a simple induction shows that $\widetilde{J}_v = \widetilde{J}_{\mathbf{v}_v}$, $\widetilde{T}_v = \widetilde{T}_{\mathbf{v}_v}$, $\widetilde{U}_v = \widetilde{U}_{\mathbf{v}_v}$ and $L_v^j = |\widetilde{C}_{\mathbf{v}_v}^j|$, so that $\widetilde{\mathscr{F}}_v = \widetilde{\mathscr{F}}_{\mathbf{v}_v}$. Applying Lemma 1 for $v = \mathrm{depth}(\mathbf{v})$ (so that $\mathbf{v}_v = \mathbf{v}$) therefore gives the following: conditionally on $\widetilde{\mathscr{F}}_{\mathbf{v}}$, the variables $\widetilde{T}_{\mathbf{v}}, \widetilde{J}_{\mathbf{v}}, \widetilde{U}_{\mathbf{v}}$ are independent, $\widetilde{T}_{\mathbf{v}} - \widetilde{\tau}_{\mathbf{v}} \sim \mathsf{Exp}(|\widetilde{C}_{\mathbf{v}}^j|)$, $\mathbb{P}(\widetilde{J}_{\mathbf{v}} = j \,|\, \widetilde{\mathscr{F}}_{\mathbf{v}}) = |\widetilde{C}_{\mathbf{v}}^j| / \big(\sum_{j'=1}^d |\widetilde{C}_{\mathbf{v}}^j|\big)$ and $\widetilde{U}_{\mathbf{v}} \sim \mathcal{U}([0,1])$, so that $(\widetilde{S}_{\mathbf{v}} | \widetilde{\mathscr{F}}_{\mathbf{v}}, \widetilde{T}_{\mathbf{v}}, \widetilde{J}_{\mathbf{v}}) \sim \mathcal{U}(\widetilde{C}_{\mathbf{v}}^{\widetilde{J}_{\mathbf{v}}})$. Hence, we have proven that, for every $\mathbf{v}$, the conditional distribution of $(T_{\mathbf{v}}, J_{\mathbf{v}}, S_{\mathbf{v}})$ given $\widetilde{\mathscr{F}}_{\mathbf{v}}$ is the same as that of $(\widetilde{T}_{\mathbf{v}}, \widetilde{J}_{\mathbf{v}}, \widetilde{S}_{\mathbf{v}})$ given $\widetilde{\mathscr{F}}_{\mathbf{v}}$. By induction on $\mathbf{v}$, since $\mathscr{F}_\epsilon = \widetilde{\mathscr{F}}_\epsilon$ is the trivial σ-algebra, this shows that $T_{\mathbf{v}}$ and $\widetilde{T}_{\mathbf{v}}$ have the same distribution for every $\mathbf{v}$. Plugging this into (3.4) and (3.5) and combining it with (3.3) completes the proof of Proposition 2. □

Proof of Lemma 1. We show by induction on $v \in \mathbf{N}$ the following property: conditionally on $\widetilde{\mathscr{F}}_{\mathbf{v}}$, $(\widetilde{T}_v^j, \widetilde{U}_v^j)_{1 \leqslant j \leqslant d}$ are independent, $\widetilde{T}_v^j - \widetilde{T}_{v-1} \sim \mathsf{Exp}(L_v^j)$ and $\widetilde{U}_v^j \sim \mathcal{U}([0,1])$.

Initialization For $v = 0$ (with $\widetilde{\mathscr{F}}_0$ the trivial σ-algebra), since $\widetilde{V}_0^j = 0$ we have $\widetilde{T}_0^j = E_0^j / a_j \sim \mathsf{Exp}(a_j) = \mathsf{Exp}(L_0^j)$, $\widetilde{U}_0^j = U_0^j \sim \mathcal{U}([0,1])$ and these random variables are independent.

Inductive step Let $v \in \mathbf{N}$, and assume the property is true up to step v. Conditionally on $\widetilde{\mathscr{F}}_{v+1}$, *i.e.* on $\widetilde{\mathscr{F}}_v, \widetilde{T}_v, \widetilde{J}_v, \widetilde{U}_v$, we have:

- for $j \neq \widetilde{J}_v$, the variables $\widetilde{T}_{v+1}^j - \widetilde{T}_{v-1} = \widetilde{T}_v^j - \widetilde{T}_{v-1}$ are independent $\mathsf{Exp}(\widetilde{L}_v^j) = \mathsf{Exp}(\widetilde{L}_{v+1}^j)$ random variables (when conditioned only on $\widetilde{\mathscr{F}}_v$, by the induction hypothesis), conditioned on $\widetilde{T}_{v+1}^j - \widetilde{T}_{v-1} \geqslant \widetilde{T}_v - \widetilde{T}_{v-1}$, so by the memory-less property of exponential random variables $\widetilde{T}_{v+1}^j - \widetilde{T}_v = (\widetilde{T}_{v+1}^j - \widetilde{T}_{v-1}) - (\widetilde{T}_v - \widetilde{T}_{v-1}) \sim \mathsf{Exp}(\widetilde{L}_{v+1}^j)$ (and those variables are independent).
- for $j \neq \widetilde{J}_v$, the variables $\widetilde{U}_{v+1}^j = \widetilde{U}_v^j$ are independent $\mathcal{U}([0,1])$ random variables (conditionally on $\widetilde{\mathscr{F}}_v$), conditioned on the independent variables $\widetilde{T}_v, \widetilde{J}_v, \widetilde{U}_v$, so they remain independent $\mathcal{U}([0,1])$ random variables.
- $(\widetilde{T}_{v+1}^{\widetilde{J}_v} - \widetilde{T}_v, \widetilde{U}_{v+1}^{\widetilde{J}_v}) = (E^{\widetilde{J}_v}_{\widetilde{V}_{v+1}^{\widetilde{J}_v}} / \widetilde{L}_{v+1}^{\widetilde{J}_v}, U^{\widetilde{J}_v}_{\widetilde{V}_{v+1}^{\widetilde{J}_v}})$ is distributed, conditionally on $\widetilde{\mathscr{F}}_{v+1}$, *i.e.* on $\widetilde{J}_v, \widetilde{T}_v, \widetilde{V}_{v+1}^{\widetilde{J}_v}, \widetilde{L}_{v+1}^{\widetilde{J}_v}$, as $\mathsf{Exp}(\widetilde{L}_{v+1}^{\widetilde{J}_v}) \otimes \mathcal{U}([0,1])$, and independent of $(\widetilde{T}_{v+1}^j, \widetilde{U}_{v+1}^j)_{j \neq \widetilde{J}_v}$.

This completes the proof by induction.

Let $v \in \mathbf{N}$. We have established that, conditionally on $\mathscr{F}_v$, the variables $(\widetilde{T}_v^j, \widetilde{U}_v^j)_{1 \leqslant j \leqslant d}$ are independent, with $\widetilde{T}_v^j - \widetilde{T}_{v-1} \sim \mathsf{Exp}(\widetilde{L}_v^j)$ and $\widetilde{U}_v^j \sim \mathcal{U}([0,1])$. In particular, conditionally on $\mathscr{F}_v$, $\widetilde{U}_v$ is independent from $(\widetilde{J}_v, \widetilde{T}_v)$, $\widetilde{U}_v \sim \mathcal{U}([0,1])$, and (by the property of the minimum of independent exponential random variables) $\widetilde{J}_v$ is independent of $\widetilde{T}_v$, $\widetilde{T}_v \sim \mathsf{Exp}(\sum_{j=1}^d \widetilde{L}_v^j)$ and $\mathbb{P}(\widetilde{J}_v = j \,|\, \mathscr{F}_v) = \widetilde{L}_v^j / (\sum_{j'=1}^d \widetilde{L}_v^{j'})$. This concludes the proof of Lemma 1. □

4. Proof of Theorem 1. Recall that a Mondrian Forest estimate with lifetime parameter λ, is defined, for all $x \in [0,1]^d$, by

$$\widehat{f}_{\lambda,n,M}(x) = \widehat{f}_{\lambda,n,M}(x, \Pi_{\lambda,M}) = \frac{1}{M} \sum_{m=1}^{M} \widehat{f}_{\lambda,n}^{(m)}(x, \Pi_\lambda^{(m)}),$$

where $\widehat{f}_{\lambda,n}^{(m)}(x, \Pi_\lambda^{(m)})$ denotes the Mondrian Tree based on the random partition $\Pi_\lambda^{(m)}$ and $\Pi_{\lambda,M} = (\Pi_\lambda^{(1)}, \ldots, \Pi_\lambda^{(M)})$. To ease notation, we will write $\widehat{f}_{\lambda,n}^{(m)}(x)$ instead of $\widehat{f}_{\lambda,n}^{(m)}(x, \Pi_\lambda^{(m)})$. First, note that, by Jensen's inequality,

$$\begin{aligned} R(\widehat{f}_{\lambda,n,M}) &= \mathbb{E}_{(X,\Pi_{\lambda,M})}[(\widehat{f}_{\lambda,n,M}(x, \Pi_{\lambda,M}) - f(X))^2] \\ &\leqslant \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{(X,\Pi_\lambda^{(m)})}[(\widehat{f}_{\lambda,n}^{(m)}(X, \Pi_\lambda^{(m)}) - f(X))^2] \\ &\leqslant \mathbb{E}_{(X,\Pi_\lambda^{(1)})}[(\widehat{f}_{\lambda,n}^{(1)}(X, \Pi_\lambda^{(1)}) - f(X))^2], \end{aligned}$$

since each Mondrian tree has the same distribution. Therefore, it is sufficient to prove that a single Mondrian tree is consistent. Now, since Mondrian partitions are independent of the dataset $\mathscr{D}_n$, we can apply Theorem 4.2 from [3], which states that a Mondrian tree estimate is consistent if

(*i*) $D_\lambda(X) \to 0$ in probability, as $n \to \infty$,
(*ii*) $K_\lambda / n \to \infty$ in probability, as $n \to \infty$,

where $D_\lambda(X)$ is the diameter of the cell of the Mondrian tree that contains X, and K_λ is the number of cells in the Mondrian tree. Note that the initial assumptions in Theorem 4.2 in [3] contains deterministic convergence, but can be relaxed to convergences in probability by a close inspection of the proof. In the sequel, we prove that an individual Mondrian tree satisfies (*i*) and (*ii*) which will conclude the proof. To prove (*i*), just note that, according to Corollary 1,

$$\mathbb{E}[D_\lambda(X)^2] = \mathbb{E}[\mathbb{E}[D_\lambda(X)^2 \,|\, X]] \leqslant \frac{4d}{\lambda^2},$$

which tends to zero, since $\lambda = \lambda_n \to \infty$, as $n \to \infty$. Thus, condition (i) holds. Now, to prove (ii), observe that

$$\mathbb{E}\Big[\frac{K_\lambda}{n}\Big] = \frac{(1+\lambda)^d}{n},$$

which tends to zero since $\lambda_n^d/n \to 0$ by assumption, as $n \to \infty$. This concludes the proof of Theorem 1. □

5. Proof of Proposition 3. Let $\Pi_\lambda^{(1)}$ be the Mondrian partition of $[0,1]$ used to construct the randomized estimator $\widehat{f}_{\lambda,n}^{(1)}$. Denote by $\bar{f}_\lambda^{(1)}$ the random function $\bar{f}_\lambda^{(1)}(x) = \mathbb{E}_X\left[f(X) \,|\, X \in C_\lambda(x)\right]$, and define $\widetilde{f}_\lambda(x) = \mathbb{E}\left[\bar{f}_\lambda^{(1)}(x)\right]$ (which is deterministic). For the seek of clarity, we will drop the exponent "(1)" in all notations, keeping in mind that we consider only one particular Mondrian partition, whose associated Mondrian Tree estimate is denoted by $\widehat{f}_{\lambda,n}$. Recall the bias-variance decomposition (7.4) for Mondrian trees:

$$R(\widehat{f}_{\lambda,n}^{(1)}) = \mathbb{E}\big[(f(X) - \bar{f}_\lambda(X))^2\big] + \mathbb{E}\big[(\bar{f}_\lambda(X) - \widehat{f}_{\lambda,n}^{(1)}(X))^2\big]\,. \tag{5.1}$$

We will provide lower bounds for the first term (the bias, depending on λ) and the second (the variance, depending on both λ and n), which will lead to the stated lower bound on the risk, valid for every value of λ.

Lower bound on the bias. As we will see, the point-wise bias $\mathbb{E}[(\bar{f}_\lambda(x) - f(x))^2]$ can be computed explicitly given our assumptions. Let $x \in [0,1]$. Since $\widetilde{f}_\lambda(x) = \mathbb{E}[\bar{f}_\lambda(x)]$, we have

$$\mathbb{E}\left[(\bar{f}_\lambda(x) - f(x))^2\right] = \mathrm{Var}(\bar{f}_\lambda(x)) + (\widetilde{f}_\lambda(x) - f(x))^2\,. \tag{5.2}$$

By Proposition 1, the cell of x in Π_λ can be written as $C_\lambda(x) = [L_\lambda(x), R_\lambda(x)]$, with $L_\lambda(x) = (x - \lambda^{-1}E_L) \vee 0$ and $R_\lambda(x) = (x + \lambda^{-1}E_R) \wedge 1$, where E_L, E_R are two independent $\mathsf{Exp}(1)$ random variables. Now, since $X \sim \mathcal{U}([0,1])$ and $f(u) = 1 + u$,

$$\bar{f}_\lambda(x) = \frac{1}{R_\lambda(x) - L_\lambda(x)} \int_{L_\lambda(x)}^{R_\lambda(x)} (1+u)\mathrm{d}u = 1 + \frac{L_\lambda(x) + R_\lambda(x)}{2}\,.$$

Since $L_\lambda(x)$ and $R_\lambda(x)$ are independent, we have

$$\mathrm{Var}(\bar{f}_\lambda(x)) = \frac{\mathrm{Var}(L_\lambda(x)) + \mathrm{Var}(R_\lambda(x))}{4}.$$

In addition,

$$\mathrm{Var}(R_\lambda(x)) = \mathrm{Var}\big(x + \lambda^{-1}[E_R \wedge \lambda(1-x)]\big) = \lambda^{-2}\mathrm{Var}(E_R \wedge [\lambda(1-x)])$$

Now, if $E \sim \mathsf{Exp}(1)$ and $a \geqslant 0$, we have

$$\mathbb{E}[E \wedge a] = \int_0^a u e^{-u} \mathrm{d}u + a\mathbb{P}(E \geqslant a) = 1 - e^{-a} \tag{5.3}$$
$$\mathbb{E}[(E \wedge a)^2] = \int_0^a u^2 e^{-u} \mathrm{d}u + a^2\mathbb{P}(E \geqslant a) = 2\left(1 - (a+1)e^{-a}\right),$$

so that

$$\mathrm{Var}(E \wedge a) = \mathbb{E}[(E \wedge a)^2] - \mathbb{E}[E \wedge a]^2 = 1 - 2ae^{-a} - e^{-2a}.$$

The formula above gives the variances of $R_\lambda(x)$ and $L_\lambda(x)$ respectively:

$$\mathrm{Var}(R_\lambda(x)) = \lambda^{-2}\left(1 - 2\lambda(1-x)e^{-\lambda(1-x)} - e^{-2\lambda(1-x)}\right)$$
$$\mathrm{Var}(L_\lambda(x)) = \lambda^{-2}\left(1 - 2\lambda x e^{-\lambda x} - e^{-2\lambda x}\right),$$

and thus

$$\mathrm{Var}(\bar{f}_\lambda(x)) = \frac{1}{4\lambda^2}\left(2 - 2\lambda x e^{-\lambda x} - 2\lambda(1-x)e^{-\lambda(1-x)} - e^{-2\lambda x} - e^{-2\lambda(1-x)}\right). \tag{5.4}$$

In addition, the formula (5.3) yields

$$\mathbb{E}[R_\lambda(x)] = x + \lambda^{-1}\left(1 - e^{-\lambda(1-x)}\right)$$
$$\mathbb{E}[L_\lambda(x)] = x - \lambda^{-1}\left(1 - e^{-\lambda x}\right),$$

and thus

$$\widetilde{f}_\lambda(x) = 1 + \frac{\mathbb{E}[L_\lambda(x)] + \mathbb{E}[R_\lambda(x)]}{2} = 1 + x + \frac{1}{2\lambda}\left(e^{-\lambda x} - e^{-\lambda(1-x)}\right). \tag{5.5}$$

Combining (5.4) and (5.5) with the decomposition (5.2) gives

$$\mathbb{E}\left[\left(\bar{f}_\lambda(x) - f(x)\right)^2\right] = \frac{1}{2\lambda^2}\left(1 - \lambda x e^{-\lambda x} - \lambda(1-x)e^{-\lambda(1-x)} - e^{-\lambda}\right). \tag{5.6}$$

Integrating over X, we obtain

$$\begin{aligned}
&\mathbb{E}\left[(\bar{f}_\lambda(X) - f(X))^2\right] \\
&= \frac{1}{2\lambda^2}\left(1 - \int_0^1 \lambda x e^{-\lambda x}\mathrm{d}x - \int_0^1 \lambda(1-x)e^{-\lambda(1-x)}\mathrm{d}x - e^{-\lambda}\right) \\
&= \frac{1}{2\lambda^2}\left(1 - 2 \times \frac{1}{\lambda}\left(1 - (\lambda+1)e^{-\lambda}\right) - e^{-\lambda}\right) \\
&= \frac{1}{2\lambda^2}\left(1 - \frac{2}{\lambda} + e^{-\lambda} + \frac{2}{\lambda}e^{-\lambda}\right). \qquad (5.7)
\end{aligned}$$

Now, note that the bias $\mathbb{E}[(\bar{f}_\lambda(X)-f(X))^2]$ is positive for $\lambda \in \mathbf{R}_+^*$ (indeed, it is nonnegative, and non-zero since f is not piecewise constant). In addition, the expression (5.7) shows that it is continuous in λ on $\mathbf{R}_+^*$, and that it admits a limit $\frac{1}{12}$ as $\lambda \to 0$ (using the fact that $e^{-\lambda} = 1-\lambda+\frac{\lambda^2}{2}-\frac{\lambda^3}{6}+o(\lambda^3)$). Hence, the function $\lambda \mapsto \mathbb{E}[(\bar{f}_\lambda(X) - f(X))^2]$ is positive and continuous on $\mathbf{R}_+$, so that it admits a minimum $C_1 > 0$ on the compact interval $[0, 6]$. In addition, the expression (5.7) shows that for $\lambda \geqslant 6$, we have

$$\mathbb{E}\left[(\bar{f}_\lambda(X) - f(X))^2\right] \geqslant \frac{1}{2\lambda^2}\left(1 - \frac{2}{6}\right) = \frac{1}{3\lambda^2}\,. \tag{5.8}$$

First lower bound on the variance. We now turn to the task of bounding the variance from below. In order to avoid restrictive conditions on λ, we will provide two separate lower bounds, valid in two different regimes.

Our first lower bound on the variance, valid for $\lambda \leqslant n/3$, controls the error of estimation of the optimal labels in nonempty cells. It depends on σ^2, and is of order $\Theta\big(\sigma^2 \frac{\lambda}{n}\big)$. We use a general bound on the variance of regressograms [1, Proposition 2] (note that while this result is stated for a fixed number of cells, it can be adapted to a random number of cells by conditioning on $K_\lambda = k$ and then by averaging):

$$\begin{aligned}&\mathbb{E}\left[\left(\widehat{f}_{\lambda,n}(X) - \widetilde{f}_\lambda(X)\right)^2\right] \\ &\geqslant \frac{\sigma^2}{n}\left(\mathbb{E}\left[K_\lambda\right] - 2\mathbb{E}_{\Pi_\lambda}\left[\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \exp(-n\mathbb{P}(X \in C_\mathbf{v}))\right]\right).\end{aligned} \tag{5.9}$$

Now, recall that the splits defining Π_λ form a Poisson point process on $[0, 1]$ of intensity $\lambda \mathrm{d}x$ (Fact 1). In particular, the splits can be described as follows. Let $(E_k)_{k\geqslant 1}$ be an i.i.d. sequence of $\mathsf{Exp}(1)$ random variables, and $S_p := \sum_{k=1}^p E_k$ for $p \geqslant 0$. Then, the (ordered) splits in Π_λ have the same distribution as $(\lambda^{-1}S_1, \ldots, \lambda^{-1}S_{K_\lambda - 1})$, where $K_\lambda := 1 + \sup\{p \geqslant 0 : S_p \leqslant \lambda\}$. In addition, the probability that $X \sim \mathcal{U}([0, 1])$ falls in the cell

$[\lambda^{-1}S_{k-1}, \lambda^{-1}S_k \wedge 1)$ $(1 \leqslant k \leqslant K_\lambda)$ is $\lambda^{-1}(S_k \wedge 1 - S_{k-1})$, so that

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \exp(-n\mathbb{P}\left(X \in C_{\mathbf{v}}\right))\Big] \\
&= \mathbb{E}\Big[\sum_{k=1}^{K_\lambda - 1} e^{-n\lambda^{-1}(S_k - S_{k-1})} + e^{-n(1-\lambda^{-1}S_{K_\lambda - 1})}\Big] \\
&\leqslant \mathbb{E}\Big[\sum_{k=1}^{\infty} \mathbf{1}(S_k \leqslant \lambda) e^{-n\lambda^{-1}E_k}\Big] + 1 \\
&= \sum_{k=1}^{\infty} \mathbb{E}\big[\mathbf{1}(S_k \leqslant \lambda)\big]\mathbb{E}\big[e^{-n\lambda^{-1}E_k}\big] + 1 \qquad (5.10) \\
&= \sum_{k=1}^{\infty} \mathbb{E}\big[\mathbf{1}(S_k \leqslant \lambda)\big] \cdot \int_0^\infty e^{-n\lambda^{-1}u} e^{-u} \mathrm{d}u + 1 \\
&= \frac{\lambda}{n+\lambda}\mathbb{E}\Big[\sum_{k=1}^{\infty} \mathbf{1}(S_k \leqslant \lambda)\Big] + 1 \\
&= \frac{\lambda}{n+\lambda}\mathbb{E}\left[K_\lambda\right] + 1 \\
&= \frac{\lambda}{n+\lambda}(1+\lambda) + 1 \qquad (5.11)
\end{aligned}$$

where (5.10) comes from the fact that E_k and S_{k-1} are independent. Plugging Equation (5.11) in the lower bound (5.9) yields

$$\begin{aligned}
\mathbb{E}\left[\big(\widehat{f}_{\lambda,n}(X) - \widetilde{f}_\lambda(X)\big)^2\right] &\geqslant \frac{\sigma^2}{n}\left((1+\lambda) - 2(1+\lambda)\frac{\lambda}{n+\lambda} - 2\right) \\
&= \frac{\sigma^2}{n}\left((1+\lambda)\frac{n-\lambda}{n+\lambda} - 2\right).
\end{aligned}$$

Now, assume that $6 \leqslant \lambda \leqslant \frac{n}{3}$. Since

$$(1+\lambda)\frac{n-\lambda}{n+\lambda} - 2 \underset{(\lambda\leqslant n/3)}{\geqslant} (1+\lambda)\frac{n-n/3}{n+n/3} - 2 = (1+\lambda)\frac{1}{2} - 2 \underset{(\lambda\geqslant 6)}{\geqslant} \frac{\lambda}{4},$$

the above lower bound implies, for $6 \leqslant \lambda \leqslant \frac{n}{3}$,

$$\mathbb{E}\left[\big(\widehat{f}_{\lambda,n}(X) - \widetilde{f}_\lambda(X)\big)^2\right] \geqslant \frac{\sigma^2\lambda}{4n}. \qquad (5.12)$$

Second lower bound on the variance. The lower bound (5.12) is only valid for $\lambda \leqslant n/3$; as λ becomes of order n or larger, the previous bound becomes vacuous. We now provide another lower bound on the variance, valid when $\lambda \geqslant n/3$, by considering the contribution of empty cells to the variance.

Let $\mathbf{v} \in \mathcal{L}(\Pi_\lambda)$. If $C_\mathbf{v}$ contains no sample point from $\mathscr{D}_n$, then for $x \in C_\mathbf{v}$: $\widehat{f}_{\lambda,n}(x) = 0$ and thus $(\widehat{f}_{\lambda,n}(x) - \bar{f}_\lambda(x))^2 = \bar{f}_\lambda(x)^2 \geqslant 1$. Hence, the variance term is lower bounded as follows, denoting $N_n(C)$ the number of $1 \leqslant i \leqslant n$ such that $X_i \in C$ and $N_{\lambda,n}(x) = N_n(C_\lambda(x))$:

$$
\begin{aligned}
\mathbb{E}\big[(\widehat{f}_{\lambda,n}(X) - \bar{f}_\lambda(X))^2\big] &\geqslant \mathbb{P}\big(N_{\lambda,n}(X) = 0\big) \\
&= \mathbb{E}\bigg[\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \mathbb{P}(X \in C_\mathbf{v})\,\mathbb{P}(N_n(C_\mathbf{v}) = 0)\bigg] \\
&= \mathbb{E}\bigg[\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \mathbb{P}(X \in C_\mathbf{v})\big(1 - \mathbb{P}(X \in C_\mathbf{v})\big)^n\bigg] \\
&\geqslant \mathbb{E}\bigg[\Big(\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \mathbb{P}(X \in C_\mathbf{v})\big(1 - \mathbb{P}(X \in C_\mathbf{v})\big)\Big)^n\bigg] \qquad (5.13) \\
&\geqslant \mathbb{E}\bigg[\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \mathbb{P}(X \in C_\mathbf{v})\big(1 - \mathbb{P}(X \in C_\mathbf{v})\big)\bigg]^n \qquad (5.14) \\
&= \bigg(1 - \mathbb{E}\bigg[\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \mathbb{P}(X \in C_\mathbf{v})^2\bigg]\bigg)^n \qquad (5.15)
\end{aligned}
$$

where (5.13) and (5.14) come from Jensen's inequality applied to the convex function $x \mapsto x^n$. Now, using the notations defined above, we have

$$
\begin{aligned}
\mathbb{E}\bigg[\sum_{\mathbf{v}\in\Pi_\lambda} \mathbb{P}(X \in C_\mathbf{v})^2\bigg] &\leqslant \mathbb{E}\bigg[\sum_{k=1}^{K_\lambda} (\lambda^{-1}E_k)^2\bigg] \\
&= \lambda^{-2}\,\mathbb{E}\bigg[\sum_{k=1}^{\infty} \mathbf{1}(S_{k-1} \leqslant \lambda) E_k^2\bigg] \\
&= \lambda^{-2}\,\mathbb{E}\bigg[\sum_{k=1}^{\infty} \mathbf{1}(S_{k-1} \leqslant \lambda)\mathbb{E}\big[E_k^2 \,|\, S_{k-1}\big]\bigg] \\
&= 2\lambda^{-2}\,\mathbb{E}\bigg[\sum_{k=1}^{\infty} \mathbf{1}(S_{k-1} \leqslant \lambda)\bigg] \qquad (5.16) \\
&= 2\lambda^{-2}\,\mathbb{E}\,[K_\lambda] \\
&= \frac{2(\lambda+1)}{\lambda^2}\,, \qquad (5.17)
\end{aligned}
$$

where the equality $\mathbb{E}[E_k^2 \,|\, S_{k-1}] = 2$ (used in Equation (5.16)) comes from the fact that $E_k \sim \mathsf{Exp}(1)$ is independent of S_{k-1}.

The bounds (5.15) and (5.17) imply that, if $2(\lambda+1)/\lambda^2 \leqslant 1$, then

$$\mathbb{E}\big[(\widehat{f}_{\lambda,n}(X) - \bar{f}_\lambda(X))^2\big] \geqslant \left(1 - \frac{2(\lambda+1)}{\lambda^2}\right)^n . \tag{5.18}$$

Now, assume that $n \geqslant 18$ and $\lambda \geqslant \frac{n}{3} \geqslant 6$. Then

$$\frac{2(\lambda+1)}{\lambda^2} \leqslant 2 \cdot \frac{3}{n}\left(1 + \frac{3}{n}\right) \leqslant 2 \cdot \frac{3}{n}\left(1 + \frac{3}{18}\right) = \frac{7}{n} \underset{(n \geqslant 18)}{\leqslant} 1 ,$$

so that, using the inequality $(1-x)^m \geqslant 1 - mx$ for $m \geqslant 0$ and $x \in \mathbf{R}$,

$$\left(1 - \frac{2(\lambda+1)}{\lambda^2}\right)^{n/8} \geqslant \left(1 - \frac{7}{n}\right)^{n/8} \geqslant 1 - \frac{n}{8} \cdot \frac{7}{n} = \frac{1}{8} .$$

Combining the above inequality with (5.18) gives, letting $C_2 := 1/8^8$,

$$\mathbb{E}\big[(\widehat{f}_{\lambda,n}(X) - \bar{f}_\lambda(X))^2\big] \geqslant C_2 . \tag{5.19}$$

Summing up. Assume that $n \geqslant 18$. Recall the bias-variance decomposition (5.1) of the risk $R(\widehat{f}_{\lambda,n})$ of the Mondrian tree.

- If $\lambda \leqslant 6$, we saw that the bias (and hence the risk) is larger than C_1;
- If $\lambda \geqslant \frac{n}{3}$, Equation (5.18) implies that the variance (and hence the risk) is larger than C_2;
- If $6 \leqslant \lambda \leqslant \frac{n}{3}$, Equations (5.8) (bias term) and (5.12) (variance term) imply that
$$R(\widehat{f}_{\lambda,n}) \geqslant \frac{1}{3\lambda^2} + \frac{\sigma^2\lambda}{4n} .$$

In particular,

$$\inf_{\lambda \in \mathbf{R}^+} R(\widehat{f}_{\lambda,n}) \geqslant C_1 \wedge C_2 \wedge \inf_{\lambda \in \mathbf{R}^+}\left(\frac{1}{3\lambda^2} + \frac{\sigma^2\lambda}{4n}\right) = C_0 \wedge \frac{1}{4}\left(\frac{3\sigma^2}{n}\right)^{2/3} \tag{5.20}$$

where we let $C_0 = C_1 \wedge C_2$.

6. Proof of Proposition 4. First, note that in all cases, since $|Y| \leqslant B$ almost surely, we also have $|\widehat{g}_n(X)| \leqslant B$ almost surely, so that $(Y - \widehat{g}_n(X))^2 \leqslant 4B^2$. Let $N_\varepsilon = |I_\varepsilon|$. Note that N_ε is a binomial variable with parameters $n - n_0 \geqslant n/2$ and $\mathbb{P}(X \in B_\varepsilon) \geqslant p_0(1-2\varepsilon)^d$ (since $p \geqslant p_0$). Now, recall Chernoff's bound: if $N \sim \mathsf{Bin}(m,p)$ and $\delta \in (0,1)$, then $\mathbb{P}(N \leqslant$

$(1-\delta)mq) \leqslant e^{-mq\delta^2/2}$; in particular, $\mathbb{P}(N \leqslant mq/2) \leqslant e^{-mq/8}$. Hence, letting $c_1 = p_0(1-2\varepsilon)^d/4$,

$$\mathbb{P}\big(N_\varepsilon \leqslant c_1 n\big) \leqslant \exp(-c_1 n/4)\,. \tag{6.1}$$

Conditionally on I_ε, the sample $\mathscr{D}' = \{(X_i, Y_i) : i \in I_\varepsilon\}$ is an i.i.d. sample of size N_ε of the conditional distribution of (X, Y) given $X \in B_\varepsilon$; it is also independent of $\mathscr{D}_{n_0}$, and thus of the estimators $\widehat{f}_\alpha$, $\alpha = 0, \dots, A$. It follows from Theorem 1 in the supplementary material "Proof of the optimality of the empirical star algorithm" of [2] that the estimator $\widehat{g}_n$ defined by (5.6) satisfies, with probability $1-\delta$ over the random sample $\mathscr{D}'$ conditionally on N_ε,

$$\begin{aligned}\mathbb{E}_{(X,Y)}\big[(\widehat{g}_n(X) - Y)^2 \,|\, X \in B_\varepsilon\big] - \min_{0\leqslant\alpha\leqslant A} \mathbb{E}_{(X,Y)}\big[(\widehat{f}_\alpha(X) - Y)^2 \,|\, X \in B_\varepsilon\big] \\ \leqslant \frac{CB^2 \log[(A+1)\delta^{-1}]}{N_\varepsilon}\end{aligned} \tag{6.2}$$

for every $\delta \in (0,1)$, where $C = 600$ and the expectation is taken with respect to an independent sample (X, Y) (the bound (6.2) is deduced from the aforementioned theorem by replacing Y by Y/B, which lies in $[-1,1]$). Since $Y = f(X) + \varepsilon$ with $\mathbb{E}[\varepsilon|X] = 0$, we have $\mathbb{E}[(g(X)-Y)^2|X] = \mathbb{E}[(g(X) - f(X))^2|X] + \mathbb{E}[\varepsilon^2 \,|\, X]$. Hence, inequality (6.2) writes

$$\begin{aligned}&\mathbb{E}_{(X,Y)}\big[(\widehat{g}_n(X) - f(X))^2 \,|\, X \in B_\varepsilon\big] \\ &\leqslant \min_{0\leqslant\alpha\leqslant A} \mathbb{E}_{(X,Y)}\big[(\widehat{f}_\alpha(X) - f(X))^2 \,|\, X \in B_\varepsilon\big] + \frac{CB^2 \log[(A+1)\delta^{-1}]}{N_\varepsilon}.\end{aligned}$$

By integrating the above inequality over the confidence level δ, we obtain

$$\begin{aligned}&\mathbb{E}_{(X,Y),\mathscr{D}'}\big[(\widehat{g}_n(X) - f(X))^2 \,|\, X \in B_\varepsilon, N_\varepsilon\big] \\ &\leqslant \min_{0\leqslant\alpha\leqslant A} \mathbb{E}_{(X,Y)}\big[(\widehat{f}_\alpha(X) - f(X))^2 \,|\, X \in B_\varepsilon\big] + \frac{CB^2[\log(A+1)+1]}{N_\varepsilon};\end{aligned}$$

by taking the expectation over $\mathscr{D}_{n_0}$, conditioning on $N_\varepsilon > c_1 n$, and recalling that $A \leqslant \log_2(n)$, we get

$$\begin{aligned}&\mathbb{E}\big[(\widehat{g}_n(X) - f(X))^2 \,|\, X \in B_\varepsilon, N_\varepsilon > c_1 n\big] \\ &\quad\leqslant \min_{0\leqslant\alpha\leqslant A} \mathbb{E}\big[(\widehat{f}_\alpha(X) - f(X))^2 \,|\, X \in B_\varepsilon\big] + \frac{CB^2[\log(1+\log_2 n)+1]}{c_1 n}.\end{aligned} \tag{6.3}$$

Finally, combining the bounds (6.1) and (6.3) yields

$$\begin{aligned}
&\mathbb{E}\big[(\widehat{g}_n(X) - f(X))^2 \,|\, X \in B_\varepsilon\big] \\
&\leqslant \mathbb{P}\left(N_\varepsilon \leqslant c_1 n\right) \cdot 4B^2 + \mathbb{E}\big[(\widehat{g}_n(X) - f(X))^2 \,|\, X \in B_\varepsilon, N_\varepsilon > c_1 n\big] \\
&\leqslant 4B^2 e^{-c_1 n/4} + \min_{0 \leqslant \alpha \leqslant A} \mathbb{E}\big[(\widehat{f}_\alpha(X) - f(X))^2 \,|\, X \in B_\varepsilon\big] \\
&\qquad + \frac{CB^2[\log(1+\log_2 n)+1]}{c_1 n}\,,
\end{aligned} \tag{6.4}$$

which is precisely inequality (5.7).

Assume that f belongs to the class $\mathscr{C}^{p,\beta}(L)$, with $p \in \{0,1\}$, $\beta \in (0,1]$ and $L > 0$; we now proceed to show that $\widehat{g}_n$ achieves the minimax rate of estimation for this class. Let $s = p + \beta \in (0,2]$. If $p = 0$ (namely, $s \leqslant 1$), it follows from Theorem 2 (with the same adaptation as in the proof of Theorem 3 to bound the variance term conditionally on $X \in B_\varepsilon$) that, for every $\lambda > 0$,

$$\mathbb{E}\big[(\widehat{f}_{\lambda,n_0,M}(X) - f(X))^2 \,|\, X \in B_\varepsilon\big] \leqslant \frac{(4d)^s L^2}{\lambda^{2s}} + \frac{11B^2(1+\lambda)^d}{p_0(1-2\varepsilon)^d n_0}$$

(note that $\sigma, \|f\|_\infty \leqslant B$ since $|Y| \leqslant B$). It follows that, for some constants C_1, C_2 independent of λ, L, n,

$$\begin{aligned}
\min_{0 \leqslant \alpha \leqslant A} \mathbb{E}\big[(\widehat{f}_\alpha(X) - f(X))^2 \,|\, X \in B_\varepsilon\big] &\leqslant \min_{0 \leqslant \alpha \leqslant A} \left[\frac{C_1 L^2}{(2^\alpha)^{2s}} + \frac{C_2(1+2^\alpha)^d}{n}\right] \\
&\leqslant 4 \min_{\lambda \in [1, n^{1/d}]} \left[\frac{C_1 L^2}{\lambda^{2s}} + \frac{C_2(1+\lambda)^d}{n}\right],
\end{aligned} \tag{6.5}$$

where we used the fact that, for every $\lambda \in [1, n^{1/d}]$, there exists some α, $0 \leqslant \alpha \leqslant A$, such that $\lambda/2 \leqslant 2^\alpha \leqslant \lambda$. It follows from (6.4) and (6.5) that

$$\begin{aligned}
&\mathbb{E}\big[(\widehat{g}_n(X) - f(X))^2 \,|\, X \in B_\varepsilon\big] \\
&\leqslant O\Big(\min_{0 \leqslant \lambda \leqslant n^{1/d}} \left[\frac{C_1 L^2}{\lambda^{2s}} + \frac{C_2(1+\lambda)^d}{n}\right] + \frac{\log\log n}{n}\Big) \\
&= O\left(L^{2d/(d+2s)} n^{-2s/(d+2s)}\right)
\end{aligned}$$

where the last bound follows from the fact that $\lambda_* = (L^2 n)^{1/(d+2s)}$ belongs to $[1, n^{1/d}]$ for n large enough (and $\log\log n / n = o(n^{2s/(d+2s)})$).

Now, consider the case $p = 1$, *i.e.*, $1 < s \leqslant 2$. It follows from Theorem 3 that for some constants C_3, C_4 independent of λ, L, n, we have for every

$\lambda \in [1, n^{1/d}]$ (using the fact that $M \geqslant n^{2/d} \geqslant \lambda^2$, so that $1/(M\lambda^2) \leqslant 1/\lambda^4 \leqslant 1/\lambda^{2s}$, and $e^{-\lambda\varepsilon}/\lambda^3 = O(1/\lambda^{2s})$)

$$\mathbb{E}\big[(\widehat{f}_{\lambda,n,M}(X) - f(X))^2 \,|\, X \in B_\varepsilon\big] \leqslant \frac{C_3 L^2}{\lambda^{2s}} + \frac{C_4(1+\lambda)^d}{n}\,. \tag{6.6}$$

From the same argument as in the case $0 < s \leqslant 1$, combining inequalities (6.6) and (6.4) yields

$$\mathbb{E}\big[(\widehat{g}_n(X) - f(X))^2 \,|\, X \in B_\varepsilon\big] = O\big(L^{2d/(d+2s)} n^{-2s/(d+2s)}\big)$$

which concludes the proof of Proposition 4. □

7. Proof of Lemma 1. According to Equation (7.15) from the main text, we have

$$F_\lambda(x,z) = \lambda^d \exp(-\lambda\|x-z\|_1) \prod_{1\leqslant j\leqslant d} G_\lambda(x_j, z_j) \tag{7.1}$$

where we defined, for $u, v \in [0,1]$,

$$\begin{aligned} G_\lambda(u,v) &= \mathbb{E}\left[(\lambda|u-v| + E_1 \wedge \lambda(u\wedge v) + E_2 \wedge \lambda(1-u\vee v))^{-1}\right] \\ &= H(\lambda|u-v|, \lambda u\wedge v, \lambda(1-u\vee v)) \end{aligned}$$

with E_1, E_2 two independent $\mathsf{Exp}(1)$ random variables, and $H : (\mathbf{R}_+^*)^3 \to \mathbf{R}$ the function defined by

$$H(a,b_1,b_2) = \mathbb{E}\left[(a + E_1 \wedge b_1 + E_2 \wedge b_2)^{-1}\right]\,;$$

also, let

$$H(a) = \mathbb{E}\left[(a + E_1 + E_2)^{-1}\right].$$

Denote

$$\begin{aligned} A &= \int_{[0,1]^d} (z-x) F_\lambda(x,z)\mathrm{d}z \\ B &= \int_{[0,1]^d} \frac{1}{2}\|z-x\|^2 F_\lambda(x,z)\mathrm{d}z. \end{aligned}$$

Since $1 = \int F_\lambda^{(1)}(u,v)\mathrm{d}v = \int \lambda\exp(-\lambda|u-v|)G_\lambda(u,v)\mathrm{d}v$, applying Fubini's theorem we obtain

$$A_j = \Phi_\lambda^1(x_j) \qquad \text{and} \qquad B = \sum_{j=1}^d \Phi_\lambda^2(x_j) \tag{7.2}$$

where we define for $u \in [0,1]$ and $k \in \mathbf{N}$

$$\Phi_\lambda^k(u) = \int_0^1 \lambda \exp(-\lambda|u-v|) G_\lambda(u,v) \frac{(v-u)^k}{k!} \mathrm{d}v \,. \tag{7.3}$$

Observe that

$$\Phi_\lambda^k(u) = \lambda^{-k} \int_{-\lambda u}^{\lambda(1-u)} \frac{v^k}{k!} \exp(-|v|) H(|v|, \lambda u + v \wedge 0, \lambda(1-u) - v \vee 0) \mathrm{d}v \,.$$

We will control $\Phi_\lambda^k(u)$ for $k = 1, 2$. First, write

$$\begin{aligned}\lambda \Phi_\lambda^1(u) = &- \int_0^{\lambda u} v e^{-v} H(v, \lambda u - v, \lambda(1-u)) \mathrm{d}v \\ &+ \int_0^{\lambda(1-u)} v e^{-v} H(v, \lambda u, \lambda(1-u) - v) \mathrm{d}v\end{aligned}$$

Now, let $\beta := \lambda \frac{u \wedge (1-u)}{2}$. We have

$$\begin{aligned}&\lambda \Phi_\lambda^1(u) - \int_0^\beta v e^{-v} \left[H(v, \lambda u, \lambda(1-u) - v) - H(v, \lambda u - v, \lambda(1-u)) \right] \mathrm{d}v = \\ &- \underbrace{\int_\beta^{\lambda u} v e^{-v} H(v, \lambda u - v, \lambda(1-u)) \mathrm{d}v}_{:= I_1 \geqslant 0} + \underbrace{\int_\beta^{\lambda(1-u)} v e^{-v} H(v, \lambda u, \lambda(1-u) - v) \mathrm{d}v}_{:= I_2 \geqslant 0}\end{aligned}$$

so that the left-hand side of the above equation is between $-I_1 \leqslant 0$ and $I_2 \geqslant 0$, and thus its absolute value is bounded by $|I_1| \vee |I_2|$. Now, note that, since $H(v, \cdot, \cdot) \leqslant v^{-1}$, we have

$$|I_2| \leqslant \int_\beta^\infty v e^{-v} v^{-1} \mathrm{d}v = e^{-\beta}$$

and similarly $|I_1| \leqslant e^{-\beta}$, so that

$$\begin{aligned}&\left| \lambda \Phi_\lambda^1(u) - \underbrace{\int_0^\beta v e^{-v} \left[H(v, \lambda u, \lambda(1-u) - v) - H(v, \lambda u - v, \lambda(1-u)) \right] \mathrm{d}v}_{:= I_3} \right| \\ &\leqslant e^{-\beta}\end{aligned} \tag{7.4}$$

It now remains to bound $|I_3|$. For that purpose, note that since H is decreasing in its second and third argument, we have

$$\begin{aligned}
&H(v) - H(v, \lambda u - v, \lambda(1-u)) \\
&\leqslant H(v, \lambda u, \lambda(1-u) - v) - H(v, \lambda u - v, \lambda(1-u)) \\
&\leqslant H(v, \lambda u, \lambda(1-u) - v) - H(v)
\end{aligned}$$

which implies

$$\begin{aligned}
&|H(v, \lambda u, \lambda(1-u) - v) - H(v, \lambda u - v, \lambda(1-u))| \\
&\leqslant \max(|H(v, \lambda u, \lambda(1-u) - v) - H(v)|, |H(v) - H(v, \lambda u - v, \lambda(1-u))|).
\end{aligned}$$

Besides, since $(a + E_1 \wedge b_1 + E_2 \wedge b_2)^{-1} \leqslant (a + E_1 + E_2)^{-1} + a^{-1}(\mathbf{1}\{E_1 \geqslant b_1\} + \mathbf{1}\{E_2 \geqslant b_2\})$,

$$H(a, b_1, b_2) - H(a) \leqslant a^{-1}(e^{-b_1} + e^{-b_2}), \tag{7.5}$$

for all a, b_1, b_2. Since $\lambda u - v \geqslant \beta$ and $\lambda(1-u) - v \geqslant \beta$ for $v \in [0, \beta]$, we have

$$|H(v) - H(v, \lambda u - v, \lambda(1-u))|, |H(v) - H(v, \lambda u, \lambda(1-u) - v)| \leqslant 2v^{-1}e^{-\beta}$$

so that for $v \in [0, \beta]$

$$|H(v, \lambda u, \lambda(1-u) - v) - H(v, \lambda u - v, \lambda(1-u))| \leqslant 2v^{-1}e^{-\beta}$$

and hence

$$\begin{aligned}
|I_3| &\leqslant \int_0^\beta v e^{-v} \, |H(v, \lambda u, \lambda(1-u) - v) - H(v, \lambda u - v, \lambda(1-u))| \, \mathrm{d}v \\
&\leqslant \int_0^\beta v e^{-v} 2v^{-1} e^{-\beta} \mathrm{d}v \\
&\leqslant 2e^{-\beta} \int_0^\infty e^{-v} \mathrm{d}v \\
&= 2e^{-\beta} \qquad (7.6)
\end{aligned}$$

Combining Equations (7.4) and (7.6) yields:

$$|\Phi_\lambda^1(u)| \leqslant \frac{3}{\lambda} e^{-\lambda[u \wedge (1-u)]/2} \tag{7.7}$$

that is,

$$\left\| \int_{[0,1]^d} (z - x) F_\lambda(x, z) \mathrm{d}z \right\|^2 = \sum_{j=1}^d \left(\Phi_\lambda^1(x_j)\right)^2 \leqslant \frac{9}{\lambda^2} \sum_{j=1}^d e^{-\lambda[x_j \wedge (1-x_j)]} .$$

Furthermore,

$$\begin{aligned}0 \leqslant \Phi_\lambda^2(u) &= \lambda^{-2} \int_{-\lambda u}^{\lambda(1-u)} \frac{v^2}{2} e^{-|v|} H(|v|, \lambda u + v \wedge 0, \lambda(1-u) - v \vee 0) \mathrm{d}v \\ &\leqslant \lambda^{-2} \int_0^\infty v^2 e^{-v} v^{-1} \mathrm{d}v \\ &= \lambda^{-2}\end{aligned}$$

so that

$$0 \leqslant \Phi_\lambda^2(u) \leqslant \frac{1}{\lambda^2},$$

which proves the second inequality by summing over $j = 1, \ldots, d$. This concludes the proof of Lemma 1. □

References.

[1] S. Arlot and R. Genuer. Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*, 2014.

[2] J.-Y. Audibert. Progressive mixture rules are deviation suboptimal. In *Advances in Neural Information Processing Systems 20*, pages 41–48. Curran Associates, Inc., 2008.

[3] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2002.

Jaouad Mourtada
CMAP, Ecole Polytechnique
Route de Saclay
91128 Palaiseau cedex, France
E-mail: jaouad.mourtada@polytechnique.edu

Stéphane Gaïffas
LPMA - Univ. Paris Diderot
Bâtiment Sophie Germain
Case courrier 7012
75205 PARIS CEDEX 13, France
E-mail: stephane.gaiffas@lpsm.paris

Erwan Scornet
CMAP, Ecole Polytechnique
Route de Saclay
91128 Palaiseau cedex, France
E-mail: erwan.scornet@polytechnique.edu