Mondrian Random forests: theory and methodology

Chapter 2

Minimax optimal rates for Mondrian trees and forests

Abstract. Introduced by Breiman (2001a), Random Forests are widely used classification and regression algorithms. While being initially designed as batch algorithms, several variants have been proposed to handle online learning. One particular instance of such forests is the *Mondrian Forest* Lakshminarayanan et al. (2014, 2016), whose trees are built using the so-called Mondrian process, therefore allowing to easily update their construction in a streaming fashion. In this chapter, we provide a thorough theoretical study of Mondrian Forests in a batch learning setting, based on new results about Mondrian partitions. Our results include consistency and convergence rates for Mondrian Trees and Forests, that turn out to be minimax optimal on the set of s-Hölder function with $s \in (0, 1]$ (for trees and forests) and $s \in (1, 2]$ (for forests only), assuming a proper tuning of their complexity parameter in both cases. Furthermore, we prove that an adaptive procedure (to the unknown $s \in (0, 2]$) can be constructed by combining Mondrian Forests with a standard model aggregation algorithm. These results are the first demonstrating that some particular random forests achieve minimax rates *in arbitrary dimension*. Owing to their remarkably simple distributional properties, which lead to minimax rates, Mondrian trees are a promising basis for more sophisticated yet theoretically sound random forests variants.

Contents

2.1 Introduction

Introduced by Breiman (2001a), *Random Forests* (RF) are state-of-the-art classification and regression algorithms that proceed by averaging the forecasts of a number of randomized decision trees grown in parallel. Many extensions of RF have been proposed to tackle quantile estimation problems (Meinshausen, 2006), survival analysis (Ishwaran et al., 2008) and ranking (Clémençon et al., 2013); improvements of original RF are provided in literature, to cite but a few, better sampling strategies (Geurts et al., 2006), new splitting methods (Menze et al., 2011) or Bayesian alternatives (Chipman et al., 2010). Despite their widespread use and remarkable success in practical applications, the theoretical properties of such algorithms are still not fully understood (for an overview of theoretical results on RF, see Biau and Scornet, 2016). As a result of the complexity of the procedure, which combines sampling steps and feature selection, Breiman's original algorithm has proved difficult to analyze. A recent line of research (Scornet et al., 2015; Wager and Walther, 2015; Mentch and Hooker, 2016; Cui et al., 2017; Wager and Athey, 2018; Athey et al., 2019) has sought to obtain some theoretical guarantees for RF variants that closely resembled the algorithm used in practice. It should be noted, however, that most of these theoretical guarantees only offer limited information on the quantitative behavior of the algorithm (guidance for parameter tuning is scarce) or come at the price of conjectures on the true behavior of the RF algorithm itself, being thus still far from explaining the excellent empirical performance of it.

In order to achieve a better understanding of the random forest algorithm, another line of research focuses on modified and stylized versions of RF. Among these methods, *Purely Random Forests* (PRF) (Breiman, 2000; Biau et al., 2008; Biau, 2012; Genuer, 2012; Arlot and Genuer, 2014; Klusowski, 2018) grow the individual trees independently of the sample, and are thus particularly amenable to theoretical analysis. The consistency of such algorithms (as well as other idealized RF procedures) was first obtained by Biau et al. (2008), as a byproduct of the consistency of individual tree estimates. These results aim at quantifying the performance guarantees by analyzing the bias/variance of simplified versions of RF, such as PRF models (Genuer, 2012; Arlot and Genuer, 2014). In particular, Genuer (2012) shows that some PRF variant achieves the minimax rate for the estimation of a Lipschitz regression function in dimension one. The bias-variance analysis is extended by Arlot and Genuer (2014), showing that PRF can also achieve minimax rates for $\mathscr{C}^2$ regression functions in dimension one. These results are much more precise than mere consistency, and offer insights on the proper tuning of the procedure. Quite surprisingly, these optimal rates are only obtained in the one-dimensional case (where decision trees reduce to histograms). In the multi-dimensional setting, where trees exhibit an intricate recursive structure, only suboptimal rates are derived. As shown by lower bounds from Klusowski (2018), this is not merely a limitation from the analysis: centered forests, a standard variant of PRF, exhibit suboptimal rates under nonparametric assumptions.

From a more practical perspective, an important limitation of the most commonly used RF algorithms, such as Breiman's Random Forests (Breiman, 2001a) and the Extra-Trees algorithm (Geurts et al., 2006), is that they are typically trained in a batch manner, where the whole dataset, available at once, is required to build the trees. In order to allow their use in situations where large amounts of data have to be analyzed in a streaming fashion, several online variants of decision trees and RF algorithms have been proposed (Domingos

and Hulten, 2000; Saffari et al., 2009; Taddy et al., 2011; Denil et al., 2013, 2014).

Of particular interest in this article is the *Mondrian Forest* (MF) algorithm, an efficient and accurate online random forest classifier introduced by Lakshminarayanan et al. (2014), see also Lakshminarayanan et al. (2016). This algorithm is based on the Mondrian process (Roy and Teh, 2009; Roy, 2011; Orbanz and Roy, 2015), a natural probability distribution on the set of recursive partitions of the unit cube $[0,1]^d$. An appealing property of Mondrian processes is that they can be updated in an online fashion. In Lakshminarayanan et al. (2014), the use of the *conditional Mondrian* process enables the authors to design an online algorithm which matches its batch counterpart: training the algorithm one data point at a time leads to the same randomized estimator as training the algorithm on the whole dataset at once. The algorithm proposed in Lakshminarayanan et al. (2014) depends on a lifetime parameter $\lambda > 0$ that guides the complexity of the trees by stopping their building process. However, a theoretical analysis of MF is lacking, in particular, the tuning of λ is unclear from a theoretical perspective. In this chapter, we show that, aside from their appealing computational properties, Mondrian Forests are amenable to a precise theoretical analysis. We study MF in a batch setting and provide theoretical guidance on the tuning of λ.

Based on a detailed analysis of Mondrian partitions, we prove consistency and convergence rates for MF *in arbitrary dimension*, that turn out to be minimax optimal on the set of s-Hölder function with $s \in (0,2]$, assuming that λ and the number of trees in the forest (for $s \in (1,2]$) are properly tuned. Furthermore, we construct a procedure that adapts to the unknown smoothness $s \in (0,2]$ by combining Mondrian Forests with a standard model aggregation algorithm. To the best of our knowledge, such results have only been proved for very specific purely random forests, where the covariate space is of dimension one (Arlot and Genuer, 2014). Our analysis also sheds light on the benefits of Mondrian Forests compared to single Mondrian Trees: the bias reduction of Mondrian Forests allow them to be minimax for $s \in (1,2]$, while a single tree fails to be minimax in this case.

Agenda. This chapter is organized as follows. In Section 2.2, we describe the considered setting and set the notations for trees and forests. Section 2.3 defines the Mondrian process introduced by Roy and Teh (2009) and describes the MF algorithm. Section 2.4 provides new sharp properties for Mondrian partitions: cells distribution in Proposition 2.1 and a control of the cells diameter in Corollary 2.1, while the expected number of cells is provided in Proposition 2.2. Building on these properties, we provide, in Section 2.5, statistical guarantees for MF: Theorem 2.1 proves consistency, while Theorems 2.2 and 2.3 provide minimax rates for $s \in (0,1]$ and $s \in (1,2]$ respectively. Finally, Proposition 2.4 proves that a combination of MF with a model aggregation algorithm adapts to the unknown smoothness $s \in (0,2]$.

2.2 Setting and notations

We first describe the setting of the chapter and set the notations related to the Mondrian tree structure. For the sake of conciseness, we consider the regression setting, and show how to extend the results to classification in Section 2.5.5.

Setting. We consider a regression framework, where the dataset $\mathscr{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ consists of i.i.d. $[0,1]^d \times \mathbf{R}$-valued random variables. We assume throughout the chapter that

the dataset is distributed as a generic pair (X, Y) such that $\mathbb{E}[Y^2] < \infty$. This unknown distribution, characterized by the distribution μ of X on $[0,1]^d$ and by the conditional distribution of $Y|X$, can be written as

$$Y = f(X) + \varepsilon, \tag{2.1}$$

where $f(X) = \mathbb{E}[Y|X]$ is the conditional expectation of Y given X, and ε is a noise satisfying $\mathbb{E}[\varepsilon|X] = 0$. Our goal is to output a *randomized estimate* $\widehat{f}_n(\cdot, Z, \mathscr{D}_n) : [0,1]^d \to \mathbf{R}$, where Z is a random variable that accounts for the randomization procedure. To simplify notation, we will denote $\widehat{f}_n(x, Z) = \widehat{f}_n(x, Z, \mathscr{D}_n)$. The quality of a randomized estimate $\widehat{f}_n$ is measured by its quadratic risk

$$R(\widehat{f}_n) = \mathbb{E}[(\widehat{f}_n(X, Z) - f(X))^2]$$

where the expectation is taken with respect to $(X, Z, \mathscr{D}_n)$. We say that a sequence $(\widehat{f}_n)_{n \geqslant 1}$ is *consistent* whenever $R(\widehat{f}_n) \to 0$ as $n \to \infty$.

Trees and Forests. A regression tree is a particular type of partitioning estimate. First, a recursive partition Π of $[0,1]^d$ is built by performing successive axis-aligned splits (see Section 2.3), then the regression tree prediction is computed by averaging the labels Y_i of observations falling in the same cell as the query point $x \in [0,1]^d$, that is

$$\widehat{f}_n(x, \Pi) = \sum_{i=1}^{n} \frac{\mathbf{1}(X_i \in C_\Pi(x))}{N_n(C_\Pi(x))} Y_i, \tag{2.2}$$

where $C_\Pi(x)$ is the cell of the tree partition containing x and $N_n(C_\Pi(x))$ is the number of observations falling into $C_\Pi(x)$, with the convention that the estimate returns 0 if the cell $C_\Pi(x)$ is empty.

A random forest estimate is obtained by averaging the predictions of M randomized decision trees; more precisely, we will consider purely random forests, where the randomization of each tree (denoted above by Z) comes exclusively from the random partition, which is independent of $\mathscr{D}_n$. Let $\Pi_M = (\Pi^{(1)}, \ldots, \Pi^{(M)})$, where $\Pi^{(m)}$ (for $m = 1, \ldots, M$) are i.i.d. random partitions of $[0,1]^d$. The random forest estimate is thus defined as

$$\widehat{f}_{n,M}(x, \Pi_M) = \frac{1}{M} \sum_{m=1}^{M} \widehat{f}_n(x, \Pi^{(m)}), \tag{2.3}$$

where $\widehat{f}_n(x, \Pi^{(m)})$ is the prediction, at point x, of the tree with random partition $\Pi^{(m)}$, defined in (2.2).

The Mondrian Forest, whose construction is described below, is a particular instance of (2.3), in which the Mondrian process plays a crucial role by specifying the randomness Π of tree partitions.

2.3 The Mondrian Forest algorithm

Given a rectangular box $C = \prod_{j=1}^{d} [a_j, b_j] \subseteq \mathbf{R}^d$, we denote $|C| := \sum_{j=1}^{d} (b_j - a_j)$ its *linear dimension*. The Mondrian process $\mathsf{MP}(C)$ is a distribution on (infinite) tree partitions of C introduced by Roy and Teh (2009), see also Roy (2011) for a rigorous construction. Mondrian

partitions are built by iteratively splitting cells at some random time, which depends on the linear dimension of the cell; the splitting probability on each side is proportional to the side length of the cell, and the position is drawn uniformly.

The Mondrian process distribution $\mathsf{MP}(\lambda, C)$ is a distribution on tree partitions of C, resulting from the pruning of partitions drawn from $\mathsf{MP}(C)$. The pruning is done by removing all splits occurring after time $\lambda > 0$. In this perspective, λ is called the lifetime parameter and controls the complexity of the partition: large values of λ corresponds to deep trees (complex partitions).

Sampling from the distribution $\mathsf{MP}(\lambda, C)$ can be done efficiently by applying the recursive procedure $\texttt{SampleMondrian}(C, \tau = 0, \lambda)$ described in Algorithm 1. Figure 2.1 below shows a particular instance of Mondrian partition on a square box, with lifetime parameter $\lambda = 3.4$. In what follows, $\mathsf{Exp}(\lambda)$ stands for the exponential distribution with intensity $\lambda > 0$.

Algorithm 1 $\texttt{SampleMondrian}(C, \tau, \lambda)$: samples a Mondrian partition of C, starting from time τ and until time λ.

Inputs: A cell $C = \prod_{1 \leqslant j \leqslant d} [a_j, b_j]$, starting time τ and lifetime parameter λ.
Sample a random variable $E_C \sim \mathsf{Exp}(|C|)$
if $\tau + E_C \leqslant \lambda$ **then**
    Sample a split dimension $J \in \{1, \ldots, d\}$, with $\mathbb{P}(J = j) = (b_j - a_j)/|C|$
    Sample a split threshold S_J uniformly in $[a_J, b_J]$
    Split C along the split (J, S_J): let $C_0 = \{x \in C : x_J \leqslant S_J\}$ and $C_1 = C \setminus C_0$
    return $\texttt{SampleMondrian}(C_0, \tau + E_C, \lambda) \cup \texttt{SampleMondrian}(C_1, \tau + E_C, \lambda)$
else
    return $\{C\}$ (*i.e.*, do not split C).
end if

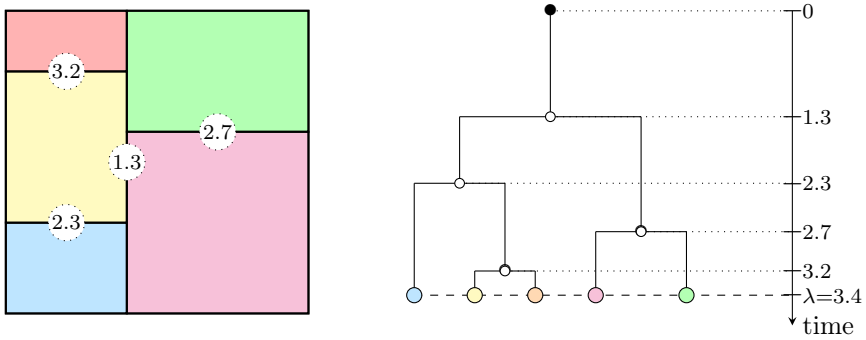

Figure 2.1: A Mondrian partition (left) with corresponding tree structure (right), which shows the evolution of the tree over time. The split times are indicated on the vertical axis, while the splits are denoted with bullets ($\circ$).

Remark 2.1. Using the fact that Exp is memoryless (if $E \sim \mathsf{Exp}(\lambda)$ and $u > 0$ then $E - u | E > u \sim \mathsf{Exp}(\lambda)$), it is possible to efficiently sample $\Pi_{\lambda'} \sim \mathsf{MP}(\lambda', C)$ given its pruning $\Pi_\lambda \sim \mathsf{MP}(\lambda, C)$ at time $\lambda \leqslant \lambda'$.

A Mondrian Tree estimator is given by Equation (2.2) where the partition $\Pi^{(m)}$ is sampled from the distribution $\mathsf{MP}(\lambda, [0,1]^d)$. The Mondrian Forest grows randomized tree partitions

$\Pi_\lambda^{(1)}, \ldots, \Pi_\lambda^{(M)}$, fits each one with the dataset $\mathscr{D}_n$ by averaging the labels falling into each leaf, then combines the resulting Mondrian Tree estimates by averaging their predictions. In accordance with Equation (2.3), we let

$$\widehat{f}_{\lambda,n,M}(x, \Pi_{\lambda,M}) = \frac{1}{M} \sum_{m=1}^{M} \widehat{f}_{\lambda,n}^{(m)}(x, \Pi_\lambda^{(m)}) \tag{2.4}$$

be the Mondrian Forest estimate described above, 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 $\widehat{f}_{\lambda,n}^{(m)}(x, \Pi_\lambda^{(m)})$. Although we use the standard definition of Mondrian processes, the way we compute the prediction in a Mondrian Tree differs from the original one. Indeed, in Lakshminarayanan et al. (2014), prediction is given by the expectation over a posterior distribution, where a hierarchical prior is assumed on the label distribution of each cell of the tree. In this chapter, we simply compute the average of the observations falling into a given cell.

2.4 Local and global properties of the Mondrian process

In this Section, we show that the properties of the Mondrian process enable us to compute explicitly some local and global quantities related to the structure of Mondrian partitions. To do so, we will need the following two facts, exposed by Roy and Teh (2009).

Fact 2.1 (Dimension 1). *For $d = 1$, the splits from a Mondrian process $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1])$ form a subset of $[0,1]$, which is distributed as a Poisson point process of intensity $\lambda \mathrm{d}x$.*

Fact 2.2 (Restriction). *Let $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$ be a Mondrian partition, and $C = \prod_{j=1}^d [a_j, b_j] \subset [0,1]^d$ be a box. Consider the* restriction $\Pi_\lambda|_C$ *of Π_λ on C, i.e. the partition on C induced by the partition Π_λ of $[0,1]^d$. Then $\Pi_\lambda|_C \sim \mathsf{MP}(\lambda, C)$.*

Fact 2.1 deals with the one-dimensional case by making explicit the distribution of splits for Mondrian process, which follows a Poisson point process. The restriction property stated in Fact 2.2 is fundamental, and enables one to precisely characterize the behavior of the Mondrian partitions.

Given any point $x \in [0,1]^d$, Proposition 2.1 below is a sharp result giving the exact distribution of the cell $C_\lambda(x)$ containing x from the Mondrian partition. Such a characterization is typically unavailable for other randomized trees partitions involving a complex recursive structure.

Proposition 2.1 (Cell distribution). *Let $x \in [0,1]^d$ and denote by*

$$C_\lambda(x) = \prod_{1 \leqslant j \leqslant d} [L_{j,\lambda}(x), R_{j,\lambda}(x)]$$

the cell containing x in a partition $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$ (this cell corresponds to a leaf). Then, the distribution of $C_\lambda(x)$ is characterized by the following properties:

(*i*) *$L_{1,\lambda}(x), R_{1,\lambda}(x), \ldots, L_{d,\lambda}(x), R_{d,\lambda}(x)$ are independent*;

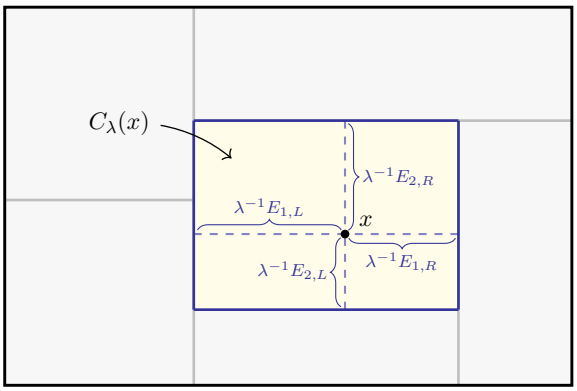

Figure 2.2: Cell distribution in a Mondrian partition (Proposition 2.1).

(*ii*) *For each* $j = 1, \dots, d$, $L_{j,\lambda}(x)$ *is distributed as* $(x - \lambda^{-1}E_{j,L}) \vee 0$ *and* $R_{j,\lambda}(x)$ *as* $(x + \lambda^{-1}E_{j,R}) \wedge 1$, *where* $E_{j,L}, E_{j,R} \sim \mathsf{Exp}(1)$.

The proof of Proposition 2.1 is given in Section 2.7. Figure 2.2 is a graphical representation of Proposition 2.1. A consequence of Proposition 2.1 is the next Corollary 2.1, which gives a precise upper bound on the diameter of the cells. In particular, this result is used in the proofs of the theoretical guarantees for Mondrian Trees and Forests from Section 2.5 below.

Corollary 2.1 (Cell diameter)**.** *Set* $\lambda > 0$ *and* $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$ *be a Mondrian partition. Let* $x \in [0,1]^d$ *and let* $D_\lambda(x)$ *be the* ℓ^2*-diameter of the cell* $C_\lambda(x)$ *containing* x *in* Π_λ. *For every* $\delta > 0$, *we have*

$$\mathbb{P}(D_\lambda(x) \geqslant \delta) \leqslant d\Big(1 + \frac{\lambda\delta}{\sqrt{d}}\Big) \exp\Big(-\frac{\lambda\delta}{\sqrt{d}}\Big) \tag{2.5}$$

and

$$\mathbb{E}\big[D_\lambda(x)^2\big] \leqslant \frac{4d}{\lambda^2}\,. \tag{2.6}$$

In order to control the risk of Mondrian Trees and Forests, we need an upper bound on the number of cells in a Mondrian partition. Quite surprisingly, the expectation of this quantity can be computed exactly, as shown in Proposition 2.2.

Proposition 2.2 (Number of cells)**.** *Set* $\lambda > 0$ *and* $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$ *be a Mondrian partition. If* K_λ *denotes the number of cells in* Π_λ, *we have* $\mathbb{E}[K_\lambda] = (1+\lambda)^d$.

The proof of Proposition 2.2 is given in Section 2.8.2, while a sketch of proof is provided in Section 2.7. Although the proof is technically involved, it relies on a natural coupling argument: we introduce a recursive modification of the construction of the Mondrian process which keeps the expected number of leaves unchanged, and for which this quantity can be computed directly using the Mondrian-Poisson equivalence in dimension one (Fact 2.1). A much simpler result is $\mathbb{E}[K_\lambda] \leqslant (e(1+\lambda))^d$, which was previously obtained in Mourtada et al. (2017). By contrast, Proposition 2.2 provides the *exact* value of this expectation, which removes a superfluous e^d factor.

Remark 2.2. Proposition 2.2 naturally extends (with the same proof) to the more general case of a Mondrian process with finite measures with no atoms $\nu_1, \dots, \nu_d$ on the sides $C^1, \dots, C^d$ of a box $C \subseteq \mathbf{R}^d$ (for a definition of the Mondrian process in this more general case, see Roy, 2011). In this case, we have $\mathbb{E}[K_\lambda] = \prod_{1 \leqslant j \leqslant d}(1 + \nu_j(C^j))$.

As illustrated in this Section, a remarkable fact with the Mondrian Forest is that the quantities of interest for the statistical analysis of the algorithm can be made explicit. In particular, we have seen in this Section that, roughly speaking, a Mondrian partition is balanced enough that it contains $O(\lambda^d)$ cells of diameter $O(1/\lambda)$, which is the minimal number of cells to cover $[0,1]^d$.

2.5 Minimax theory for Mondrian Forests

This Section gathers several theoretical guarantees for Mondrian Trees and Forests. Section 2.5.1 states the universal consistency of the procedure, provided that the lifetime λ_n belongs to an appropriate range. We provide convergence rates which turn out to be minimax optimal for s-Hölder regression functions with $s \in (0,1]$ in Section 2.5.2 and with $s \in (1,2]$ in Section 2.5.3, provided in both cases that λ_n is properly tuned. Note that in particular, we illustrate in Section 2.5.3 the fact that Mondrian Forests improve over Mondrian trees, when $s \in (1,2]$. In Section 2.5.4, we prove that a combination of MF with a model aggregation algorithm adapts to the unknown $s \in (0,2]$. Finally, results for classification are given in Section 2.5.5.

2.5.1 Consistency of Mondrian Forests

The consistency of the Mondrian Forest estimator is established in Theorem 2.1 below, assuming a proper tuning of the lifetime parameter λ_n.

Theorem 2.1 (Universal consistency)**.** *Let $M \geqslant 1$. Consider Mondrian Trees $\widehat{f}^{(m)}_{\lambda_n,n}$ (for $m = 1,\dots,M$) and Mondrian Forest $\widehat{f}_{\lambda_n,n,M}$ given by Equation (2.4) for a sequence $(\lambda_n)_{n\geqslant 1}$ satisfying $\lambda_n \to \infty$ and $\lambda_n^d/n \to 0$. Then, under the setting described in Section 2.2 above, the individual trees $\widehat{f}^{(m)}_{\lambda_n,n}$ (for $m = 1,\dots,M$) are consistent, and as a consequence, the forest $\widehat{f}_{\lambda_n,n,M}$ is consistent for any $M \geqslant 1$.*

The proof of Theorem 2.1 is given in Section 2.8.3. It uses the properties of Mondrian partitions established in Section 2.4 together with general consistency results for histograms. This result is universal, in the sense that it makes no assumption on the joint distribution of (X,Y), apart from $\mathbb{E}[Y^2] < \infty$ in order to ensure that the quadratic risk is well-defined (see Section 2.2).

The only tuning parameter of a Mondrian Tree is the lifetime λ_n, which encodes the complexity of the trees. Requiring an assumption on this parameter is natural, and confirmed by the well-known fact that the tree-depth is an important tuning parameter for Random Forests, see Biau and Scornet (2016). However, Theorem 2.1 does not address the question of a theoretically optimal tuning of λ_n under additional assumptions on the regression function f, which we consider in the following sections.

2.5.2 Mondrian Trees and Forests are minimax over s-Hölder classes for $s \in (0,1]$

The bounds obtained in Corollary 2.1 and Proposition 2.2 are explicit and sharp in their dependency on λ. Based on these properties, we now establish a theoretical upper bound on the risk of Mondrian Trees, which gives the optimal theoretical tuning of the lifetime parameter λ_n. To pursue the analysis, we need the following assumption.

Assumption 2.1. Consider (X, Y) from the setting described in Section 2.2 and assume also that $\mathbb{E}[\varepsilon|X] = 0$ and $\mathrm{Var}(\varepsilon|X) \leqslant \sigma^2 < \infty$ almost surely, where ε is given by Equation (2.1).

Our minimax results hold for a class of s-Hölder regression functions defined below.

Definition 2.1. Let $p \in \mathbf{N}$, $\beta \in (0, 1]$ and $L > 0$. The (p, β)-Hölder ball of norm L, denoted $\mathscr{C}^{p,\beta}(L) = \mathscr{C}^{p,\beta}([0,1]^d, L)$, is the set of p times differentiable functions $f : [0,1]^d \to \mathbf{R}$ such that

$$\|\nabla^p f(x) - \nabla^p f(x')\| \leqslant L\|x - x'\|^\beta \quad \text{and} \quad \|\nabla^k f(x)\| \leqslant L$$

for every $x, x' \in [0,1]^d$ and $k \in \{1, \dots, p\}$. Whenever $f \in \mathscr{C}^{p,\beta}(L)$, we say that f is s-Hölder with $s = p + \beta$.

Note that in what follows we will assume $s \in (0, 2]$, so that $p \in \{0, 1\}$. Theorem 2.2 below states an upper bound on the risk of Mondrian Trees and Forests, which explicitly depends on the lifetime parameter λ. Selecting λ that minimizes this bound leads to a convergence rate which turns out to be minimax optimal over the class of s-Hölder functions for $s \in (0, 1]$ (see for instance Stone, 1982, Chapter I.3 in Nemirovski, 2000 or Theorem 3.2 in Györfi et al., 2002).

Theorem 2.2. *Grant Assumption 2.1 and assume that $f \in \mathscr{C}^{0,\beta}(L)$, where $\beta \in (0, 1]$ and $L > 0$. Let $M \geqslant 1$. The quadratic risk of the Mondrian Forest $\widehat{f}_{\lambda,n,M}$ with lifetime parameter $\lambda > 0$ satisfies*

$$\mathbb{E}\big[(\widehat{f}_{\lambda,n,M}(X) - f(X))^2\big] \leqslant \frac{(4d)^\beta L^2}{\lambda^{2\beta}} + \frac{(1+\lambda)^d}{n}\big(2\sigma^2 + 9\|f\|_\infty^2\big). \tag{2.7}$$

In particular, as $n \to \infty$, the choice $\lambda := \lambda_n \asymp L^{2/(d+2\beta)} n^{1/(d+2\beta)}$ gives

$$\mathbb{E}\big[(\widehat{f}_{\lambda_n,n,M}(X) - f(X))^2\big] = O(L^{2d/(d+2\beta)} n^{-2\beta/(d+2\beta)}), \tag{2.8}$$

which corresponds to the minimax rate over the class $\mathscr{C}^{0,\beta}(L)$.

The proof of Theorem 2.2 is given in Section 2.7. It relies on the properties about Mondrian partitions stated in Section 2.4. Namely, Corollary 2.1 allows to control the bias of Mondrian Trees (first term on the right-hand side of Equation 2.7), while Proposition 2.2 helps in controlling the variance of Mondrian Trees (second term on the right-hand side of Equation 2.7).

To the best of our knowledge, Theorem 2.2 is the first to prove that a purely random forest (Mondrian Forest in this case) can be minimax optimal *in arbitrary dimension*. Minimax optimal upper bounds are obtained for $d = 1$ in Genuer (2012) and Arlot and Genuer (2014) for models of purely random forests such as Toy-PRF (where the individual partitions correspond to random shifts of the regular partition of $[0, 1]$ in k intervals) and PURF (Purely Uniformly Random Forests, where the partitions are obtained by drawing k random thresholds uniformly in $[0, 1]$). However, for $d = 1$, tree partitions reduce to partitions of $[0, 1]$ in intervals, and do not possess the recursive structure that appears in higher dimensions, which makes their analysis challenging. For this reason, the analysis of purely random forests for $d > 1$ has typically produced sub-optimal results: for example, Biau (2012) exhibit an upper bound on the risk of the centered random forests (a particular instance of PRF) which turns out to be much slower than the minimax rate for Lipschitz regression functions. A more in-depth

analysis of the same random forest model in Klusowski (2018) exhibits a new upper and lower bound of the risk, which is still slower than minimax rates for Lipschitz functions. A similar result was proved by Arlot and Genuer (2014), who studied the BPRF (Balanced Purely Random Forests algorithm, where all leaves are split, so that the resulting tree is complete), and obtained suboptimal rates. In our approach, the convenient properties of the Mondrian process enable us to bypass the inherent difficulties met in previous attempts. One specificity of Mondrian forests compared to other PRF variants is that the largest sides of cells are more likely to be split. By contrast, variants of PRF (such as centered forests) where the coordinate of the split is chosen with equal probability, may give rise to unbalanced cells with large diameter.

Theorem 2.2 provides theoretical guidance on the choice of the lifetime parameter, and suggests to set $\lambda := \lambda_n \asymp n^{1/(d+2)}$. Such an insight cannot be gleaned from an analysis that focuses on consistency alone. Theorem 2.2 is valid for Mondrian Forests with any number of trees, and thus in particular for a Mondrian Tree (this is also true for Theorem 2.1). However, it is a well-known fact that forests outperform single trees in practice (Fernández-Delgado et al., 2014). Section 2.5.3 proposes an explanation for this phenomenon, by assuming $f \in \mathscr{C}^{1,\beta}(L)$.

2.5.3 Improved rates for Mondrian Forests compared to a Mondrian Tree

The convergence rate stated in Theorem 2.2 for $f \in \mathscr{C}^{0,\beta}(L)$ is valid for both trees and forests, and the risk bound does not depend on the number M of trees that compose the forest. In practice, however, forests exhibit much better performances than individual trees. In this Section, we provide a result that illustrates the benefits of forests over trees by assuming that $f \in \mathscr{C}^{1,\beta}(L)$. As the counterexample in Proposition 2.3 below shows, single Mondrian trees do not benefit from this additional smoothness assumption, and achieve the same rate as in the Lipschitz case. This comes from the fact that the bias of trees is highly sub-optimal for such functions.

Proposition 2.3. *Assume that $Y = f(X) + \varepsilon$ with $f(x) = 1 + x$, where $X \sim \mathcal{U}([0,1])$ and ε is independent of X with variance σ^2. Consider a single Mondrian Tree estimate $\widehat{f}^{(1)}_{\lambda,n}$. Then, there exists a constant $C_0 > 0$ such that*

$$\inf_{\lambda \in \mathbf{R}_+^*} \mathbb{E}\big[(\widehat{f}^{(1)}_{\lambda,n}(X) - f(X))^2\big] \geqslant C_0 \wedge \frac{1}{4}\Big(\frac{3\sigma^2}{n}\Big)^{2/3}$$

for any $n \geqslant 18$.

The proof of Proposition 2.3 is given in Section 2.8.4. Since the minimax rate over $\mathscr{C}^{1,1}$ in dimension 1 is $O(n^{-4/5})$, Proposition 2.3 proves that a single Mondrian Tree is not minimax optimal over this function class. However, it turns out that large enough Mondrian Forests, which average Mondrian trees, are minimax optimal over $\mathscr{C}^{1,1}$. Therefore, Theorem 2.3 below highlights the benefits of a forest compared to a single tree.

Theorem 2.3. *Grant Assumption 2.1 and assume that $f \in \mathscr{C}^{1,\beta}(L)$, with $\beta \in (0,1]$ and $L > 0$. In addition, assume that X has a positive and C_p-Lipschitz density p w.r.t the Lebesgue measure on $[0,1]^d$. Let $\widehat{f}_{\lambda,n,M}$ be the Mondrian Forest estimate given by (2.4). Set $\varepsilon \in (0, 1/2)$*

and $B_\varepsilon = [\varepsilon, 1-\varepsilon]^d$. *Then, we have*

$$\mathbb{E}\big[(\widehat{f}_{\lambda,n,M}(X) - f(X))^2 | X \in B_\varepsilon\big] \leqslant \frac{2(1+\lambda)^d}{n} \frac{2\sigma^2 + 9\|f\|_\infty^2}{p_0(1-2\varepsilon)^d} + \frac{144L^2 d p_1}{p_0(1-2\varepsilon)^d} \frac{e^{-\lambda\varepsilon}}{\lambda^3} + \\ + \frac{72L^2 d^3}{\lambda^4} \Big(\frac{p_1 C_p}{p_0^2}\Big)^2 + \frac{16L^2 d^{1+\beta}}{\lambda^{2(1+\beta)}} \Big(\frac{p_1}{p_0}\Big)^2 + \frac{8dL^2}{M\lambda^2}, \quad (2.9)$$

where $p_0 = \inf_{x\in[0,1]^d} p(x)$ *and* $p_1 = \sup_{x\in[0,1]^d} p(x)$. *In particular, letting* $s = 1+\beta$, *the choices*

$$\lambda_n \asymp L^{2/(d+2s)} n^{1/(d+2s)} \quad \text{and} \quad M_n \gtrsim L^{4\beta/(d+2s)} n^{2\beta/(d+2s)}$$

give

$$\mathbb{E}\big[(\widehat{f}_{\lambda_n,n,M_n}(X) - f(X))^2 | X \in B_\varepsilon\big] = O(L^{2d/(d+2s)} n^{-2s/(d+2s)}), \quad (2.10)$$

which corresponds to the minimax risk over the class $\mathscr{C}^{1,\beta}(L)$.

In the case where $\varepsilon = 0$, *which corresponds to integrating over the whole hypercube, the bound* (2.10) *holds if* $2s \leqslant 3$. *On the other hand, if* $2s > 3$, *letting*

$$\lambda_n \asymp L^{2/(d+3)} n^{1/(d+3)} \quad \text{and} \quad M_n \gtrsim L^{4/(d+3)} n^{2/(d+3)}$$

yields the following upper bound on the integrated risk of the Mondrian Forest estimate over B_0

$$\mathbb{E}\big[(\widehat{f}_{\lambda_n,n,M_n}(X) - f(X))^2\big] = O(L^{2d/(d+3)} n^{-3/(d+3)}). \quad (2.11)$$

The proof of Theorem 2.3 is given in Section 2.7 below. It relies on an improved control of the bias, compared to the one used in Theorem 2.2 in the Lipschitz case: it exploits the knowledge of the distribution of the cell $C_\lambda(x)$ given in Proposition 2.1 instead of merely the cell diameter given in Corollary 2.1 (which was enough for Theorem 2.2). The improved rate for Mondrian Forests compared to Mondrian Trees comes from the fact that large enough forests have a smaller bias than single trees for smooth regression functions. This corresponds to the fact that averaging randomized trees tends to smooth the decision function of single trees, which are discontinuous piecewise constant functions that approximate smooth functions sub-optimally. Such an effect was already noticed by Arlot and Genuer (2014) for purely random forests.

Remark 2.3. While (2.10) gives the minimax rate for $\mathscr{C}^{1,1}$ functions, it suffers from an unavoidable standard artifact, namely a boundary effect which impacts local averaging estimates, such as kernel estimators (Wasserman, 2006; Arlot and Genuer, 2014). It is however possible to set $\varepsilon = 0$ in (2.9), which leads to the sub-optimal rate stated in (2.11).

2.5.4 Adaptation to smoothness

The minimax rates of Theorems 2.2 and 2.3 for trees and forests are achieved through a specific tuning of the lifetime parameter λ, which depends on the considered smoothness class $\mathscr{C}^{p,\beta}(L)$ through $s = p+\beta$ and $L > 0$, while on the other hand, the number of trees M simply needs to be large enough in the statement of Theorem 2.3. Since in practice such smoothness parameters are unknown, it is of interest to obtain a single method that *adapts* to them.

In order to achieve this, we adopt a standard approach based on model aggregation (Nemirovski, 2000). More specifically, we split the dataset into two part: the first is used to fit

Mondrian Forest estimators with λ varying in an exponential grid, while the second part is used to fit the STAR procedure for model aggregation, introduced by Audibert (2008). The appeals of this aggregation procedure are its simplicity, its optimal guarantee and the lack of parameter to tune.

Let $n_0 = \lfloor n/2 \rfloor$, $\mathscr{D}_{n_0} = \{(X_i, Y_i) : 1 \leqslant i \leqslant n_0\}$ and $\mathscr{D}_{n_0+1:n} = \{(X_i, Y_i) : n_0 + 1 \leqslant i \leqslant n\}$. Also, let $I_\varepsilon = \{i \in \{n_0 + 1, \dots, n\} : X_i \in [\varepsilon, 1-\varepsilon]^d\}$ for some $\varepsilon \in (0, 1/2)$. If I_ε is empty, we let the estimator be $\widehat{g}_n = 0$. We define $A = \lfloor \log_2(n^{1/d}) \rfloor$ and $M = \lceil n^{2/d} \rceil$ and consider the geometric grid $\Lambda = \{2^\alpha : \alpha = 0, \dots, A\}$. Now, let

$$\Pi^{(1)}_{n^{1/d}}, \dots, \Pi^{(M)}_{n^{1/d}} \sim \mathsf{MP}(n^{1/d}, [0,1]^d)$$

be i.i.d. Mondrian partitions. For $m = 1, \dots, M$, we let $\Pi^{(m)}_\lambda$ be the pruning of $\Pi^{(m)}_{n^{1/d}}$ in which only splits occurring before time λ have been kept. We consider now the Mondrian Forest estimators

$$\widehat{f}_\alpha = \widehat{f}_{2^\alpha, n_0, M}$$

for every $\alpha = 0, \dots, A$, where we recall that these estimators are given by (2.4). The estimators $\widehat{f}_\alpha$ are computed using the sample $\mathscr{D}_{n_0}$ and the Mondrian partitions $\Pi^{(m)}_{2^\alpha}$, $1 \leqslant m \leqslant M$. Let

$$\widehat{\alpha} = \operatorname*{arg\,min}_{\alpha=0,\dots,A} \frac{1}{|I_\varepsilon|} \sum_{i \in I_\varepsilon} (\widehat{f}_\alpha(X_i) - Y_i)^2$$

be a risk minimizer and let $\widehat{\mathcal{G}} = \bigcup_\alpha [\widehat{f}_{\widehat{\alpha}}, \widehat{f}_\alpha]$ where $[f, g] = \{(1-t)f + tg : t \in [0,1]\}$. Note that $\widehat{\mathcal{G}}$ is a star domain with origin at the empirical risk minimizer $\widehat{f}_{\widehat{\alpha}}$, hence the name STAR (Audibert, 2008). Then, the adaptive estimator is a convex combination of two Mondrian forests estimates with different lifetime parameters, given by

$$\widehat{g}_n = \operatorname*{arg\,min}_{g \in \widehat{\mathcal{G}}} \left\{ \frac{1}{|I_\varepsilon|} \sum_{i \in I_\varepsilon} (g(X_i) - Y_i)^2 \right\}. \tag{2.12}$$

Proposition 2.4. *Grant Assumption 2.1, with $|Y| \leqslant B$ almost surely and $f \in \mathscr{C}^{p,\beta}(L)$ with $p \in \{0, 1\}$, $\beta \in (0, 1]$ and $L > 0$. Also, assume that the density p of X is C_p-Lipschitz and satisfies $p_0 \leqslant p \leqslant p_1$. Then, the estimator $\widehat{g}_n$ defined by (2.12) satisfies:*

$$\begin{aligned} \mathbb{E}\big[(\widehat{g}_n(X) - f(X))^2 | X \in B_\varepsilon\big] \leqslant \min_{\alpha=0,\dots,A} \mathbb{E}\big[(\widehat{f}_\alpha(X) - f(X))^2 | X \in B_\varepsilon\big] \\ + 4B^2 e^{-c_1 n/4} + \frac{600 B^2 (\log(1 + \log_2 n) + 1)}{c_1 n} \end{aligned} \tag{2.13}$$

where $B_\varepsilon = [\varepsilon, 1-\varepsilon]^d$ and $c_1 = p_0(1-2\varepsilon)^d/4$. In particular, we have

$$\mathbb{E}\big[(\widehat{g}_n(X) - f(X))^2 | X \in B_\varepsilon\big] = O\big(L^{2d/(d+2s)} n^{-2s/(d+2s)}\big), \tag{2.14}$$

where $s = p + \beta$.

The proof of Proposition 2.4 is to be found in the Supplementary Material. Proposition 2.4 proves that the estimator $\widehat{g}_n$, which is a STAR aggregation of Mondrian Forests, is adaptive to the smoothness of f, whenever f is s-Hölder with $s \in (0, 2]$.

2.5.5 Results for binary classification

We now consider, as a by-product of the analysis conducted for regression estimation, the setting of binary classification. Assume that we are given a dataset $\mathscr{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ of i.i.d. random variables with values in $[0,1]^d \times \{0,1\}$, distributed as a generic pair (X, Y) and define $\eta(x) = \mathbb{P}[Y = 1 | X = x]$. We define the Mondrian Forest classifier $\widehat{g}_{\lambda,n,M}$ as a plug-in estimator of the regression estimator. Namely, we introduce

$$\widehat{g}_{\lambda,n,M}(x) = \mathbf{1}(\widehat{f}_{\lambda,n,M}(x) \geqslant 1/2)$$

for all $x \in [0,1]^d$, where $\widehat{f}_{\lambda,n,M}$ is the Mondrian Forest estimate defined in the regression setting. The performance of $\widehat{g}_{\lambda,n,M}$ is assessed by the 0-1 classification error defined as

$$L(\widehat{g}_{\lambda,n,M}) = \mathbb{P}(\widehat{g}_{\lambda,n,M}(X) \neq Y), \tag{2.15}$$

where the probability is taken with respect to $(X, Y, \Pi_{\lambda,M}, \mathscr{D}_n)$, where $\Pi_{\lambda,M}$ is the set sampled Mondrian partitions, see (2.4). Note that (2.15) is larger than the Bayes risk defined as

$$L(g^*) = \mathbb{P}(g^*(X) \neq Y),$$

where $g^*(x) = \mathbf{1}(\eta(x) \geqslant 1/2)$. A general theorem (Devroye et al., 1996, Theorem 6.5) allows us to derive an upper bound on the distance between the classification risk of $\widehat{g}_{\lambda,n,M}$ and the Bayes risk, based on Theorem 2.2.

Corollary 2.2. *Let $M \geqslant 1$ and assume that $\eta \in \mathscr{C}^{0,1}(L)$. Then, the Mondrian Forest classifier $\widehat{g}_n = \widehat{g}_{\lambda_n,n,M}$ with parameter $\lambda_n \asymp n^{1/(d+2)}$ satisfies*

$$L(\widehat{g}_n) - L(g^*) = o(n^{-1/(d+2)}).$$

The rate of convergence $o(n^{-1/(d+2)})$ for the error probability with a Lipschitz conditional probability η is optimal (Yang, 1999). In the same way, Theorem 2.3 extends to the context of classification:

Corollary 2.3. *Assume that X has a positive and Lipschitz density p w.r.t the Lebesgue measure on $[0,1]^d$ and that $\eta \in \mathscr{C}^{1,1}(L)$. Let $\widehat{g}_n = \widehat{g}_{\lambda_n,n,M_n}$ be the Mondrian Forest classifier composed of $M_n \gtrsim n^{2/(d+4)}$ trees, with lifetime $\lambda_n \asymp n^{1/(d+4)}$. Then, we have*

$$\mathbb{P}(\widehat{g}_n(X) \neq Y | X \in B_\varepsilon) - \mathbb{P}(g^*(X) \neq Y | X \in B_\varepsilon) = o(n^{-2/(d+4)}) \tag{2.16}$$

for all $\varepsilon \in (0, 1/2)$, where $B_\varepsilon = [\varepsilon, 1-\varepsilon]^d$.

This shows that Mondrian Forests achieve an improved rate compared to Mondrian trees for classification.

2.6 Conclusion

Despite their widespread use in practice, the theoretical understanding of Random Forests is still incomplete. In this chapter, we show that the Mondrian Forest, originally introduced to provide an efficient online algorithm, leads to an algorithm that is not only consistent, but in fact minimax optimal under nonparametric assumptions in arbitrary dimension. This

provides, to the best of our knowledge, the first results of this nature for a random forest method in arbitrary dimension. Besides, our analysis allows to illustrate improved rates for forests compared to individual trees. Mondrian partitions possess nice geometric properties, which can be controlled in an exact and direct fashion, while previous approaches (Biau et al., 2008; Arlot and Genuer, 2014) require arguments that work conditionally on the structure of the tree. Since Random forests are usually black-box procedures that are hard to analyze, it would be interesting to see whether the simple properties of the Mondrian process could be leveraged to design more sophisticated variants of RF that remain amenable to precise analysis.

The minimax rate $O(n^{-2s/(2s+d)})$ for a s-Hölder regression with $s \in (0, 2]$ obtained in this study is very slow when the number of features d is large. This comes from the well-known curse of dimensionality phenomenon, a problem affecting all fully nonparametric algorithms. A standard approach used in high-dimensional settings is to work under a sparsity assumption, where only $s \ll d$ features are informative. In this case, a procedure such as the Mondrian forests estimator should be used after a variable selection step. From a theoretical perspective, it would be interesting to see if the variable selection and function estimation steps could be combined, using results on the ability of forests to select informative variables (see, for instance, Scornet et al., 2015).

2.7 Proofs

This Section gathers the proofs of Proposition 2.1 and Corollary 2.1 (cell distribution and cell diameter). Then, a sketch of the proof of Proposition 2.2 is described in this Section (the full proof, which involves some technicalities, can be found in the Supplementary Material). Finally, we provide the proofs of Theorem 2.2 and Theorem 2.3.

Proof of Proposition 2.1. Let $0 \leqslant a_1, \dots, a_n, b_1, \dots, b_n \leqslant 1$ be such that $a_j \leqslant x_j \leqslant b_j$ for $1 \leqslant j \leqslant d$. Let $C := \prod_{j=1}^{d} [a_j, b_j]$. Note that the event

$$E_\lambda(C, x) = \left\{ L_{1,\lambda}(x) \leqslant a_1, R_{1,\lambda}(x) \geqslant b_1, \dots, L_{d,\lambda}(x) \leqslant a_d, R_{d,\lambda}(x) \geqslant b_d \right\}$$

coincides — up to the negligible event that one of the splits of Π_λ occurs on coordinate j at a_j or b_j — with the event that Π_λ does not cut C, *i.e.* that the restriction $\Pi_\lambda|_C$ of Π_λ to C contains no split. Now, by the restriction property of the Mondrian process (Fact 2.2), $\Pi_\lambda|_C$ is distributed as $\mathsf{MP}(\lambda, C)$; in particular, the probability that $\Pi_\lambda|_C$ contains no split is $\exp(-\lambda|C|)$. Hence, we have

$$\mathbb{P}(E_\lambda(C, x)) = e^{-\lambda(x-a_1)} e^{-\lambda(b_1-x)} \times \cdots \times e^{-\lambda(x-a_d)} e^{-\lambda(b_d-x)} . \tag{2.17}$$

In particular, setting $a_j = b_j = x$ in (2.17) except for one a_j or b_j, and using that $L_{j,\lambda}(x) \leqslant x$ and $R_{j,\lambda}(x) \geqslant x$, we obtain

$$\mathbb{P}(R_{j,\lambda}(x) \geqslant b_j) = e^{-\lambda(b_j-x)} \quad \text{and} \quad \mathbb{P}(L_{j,\lambda}(x) \leqslant a_j) = e^{-\lambda(x-a_j)}. \tag{2.18}$$

Since clearly $R_{j,\lambda}(x) \leqslant 1$ and $L_{j,\lambda}(x) \geqslant 0$, Equation (2.18) implies (ii). Additionally, plugging (2.18) back into Equation (2.17) shows that $L_{1,\lambda}(x), R_{1,\lambda}(x), \dots, L_{d,\lambda}(x), R_{d,\lambda}(x)$ are independent, *i.e.* point (i). This completes the proof. □

Proof of Corollary 2.1. Using Proposition 2.1, for $1 \leqslant j \leqslant d$, $D_{j,\lambda}(x) = R_{j,\lambda}(x) - x_j + x_j - L_{j,\lambda}(x)$ is stochastically upper bounded by $\lambda^{-1}(E_1 + E_2)$ with E_1, E_2 two independent $\mathsf{Exp}(1)$ random variables, which is distributed as $\mathsf{Gamma}(2, \lambda)$. This implies that

$$\mathbb{P}(D_{j,\lambda}(x) \geqslant \delta) \leqslant (1 + \lambda\delta)e^{-\lambda\delta} \tag{2.19}$$

for every $\delta > 0$ (with equality if $\delta \leqslant x_j \wedge (1 - x_j)$) and $\mathbb{E}[D_{j,\lambda}(x)^2] \leqslant \lambda^{-2}(\mathbb{E}[E_1^2] + \mathbb{E}[E_2^2]) = 4/\lambda^2$. The bound (2.5) for the diameter $D_\lambda(x) = [\sum_{j=1}^d D_{j,\lambda}(x)^2]^{1/2}$ is obtained by noting that

$$\mathbb{P}(D_\lambda(x) \geqslant \delta) \leqslant \mathbb{P}\Big(\exists j : D_{j,\lambda}(x) \geqslant \frac{\delta}{\sqrt{d}}\Big) \leqslant \sum_{j=1}^d \mathbb{P}\Big(D_{j,\lambda}(x) \geqslant \frac{\delta}{\sqrt{d}}\Big),$$

while (2.6) follows from the identity $\mathbb{E}[D_\lambda(x)^2] = \sum_{j=1}^d \mathbb{E}[D_{j,\lambda}(x)^2]$. □

Sketch of Proof of Proposition 2.2. Let us provide here an outline of the argument; a fully detailed proof is available in the Supplementary Material. The general idea of the proof is to modify the construction of Mondrian partitions (and hence their distribution) in a way that leaves the expected number of cells unchanged, while making this quantity directly computable.

Consider a Mondrian partition $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$ and a cell C formed at time τ in it (e.g., $C = [0,1]^d$ for $\tau = 0$). By the properties of exponential distributions, the split of C (if it exists) from Algorithm 1 can be obtained as follows. Sample independent variables E_j, U_j with $E_j \sim \mathsf{Exp}(1)$ and $U_j \sim \mathcal{U}([0,1])$ for $j = 1, \dots, d$. Let $T_j = (b_j - a_j)^{-1} E_j$ and $S_j = a_j + (b_j - a_j)U_j$, where $C = \prod_{j=1}^d [a_j, b_j]$, and set $J = \arg\min_{1 \leqslant j \leqslant d} T_j$. If $\tau + T_J > \lambda$ then C is not split (and is thus a cell of Π_λ). On the other hand, if $\tau + T_J \leqslant \lambda$ then C is split along coordinate J at S_J (and at time $\tau + T_J$) into $C' = \{x \in C : x_J \leqslant S_J\}$ and $C'' = C \setminus C'$. This process is then repeated for the cells C' and C'', by using independent random variables E'_j, U'_j and E''_j, U''_j respectively.

Now, note that the number of cells $K_\lambda(C)$ in Π_λ contained in C is the sum of the number of cells in C' and C'', namely $K_\lambda(C')$ and $K_\lambda(C'')$. Hence, the expectation of $K_\lambda(C)$ (conditionally on previous splits) only depends on the distribution of the split (J, S_J, T_J), as well as on the marginal distributions of $K_\lambda(C')$ and $K_\lambda(C'')$, but not on the joint distribution of $(K_\lambda(C'), K_\lambda(C''))$.

Consider the following change: instead of splitting C' and C'' based on the independent random variables E'_j, U'_j and E''_j, U''_j respectively, we reuse for both C' and C'' the variables E_j, U_j (and thus S_j, T_j) for $j \neq J$, which were not used to split C. It can be seen that, for both C' and C'', these variables have the same conditional distribution given J, S_J, T_J as the independent ones. One can then form the modified random partition $\widetilde{\Pi}_\lambda$ by recursively applying this change to the construction of Π_λ, starting with the root and propagating the unused variables at each split. By the above outlined argument, its number of cells $\widetilde{K}_\lambda$ satisfies $\mathbb{E}[\widetilde{K}_\lambda] = \mathbb{E}[K_\lambda]$.

On the other hand, one can show that the partition $\widetilde{\Pi}_\lambda$ is a "product" of independent one-dimensional Mondrian partition $\Pi_\lambda^j \sim \mathsf{MP}(\lambda, [0,1])$ along the coordinates $j = 1, \dots, d$ (this means that the cells of $\widetilde{\Pi}_\lambda$ are the Cartesian products of cells of the Π_λ^j). Since the splits of a one-dimensional Mondrian partition of $[0,1]$ form a Poisson point process of intensity $\lambda \mathrm{d}x$ (Fact 2.1), the expected number of cells of Π_λ^j is $1 + \lambda$. Since the Π_λ^j for $j = \{1, \dots, d\}$ are independent, this implies that $\mathbb{E}[\widetilde{K}_\lambda] = (1 + \lambda)^d$. Once again, the full proof is provided in the Supplementary Material. □

Proof of Theorem 2.2. Recall that the Mondrian Forest estimate at x is given by

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

By convexity of the function $y' \mapsto (y-y')^2$ for any $y \in \mathbf{R}$, we have

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

since the random trees estimators $\widehat{f}^{(m)}_{\lambda,n}$ have the same distribution for $m = 1 \dots M$. Hence, it suffices to prove Theorem 2.2 for the tree estimator $\widehat{f}^{(1)}_{\lambda,n}$. We will denote for short $\widehat{f}_\lambda := \widehat{f}^{(1)}_{\lambda,n}$ all along this proof.

Bias-variance decomposition. We establish a *bias-variance* decomposition of the risk of a Mondrian tree, akin to the one stated for purely random forests by Genuer (2012). Denote $\bar{f}_\lambda(x) := \mathbb{E}[f(X)|X \in C_\lambda(x)]$ (which depends on Π_λ) for every x in the support of μ. Given Π_λ, the function $\bar{f}_\lambda$ is the orthogonal projection of $f \in L^2([0,1]^d, \mu)$ on the subspace of functions that are constant on the cells of Π_λ. Since $\widehat{f}_\lambda$ belongs to this subspace given $\mathscr{D}_n$, we have conditionally on $(\Pi_\lambda, \mathscr{D}_n)$:

$$\mathbb{E}_X\big[(f(X) - \widehat{f}_\lambda(X))^2\big] = \mathbb{E}_X\big[(f(X) - \bar{f}_\lambda(X))^2\big] + \mathbb{E}_X\big[(\bar{f}_\lambda(X) - \widehat{f}_\lambda(X))^2\big].$$

This gives the following decomposition of the risk of $\widehat{f}_\lambda$ by taking the expectation over $(\Pi_\lambda, \mathscr{D}_n)$:

$$R(\widehat{f}_\lambda) = \mathbb{E}\big[(f(X) - \bar{f}_\lambda(X))^2\big] + \mathbb{E}\big[(\bar{f}_\lambda(X) - \widehat{f}_\lambda(X))^2\big]\,. \tag{2.20}$$

The first term of the sum, the *bias*, measures how close f is to its best approximation $\bar{f}_\lambda$ that is constant on the leaves of Π_λ (on average over Π_λ). The second term, the *variance*, measures how well the expected value $\bar{f}_\lambda(x)$ is estimated by the empirical average $\widehat{f}_\lambda(x)$ (on average over $\mathscr{D}_n, \Pi_\lambda$).

Note that (2.20) holds for the estimation risk *integrated over the hypercube* $[0,1]^d$, and not for the pointwise estimation risk. This is because in general, we have $\mathbb{E}_{\mathscr{D}_n}\big[\widehat{f}_\lambda(x)\big] \neq \bar{f}_\lambda(x)$: indeed, the cell $C_\lambda(x)$ may contain no data point in $\mathscr{D}_n$, in which case the estimate $\widehat{f}_\lambda(x)$ equals 0. It seems that a similar difficulty occurs for the decomposition in Genuer (2012); Arlot and Genuer (2014), which should only hold for the integrated risk.

Bias term. For each $x \in [0,1]^d$ in the support of μ, we have

$$|f(x) - \bar{f}_\lambda(x)| = \Big|\frac{1}{\mu(C_\lambda(x))}\int_{C_\lambda(x)} (f(x) - f(z))\mu(\mathrm{d}z)\Big| \leqslant \sup_{z \in C_\lambda(x)} |f(x) - f(z)| \leqslant L D_\lambda(x)^\beta,$$

where $D_\lambda(x)$ is the ℓ^2-diameter of $C_\lambda(x)$, since $f \in \mathscr{C}^{0,\beta}(L)$. By concavity of $x \mapsto x^\beta$ for $\beta \in (0,1]$ and Corollary 2.1, this implies

$$\mathbb{E}\big[(f(x) - \bar{f}_\lambda(x))^2\big] \leqslant L^2\mathbb{E}[D_\lambda(x)^{2\beta}] \leqslant L^2\mathbb{E}[D_\lambda(x)^2]^\beta \leqslant L^2\Big(\frac{4d}{\lambda^2}\Big)^\beta. \tag{2.21}$$

Integrating (2.21) with respect to μ yields the following bound on the bias:

$$\mathbb{E}\big[(f(X) - \bar{f}_\lambda(X))^2\big] \leqslant \frac{(4d)^\beta L^2}{\lambda^{2\beta}}\,. \tag{2.22}$$

Variance term. In order to bound the variance term, we use Proposition 2 in Arlot and Genuer (2014): if Π is a random tree partition of the unit cube in k cells (with $k \in \mathbf{N}^*$ deterministic) formed independently of the dataset $\mathscr{D}_n$, then

$$\mathbb{E}\big[(\bar{f}_\Pi(X) - \widehat{f}_\Pi(X))^2\big] \leqslant \frac{k}{n}(2\sigma^2 + 9\|f\|_\infty^2)\,. \tag{2.23}$$

Note that Proposition 2 in Arlot and Genuer (2014), stated in the case where the noise variance is constant, still holds when the noise variance is just upper bounded, based on Proposition 1 in Arlot (2008). For every $k \in \mathbf{N}^*$, applying (2.23) to the random partition $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$ conditionally on the event $\{K_\lambda = k\}$, we get

$$\begin{aligned}
\mathbb{E}\big[(\bar{f}_\lambda(X) - \widehat{f}_{\lambda,n}(X))^2\big] &= \sum_{k=1}^{\infty} \mathbb{P}(K_\lambda = k)\, \mathbb{E}[(\bar{f}_\lambda(X) - \widehat{f}_\lambda(X))^2 | K_\lambda = k] \\
&\leqslant \sum_{k=1}^{\infty} \mathbb{P}(K_\lambda = k)\, \frac{k}{n}\left(2\sigma^2 + 9\|f\|_\infty^2\right) \\
&= \frac{\mathbb{E}[K_\lambda]}{n}\left(2\sigma^2 + 9\|f\|_\infty^2\right).
\end{aligned}$$

Using Proposition 2.2, we obtain an upper bound of the variance term:

$$\mathbb{E}\big[(\bar{f}_\lambda(X) - \widehat{f}_\lambda(X))^2\big] \leqslant \frac{(1+\lambda)^d}{n}\left(2\sigma^2 + 9\|f\|_\infty^2\right). \tag{2.24}$$

Combining (2.22) and (2.24) leads to (2.7). Finally, the bound (2.8) follows by using $\lambda = \lambda_n$ in (2.7), which concludes the proof of Theorem 2.2. □

Proof of Theorem 2.3. Consider a Mondrian Forest

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

where the Mondrian Trees $\widehat{f}_\lambda^{(m)}$ for $m = 1, \ldots, M$ are based on independent partitions $\Pi_\lambda^{(m)} \sim \mathsf{MP}(\lambda, [0,1]^d)$. Also, for x in the support of μ let

$$\bar{f}_\lambda^{(m)}(x) = \mathbb{E}_X[f(X) | X \in C_\lambda^{(m)}(x)],$$

which depends on $\Pi_\lambda^{(m)}$. Let $\widetilde{f}_\lambda(x) = \mathbb{E}[\bar{f}_\lambda^{(m)}(x)]$, which is deterministic and does not depend on m. Denoting $\bar{f}_{\lambda,M}(x) = \frac{1}{M}\sum_{m=1}^{M} \bar{f}_\lambda^{(m)}(x)$, we have

$$\mathbb{E}[(\widehat{f}_{\lambda,M}(x) - f(x))^2] \leqslant 2\mathbb{E}[(\widehat{f}_{\lambda,M}(x) - \bar{f}_{\lambda,M}(x))^2] + 2\mathbb{E}[(\bar{f}_{\lambda,M}(x) - f(x))^2].$$

In addition, Jensen's inequality implies that

$$\mathbb{E}\big[(\widehat{f}_{\lambda,M}(x) - \bar{f}_{\lambda,M}(x))^2\big] \leqslant \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\big[(\widehat{f}_\lambda^{(m)}(x) - \bar{f}_\lambda^{(m)}(x))^2\big] = \mathbb{E}\big[(\widehat{f}_\lambda^{(1)}(x) - \bar{f}_\lambda^{(1)}(x))^2\big].$$

For every x we have that $\bar{f}_\lambda^{(m)}(x)$ are i.i.d. for $m = 1, \dots, M$ with expectation $\widetilde{f}_\lambda(x)$, so that

$$\mathbb{E}\big[(\bar{f}_{\lambda,M}(x) - f(x))^2\big] = (\widetilde{f}_\lambda(x) - f(x))^2 + \frac{\mathrm{Var}(\bar{f}_\lambda^{(1)}(x))}{M}\,.$$

Since $f \in \mathscr{C}^{1,\beta}(L)$ we have in particular that f is L-Lipschitz, hence

$$\mathrm{Var}(\bar{f}_\lambda^{(1)}(x)) \leqslant \mathbb{E}\big[(\bar{f}_\lambda^{(1)}(x) - f(x))^2\big] \leqslant L^2\mathbb{E}[D_\lambda(x)^2] \leqslant \frac{4dL^2}{\lambda^2}$$

for all $x \in [0,1]^d$, where we used Corollary 2.1 and where $D_\lambda(x)$ stands for the diameter of $C_\lambda(x)$. Consequently, taking the expectation with respect to X, we obtain

$$\mathbb{E}\big[(\widehat{f}_{\lambda,M}(X) - f(X))^2\big] \leqslant \frac{8dL^2}{M\lambda^2} + 2\mathbb{E}\big[(\widehat{f}_\lambda^{(1)}(X) - \bar{f}_\lambda^{(1)}(X))^2\big] + 2\mathbb{E}\big[(\widetilde{f}_\lambda(X) - f(X))^2\big].$$

The same upper bound holds also conditionally on $X \in B_\varepsilon := [\varepsilon, 1-\varepsilon]^d$:

$$\begin{aligned}\mathbb{E}\big[(\widehat{f}_{\lambda,M}(X) - f(X))^2 | X \in B_\varepsilon\big] \leqslant \frac{8dL^2}{M\lambda^2} &+ 2\mathbb{E}\big[(\widehat{f}_\lambda^{(1)}(X) - \bar{f}_\lambda^{(1)}(X))^2 | X \in B_\varepsilon\big] \\ &+ 2\mathbb{E}\big[(\widetilde{f}_\lambda(X) - f(X))^2 | X \in B_\varepsilon\big].\end{aligned} \tag{2.25}$$

Variance term. Recall that the distribution μ of X has a positive density $p : [0,1]^d \to \mathbf{R}_+^*$ which is C_p-Lipschitz, and recall that $p_0 = \inf_{x\in[0,1]^d} p(x)$ and $p_1 = \sup_{x\in[0,1]^d} p(x)$, both of which are positive and finite, since the continuous function p reaches its maximum and minimum over the compact set $[0,1]^d$. As shown in the proof of Theorem 2.2, the variance term satisfies

$$\mathbb{E}\big[(\bar{f}_\lambda^{(1)}(X) - \widehat{f}_{\lambda,n}^{(1)}(X))^2\big] \leqslant \frac{(1+\lambda)^d}{n}\left(2\sigma^2 + 9\|f\|_\infty^2\right).$$

Hence, the conditional variance in the decomposition (2.25) satisfies

$$\begin{aligned}\mathbb{E}\big[(\bar{f}_\lambda^{(1)}(X) - \widehat{f}_\lambda^{(1)}(X))^2 | X \in B_\varepsilon\big] &\leqslant \mathbb{P}(X \in B_\varepsilon)^{-1}\mathbb{E}\big[(\bar{f}_\lambda^{(1)}(X) - \widehat{f}_\lambda^{(1)}(X))^2\big] \\ &\leqslant p_0^{-1}(1-2\varepsilon)^{-d}\frac{(1+\lambda)^d}{n}(2\sigma^2 + 9\|f\|_\infty^2).\end{aligned} \tag{2.26}$$

Expression of $\widetilde{f}_\lambda$. It remains to control the bias term in the decomposition (2.25), which is the most involved part of the proof. Let us recall that $C_\lambda(x)$ stands for the cell of Π_λ which contains $x \in [0,1]^d$. We have

$$\widetilde{f}_\lambda(x) = \mathbb{E}\Big[\frac{1}{\mu(C_\lambda(x))}\int_{[0,1]^d} f(z)p(z)\mathbf{1}(z \in C_\lambda(x))\,\mathrm{d}z\Big] = \int_{[0,1]^d} f(z)\,F_{p,\lambda}(x,z)\,\mathrm{d}z, \tag{2.27}$$

where we defined

$$F_{p,\lambda}(x,z) = \mathbb{E}\left[\frac{p(z)\mathbf{1}(z \in C_\lambda(x))}{\mu(C_\lambda(x))}\right].$$

In particular, $\int_{[0,1]^d} F_{p,\lambda}(x,z)\mathrm{d}z = 1$ for any $x \in [0,1]^d$ (letting $f \equiv 1$ above). Let us also define the function F_λ, which corresponds to the case $p \equiv 1$:

$$F_\lambda(x,z) = \mathbb{E}\Big[\frac{\mathbf{1}(z \in C_\lambda(x))}{\mathrm{vol}(C_\lambda(x))}\Big],$$

where $\mathrm{vol}(C)$ stands for the volume of a box C.

Second order expansion. Assume that $f \in \mathscr{C}^{1+\beta}([0,1]^d)$ for some $\beta \in (0,1]$. This implies that

$$|f(z) - f(x) - \nabla f(x)^\top (z-x)| = \Big| \int_0^1 [\nabla f(x + t(z-x)) - \nabla f(x)]^\top (z-x) \mathrm{d}t \Big| \\ \leqslant \int_0^1 L(t\|z-x\|)^\beta \|z-x\| \mathrm{d}t \leqslant L\|z-x\|^{1+\beta} .$$

Now, by the triangle inequality,

$$\Big| \Big| \int_{[0,1]^d} (f(z) - f(x)) F_{p,\lambda}(x,z) \mathrm{d}z \Big| - \Big| \int_{[0,1]^d} \nabla f(x)^\top (z-x) F_{p,\lambda}(x,z) \mathrm{d}z \Big| \Big| \\ \leqslant \Big| \int_{[0,1]^d} (f(z) - f(x) - \nabla f(x)^\top (z-x)) F_{p,\lambda}(x,z) \mathrm{d}z \Big| \\ \leqslant L \int_{[0,1]^d} \|z-x\|^{1+\beta} F_{p,\lambda}(x,z) \mathrm{d}z ,$$

so that, using together $\int F_{p,\lambda}(x,z)\mathrm{d}z = 1$ and (2.27), we obtain

$$|\widetilde{f}_\lambda(x) - f(x)| \leqslant \Big| \nabla f(x)^\top \underbrace{\int_{[0,1]^d} (z-x) F_{p,\lambda}(x,z) \mathrm{d}z}_{:=A} \Big| + L \underbrace{\int_{[0,1]^d} \|z-x\|^{1+\beta} F_{p,\lambda}(x,z) \mathrm{d}z}_{:=B} . \quad (2.28)$$

Hence, it remains to control the two terms A, B from Equation (2.28). We will start by expressing $F_{p,\lambda}$ in terms of p, using the distribution of the cell $C_\lambda(x)$ given by Proposition 2.1 above. Next, both terms will be bounded by approximating $F_{p,\lambda}$ by F_λ and controlling these terms for F_λ (this is done in Technical Lemma 2.1 below).

Explicit form of $F_{p,\lambda}$. First, we provide an explicit form of $F_{p,\lambda}$ in terms of p. We start by determining the distribution of the cell $C_\lambda(x)$ conditionally on the event $z \in C_\lambda(x)$. Let $C = C(x,z) = \prod_{1 \leqslant j \leqslant d} [x_j \wedge z_j, x_j \vee z_j] \subseteq [0,1]^d$ be the smallest box containing both x and z; also, let $a_j = x_j \wedge z_j$, $b_j = x_j \vee z_j$, $a = (a_j)_{1 \leqslant j \leqslant d}$ and $b = (b_j)_{1 \leqslant j \leqslant d}$. Note that $z \in C_\lambda(x)$ if and only if Π_λ does not cut C. Since $C = C(x,z) = C(a,b)$, we have that $z \in C_\lambda(x)$ if and only if $b \in C_\lambda(a)$, and in this case $C_\lambda(x) = C_\lambda(a)$. In particular, the conditional distribution of $C_\lambda(x)$ given $z \in C_\lambda(x)$ equals the conditional distribution of $C_\lambda(a)$ given $b \in C_\lambda(a)$.

Write $C_\lambda(a) = \prod_{j=1}^d [L_{\lambda,j}(a), R_{\lambda,j}(a)]$; by Proposition 2.1, we have $L_{\lambda,j}(a) = (a_j - \lambda^{-1} E_{j,L}) \vee 0$, $R_{\lambda,j}(a) = (a_j + \lambda^{-1} E_{j,R}) \wedge 1$, where $E_{j,L}, E_{j,R}$, $1 \leqslant j \leqslant d$ are i.i.d. $\mathsf{Exp}(1)$ random variables. Note that $b \in C_\lambda(a)$ is equivalent to $R_{\lambda,j}(a) \geqslant b_j$ for $j = 1, \dots, d$, *i.e.* to $E_{j,R} \geqslant \lambda(b_j - a_j)$. By the memory-less property of the exponential distribution, the distribution of $E_{j,R} - \lambda(b_j - a_j)$ conditionally on $E_{j,R} \geqslant \lambda(b_j - a_j)$ is $\mathsf{Exp}(1)$. As a result (using the independence of the variables $E_{j,L}, E_{j,R}$), we obtain the following statement:

> Conditionally on $b \in C_\lambda(a)$, the coordinates $L_{\lambda,j}(a), R_{\lambda,j}(a)$, $1 \leqslant j \leqslant d$, are distributed as $(a_j - \lambda^{-1} E'_{j,L}) \vee 0, (b_j + \lambda^{-1} E'_{j,R}) \wedge 1$, where $E'_{j,L}, E'_{j,R}$ are i.i.d. $\mathsf{Exp}(1)$ random variables.

Hence, the distribution of $C_\lambda(x)$ conditionally on $z \in C_\lambda(x)$ has the same distribution as

$$C_\lambda(x,z) := \prod_{j=1}^d \left[(x_j \wedge z_j - \lambda^{-1} E_{j,L}) \vee 0, (x_j \vee z_j + \lambda^{-1} E_{j,R}) \wedge 1 \right] \quad (2.29)$$

where $E_{1,L}, E_{1,R}, \dots, E_{d,L}, E_{d,R}$ are i.i.d. $\mathsf{Exp}(1)$ random variables. In addition, note that $z \in C_\lambda(x)$ if and only if the restriction of Π_λ to $C(x,z)$ has no split (*i.e.*, its first sampled split occurs after time λ). Since this restriction is distributed as $\mathsf{MP}(\lambda, C(x,z))$ using Fact 2.2, this occurs with probability $\exp(-\lambda|C(x,z)|) = \exp(-\lambda\|x-z\|_1)$. Therefore,

$$\begin{aligned} F_{p,\lambda}(x,z) &= \mathbb{P}(z \in C_\lambda(x))\,\mathbb{E}\Big[\frac{p(z)}{\mu(C_\lambda(x))}\,\Big|\, z \in C_\lambda(x)\Big] \\ &= e^{-\lambda\|x-z\|_1}\,\mathbb{E}\bigg[\Big\{\int_{C_\lambda(x,z)} \frac{p(y)}{p(z)}\mathrm{d}y\Big\}^{-1}\bigg], \end{aligned} \tag{2.30}$$

where $C_\lambda(x,z)$ is as in (2.29). In addition, applying (2.30) to $p \equiv 1$ yields

$$F_\lambda(x,z) = \lambda^d e^{-\lambda\|x-z\|_1} \prod_{1\leqslant j\leqslant d} \mathbb{E}\Big[\big\{\lambda|x_j - z_j| + E_{j,L} \wedge \lambda(x_j \wedge z_j) + E_{j,R} \wedge \lambda(1 - x_j \vee z_j)\big\}^{-1}\Big]. \tag{2.31}$$

The following technical Lemma, whose proof is given in Section 2.8.6, will prove useful in what follows.

Lemma 2.1. *The function $F_{p,\lambda}$ given by (2.31) satisfies, for any $x \in [0,1]^d$,*

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

Control of the term B in Equation (2.28). It follows from (2.30) and from the bound $p(y)/p(z) \geqslant p_0/p_1$ that

$$F_{p,\lambda}(x,z) \leqslant \frac{p_1}{p_0} F_\lambda(x,z), \tag{2.32}$$

so that

$$\begin{aligned} \int_{[0,1]^d} \|z-x\|^{1+\beta} F_{p,\lambda}(x,z)\mathrm{d}z &\leqslant \frac{p_1}{p_0}\int_{[0,1]^d} \|z-x\|^{1+\beta} F_\lambda(x,z)\mathrm{d}z \\ &\leqslant \frac{p_1}{p_0}\Big(\int_{[0,1]^d} \|z-x\|^2 F_\lambda(x,z)\mathrm{d}z\Big)^{(1+\beta)/2} \qquad (2.33) \\ &\leqslant \frac{p_1}{p_0}\Big(\frac{2d}{\lambda^2}\Big)^{(1+\beta)/2}, \qquad (2.34) \end{aligned}$$

where (2.33) follows from the concavity of $x \mapsto x^{(1+\beta)/2}$ for $\beta \in (1,2]$, while (2.34) comes from Lemma 2.1.

Control of the term A in Equation (2.28). It remains to control $A = \int_{[0,1]^d}(z-x)F_{p,\lambda}(x,z)\mathrm{d}z$. Again, this quantity is controlled in the case of a uniform density ($p \equiv 1$) in Lemma 2.1. However, this time the crude bound (2.32) is no longer sufficient, since we need first-order terms to compensate in order to obtain the optimal rate. Rather, we will show that $F_{p,\lambda}(x,z) = (1 + O(\|x-z\|) + O(1/\lambda))F_\lambda(x,z)$.

A first upper bound on $|F_{p,\lambda}(x,z) - F_\lambda(x,z)|$. Since p is C_p-Lipschitz and lower bounded by p_0, we have

$$\left|\frac{p(y)}{p(z)} - 1\right| = \frac{|p(y) - p(z)|}{p(z)} \leqslant \frac{C_p}{p_0}\|y - z\| \leqslant \frac{C_p}{p_0}\operatorname{diam} C_\lambda(x,z) \tag{2.35}$$

for every $y \in C_\lambda(x,z)$, so that

$$1 - \frac{C_p}{p_0}\operatorname{diam} C_\lambda(x,z) \leqslant \frac{p(y)}{p(z)} \leqslant 1 + \frac{C_p}{p_0}\operatorname{diam} C_\lambda(x,z).$$

Integrating over $C_\lambda(x,z)$ and using $p(y)/p(z) \geqslant p_0/p_1$ gives

$$\begin{aligned}\left\{1 + \frac{C_p}{p_0}\operatorname{diam} C_\lambda(x,z)\right\}^{-1} \operatorname{vol} C_\lambda(x,z)^{-1} &\leqslant \left\{\int_{C_\lambda(x,z)} \frac{p(y)}{p(z)}\mathrm{d}y\right\}^{-1} \\ &\leqslant \left\{\left(1 - \frac{C_p}{p_0}\operatorname{diam} C_\lambda(x,z)\right) \vee \frac{p_0}{p_1}\right\}^{-1} \operatorname{vol} C_\lambda(x,z)^{-1}. \end{aligned} \tag{2.36}$$

In addition, since $(1+u)^{-1} \geqslant 1 - u$ for $u \geqslant 0$, we have

$$\left\{1 + \frac{C_p}{p_0}\operatorname{diam} C_\lambda(x,z)\right\}^{-1} \geqslant 1 - \frac{C_p}{p_0}\operatorname{diam} C_\lambda(x,z)\,,$$

so that setting $a := \left(1 - \frac{C_p}{p_0}\operatorname{diam} C_\lambda(x,z)\right) \vee \frac{p_0}{p_1} \in (0,1]$ gives

$$a^{-1} - 1 = \frac{1-a}{a} \leqslant \frac{(C_p/p_0)\operatorname{diam} C_\lambda(x,z)}{p_0/p_1} = \frac{p_1 C_p}{p_0^2}\operatorname{diam} C_\lambda(x,z).$$

Now, Equation (2.36) implies that

$$\begin{aligned}-\frac{C_p}{p_0}\operatorname{diam} C_\lambda(x,z) \operatorname{vol} C_\lambda(x,z)^{-1} &\leqslant \left\{\int_{C_\lambda(x,z)} \frac{p(y)}{p(z)}\mathrm{d}y\right\}^{-1} - \operatorname{vol} C_\lambda(x,z)^{-1} \\ &\leqslant \frac{p_1 C_p}{p_0^2}\operatorname{diam} C_\lambda(x,z) \operatorname{vol} C_\lambda(x,z)^{-1}\,.\end{aligned}$$

Taking the expectation over $C_\lambda(x,z)$ and using (2.30) leads to

$$\begin{aligned}-\frac{C_p}{p_0}\mathbb{E}\left[\operatorname{diam} C_\lambda(x,z) \operatorname{vol} C_\lambda(x,z)^{-1}\right] &\leqslant e^{\lambda\|x-z\|_1}(F_{p,\lambda}(x,z) - F_\lambda(x,z)) \\ &\leqslant \frac{p_1 C_p}{p_0^2}\mathbb{E}\left[\operatorname{diam} C_\lambda(x,z) \operatorname{vol} C_\lambda(x,z)^{-1}\right]\end{aligned}$$

so that

$$|F_{p,\lambda}(x,z) - F_\lambda(x,z)| \leqslant \frac{p_1 C_p}{p_0^2} e^{-\lambda\|x-z\|_1} \times \mathbb{E}\left[\operatorname{diam} C_\lambda(x,z) \operatorname{vol} C_\lambda(x,z)^{-1}\right]. \tag{2.37}$$

Control of $\mathbb{E}\big[\operatorname{diam} C_\lambda(x,z)\operatorname{vol} C_\lambda(x,z)^{-1}\big]$. Let us define the interval

$$C_\lambda^j(x,z) := \big[(x_j \wedge z_j - \lambda^{-1}E_{j,L}) \vee 0, (x_j \vee z_j + \lambda^{-1}E_{j,R}) \wedge 1\big]$$

and let $|C_\lambda^j(x,z)| = (x_j \vee z_j + \lambda^{-1}E_{j,R}) \wedge 1 - (x_j \wedge z_j - \lambda^{-1}E_{j,L}) \vee 0$ be its length. We have $\operatorname{diam} C_\lambda(x,z) \leqslant \operatorname{diam}_{\ell^1} C_\lambda(x,z)$ using the triangular inequality, so that

$$\begin{aligned}
\mathbb{E}\big[\operatorname{diam} C_\lambda(x,z)\operatorname{vol} C_\lambda(x,z)^{-1}\big] &\leqslant \mathbb{E}\Big[\sum_{j=1}^d |C_\lambda^j(x,z)|\operatorname{vol} C_\lambda(x,z)^{-1}\Big] \\
&= \sum_{j=1}^d \mathbb{E}\Big[|C_\lambda^j(x,z)|\prod_{l=1}^d |C_\lambda^l(x,z)|^{-1}\Big] = \sum_{j=1}^d \mathbb{E}\Big[\prod_{l\neq j}|C_\lambda^l(x,z)|^{-1}\Big] \\
&\leqslant \sum_{j=1}^d \mathbb{E}\Big[|C_\lambda^j(x,z)|\Big]\mathbb{E}\Big[|C_\lambda^j(x,z)|^{-1}\Big]\mathbb{E}\Big[\prod_{l\neq j}|C_\lambda^l(x,z)|^{-1}\Big] \qquad (2.38)\\
&= \sum_{j=1}^d \mathbb{E}\Big[|C_\lambda^j(x,z)|\Big] \times \mathbb{E}\Big[\prod_{l=1}^d |C_\lambda^l(x,z)|^{-1}\Big] \qquad (2.39)\\
&= \mathbb{E}\big[\operatorname{diam}_{\ell^1} C_\lambda(x,z)\big] \times \exp(\lambda\|x-z\|_1)F_\lambda(x,z). \qquad (2.40)
\end{aligned}$$

Inequality (2.38) relies on the fact that $\mathbb{E}[X]\mathbb{E}[X^{-1}] \geqslant 1$ for any positive random variable X with $X = |C_\lambda^j(x,z)|$. Equality (2.39) comes from the independence of $|C_\lambda^1(x,z)|, \dots, |C_\lambda^d(x,z)|$. Multiplying both sides of (2.40) by $e^{-\lambda\|x-z\|_1}$ leads to

$$e^{-\lambda\|x-z\|_1}\mathbb{E}\big[\operatorname{diam} C_\lambda(x,z)\operatorname{vol} C_\lambda(x,z)^{-1}\big] \leqslant \mathbb{E}\big[\operatorname{diam}_{\ell^1} C_\lambda(x,z)\big]F_\lambda(x,z)\,. \qquad (2.41)$$

In addition,

$$\begin{aligned}
\mathbb{E}\big[\operatorname{diam}_{\ell^1} C_\lambda(x,z)\big] &\leqslant \sum_{j=1}^d \mathbb{E}\big[|x_j - z_j| + \lambda^{-1}(E_{j,R}+E_{j,L})\big] \\
&= \|x-z\|_1 + \frac{2d}{\lambda}\,. \qquad (2.42)
\end{aligned}$$

Finally, combining Equations (2.37), (2.41) and (2.42) gives

$$|F_{p,\lambda}(x,z) - F_\lambda(x,z)| \leqslant \frac{p_1 C_p}{p_0^2}\Big(\|x-z\|_1 + \frac{2d}{\lambda}\Big)F_\lambda(x,z)\,. \qquad (2.43)$$

Control of A. From (2.43), we can control $\int_{[0,1]^d}(z-x)F_{p,\lambda}(x,z)\mathrm{d}z$ by approximating $F_{p,\lambda}$ by F_λ. Indeed, we have

$$\Big\|\int_{[0,1]^d}(z-x)F_{p,\lambda}(x,z)\mathrm{d}z - \int_{[0,1]^d}(z-x)F_\lambda(x,z)\mathrm{d}z\Big\| \leqslant \int_{[0,1]^d}\|z-x\|\cdot|F_{p,\lambda}(x,z)-F_\lambda(x,z)|\mathrm{d}z, \qquad (2.44)$$

with

$$\begin{aligned}
&\int_{[0,1]^d} \|z-x\| \times |F_{p,\lambda}(x,z) - F_\lambda(x,z)| \mathrm{d}z \\
&\leqslant \frac{p_1 C_p}{p_0^2} \int_{[0,1]^d} \|z-x\| \Big[\|x-z\|_1 + \frac{2d}{\lambda} \Big] F_\lambda(x,z) \mathrm{d}z \qquad \text{(by (2.43))} \\
&\leqslant \frac{p_1 C_p}{p_0^2} \Big[\sqrt{d} \int_{[0,1]^d} \|z-x\|^2 F_\lambda(x,z) \mathrm{d}z + \frac{2d}{\lambda} \int_{[0,1]^d} \|z-x\| F_\lambda(x,z) \mathrm{d}z \Big] \\
&\leqslant \frac{p_1 C_p}{p_0^2} \Big[\frac{d\sqrt{d}}{\lambda^2} + \frac{2d}{\lambda} \Big(\int_{[0,1]^d} \|z-x\|^2 F_\lambda(x,z) \mathrm{d}z \Big)^{1/2} \Big],
\end{aligned}$$

where we used the inequalities $\|v\| \leqslant \|v\|_1 \leqslant \sqrt{d}\|v\|$ as well as the Cauchy-Schwarz inequality. Hence, using Lemma 2.1, we end up with

$$\begin{aligned}
\int_{[0,1]^d} \|z-x\| \times |F_{p,\lambda}(x,z) - F_\lambda(x,z)| \mathrm{d}z &\leqslant \frac{p_1 C_p}{p_0^2} \Big[\frac{d\sqrt{d}}{\lambda^2} + \frac{2d}{\lambda} \sqrt{\frac{d}{\lambda^2}} \Big] \\
&= \frac{p_1 C_p}{p_0^2} \frac{3d\sqrt{d}}{\lambda^2} \,. \qquad (2.45)
\end{aligned}$$

Inequalities (2.44) and (2.45) together with Lemma 2.1 entail that

$$\begin{aligned}
\Big\| \int_{[0,1]^d} (z-x) F_{p,\lambda}(x,z) \mathrm{d}z \Big\|^2 &\leqslant 2 \Big\| \int_{[0,1]^d} (z-x) F_\lambda(x,z) \mathrm{d}z \Big\|^2 \\
&\quad + 2 \Big(\int_{[0,1]^d} \|z-x\| |F_{p,\lambda}(x,z) - F_\lambda(x,z)| \mathrm{d}z \Big)^2 \\
&\leqslant \frac{18}{\lambda^2} \sum_{j=1}^d e^{-\lambda[x_j \wedge (1-x_j)]} + 2 \Big(\frac{p_1 C_p}{p_0^2} \frac{3d\sqrt{d}}{\lambda^2} \Big)^2 . \qquad (2.46)
\end{aligned}$$

Control of the bias. The upper bound (2.28) on the bias writes

$$(\widetilde{f}_\lambda(x) - f(x))^2 \leqslant (|\nabla f(x)^\top A| + LB)^2 \leqslant 2(|\nabla f(x)\|^2 \times \|A\|^2 + L^2 B^2) \,,$$

so that plugging the bounds (2.34) of B and (2.46) of $\|A\|$ gives

$$\begin{aligned}
(\widetilde{f}_\lambda(x) - f(x))^2 &\leqslant 2L^2 \Big[\frac{18}{\lambda^2} \sum_{j=1}^d e^{-\lambda[x_j \wedge (1-x_j)]} + 2 \Big(\frac{p_1 C_p}{p_0^2} \frac{3d\sqrt{d}}{\lambda^2} \Big)^2 \Big] + 2L^2 \frac{p_1}{p_0} \Big(\frac{2d}{\lambda^2} \Big)^{(1+\beta)/2} \\
&\leqslant \frac{36L^2}{\lambda^2} \sum_{j=1}^d e^{-\lambda[x_j \wedge (1-x_j)]} + \frac{36L^2 d^3}{\lambda^4} \Big(\frac{p_1 C_p}{p_0^2} \Big)^2 + \frac{8L^2 d^{1+\beta}}{\lambda^{2(1+\beta)}} \Big(\frac{p_1}{p_0} \Big)^2 .
\end{aligned}$$

By integrating over X conditionally on $X \in B_\varepsilon$, this implies

$$\mathbb{E}\big[(\widetilde{f}_\lambda(X) - f(X))^2 | X \in B_\varepsilon \big] \leqslant \frac{36L^2}{\lambda^2} \psi_\varepsilon(\lambda) + \frac{36L^2 d^3}{\lambda^4} \Big(\frac{p_1 C_p}{p_0^2} \Big)^2 + \frac{8L^2 d^{1+\beta}}{\lambda^{2(1+\beta)}} \Big(\frac{p_1}{p_0} \Big)^2 , \qquad (2.47)$$

where we have, using the fact that $p_0 \leqslant p(x) \leqslant p_1$ for any $x \in [0,1]$,

$$\psi_\varepsilon(\lambda) := \sum_{j=1}^{d} \mathbb{E}\big[e^{-\lambda[X_j \wedge (1-X_j)]} | X \in B_\varepsilon\big] \leqslant \frac{dp_1}{p_0(1-2\varepsilon)^d} \int_\varepsilon^{1-\varepsilon} e^{-\lambda[u\wedge(1-u)]} \mathrm{d}u$$
$$= \frac{dp_1}{p_0(1-2\varepsilon)^d} \times 2\int_\varepsilon^{1/2} e^{-\lambda u} \mathrm{d}u \leqslant \frac{e^{-\lambda\varepsilon}}{\lambda} \frac{2dp_1}{p_0(1-2\varepsilon)^d}\,.$$

Conclusion. The decomposition (2.25), together with the bounds (2.26) on the variance and (2.47) on the bias lead to inequality (2.9) from the statement of Theorem 2.3. In particular, if $\varepsilon \in (0, \frac{1}{2})$ is fixed, inequality (2.9) writes

$$\mathbb{E}\big[(\widehat{f}_{\lambda,M}(X) - f(X))^2 | X \in B_\varepsilon\big] = O\Big(\frac{\lambda^d}{n} + \frac{L^2}{\lambda^{2(1+\beta)}} + \frac{L^2}{M\lambda^2}\Big).$$

One can optimize the right-hand side by setting $\lambda = \lambda_n \asymp L^{2/(d+2s)} n^{1/(d+2s)}$ and $M = M_n \gtrsim \lambda_n^{2\beta} \asymp L^{4\beta/(d+2s)} n^{2\beta/(d+2s)}$ with $s = 1+\beta \in (1,2]$. This leads to the minimax rate $O(L^{2d/(d+2s)} n^{-2s/(d+2s)})$ for $f \in \mathscr{C}^{1,\beta}(L)$ as announced in the statement of Theorem 2.3.

On the other hand, we have $e^{-\lambda\varepsilon} = 1$ whenever $\varepsilon = 0$, so that inequality (2.9) becomes in this case

$$\mathbb{E}[(\widehat{f}_{\lambda,M}(X) - f(X))^2] \leqslant O\Big(\frac{\lambda^d}{n} + \frac{L^2}{\lambda^{3\wedge(2s)}} + \frac{L^2}{M\lambda^2}\Big).$$

When $2s \leqslant 3$ (*i.e.* $\beta \leqslant 1/2$), this leads to the same rate as above, with the same choice of parameters. When $2s > 3$, this leads to the suboptimal rate $O(L^{2d/(d+3)} n^{-3/(d+3)})$ with the choice $M_n \gtrsim \lambda_n \asymp L^{2/(d+3)} n^{1/(d+3)}$. This concludes the proof of all the claims from Theorem 2.3. □

Notation or formula	Description
$\mathbf{v} \in \{0,1\}^*$	A node
$\mathscr{D}_n$	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. $\mathsf{MP}(\lambda, C)$
$C_\lambda(x)$	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
f	True regression function: $f(X) = \mathbb{E}[Y\lvert X]$
$\widehat{f}^{(m)}_{\lambda,n}(x)$	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 $\Pi_{\lambda,M} = (\Pi^{(1)}_\lambda, \dots, \Pi^{(M)}_\lambda)$
$\bar{f}^{(m)}_\lambda(x)$	Expected value of f inside the cell $C^{(m)}_\lambda(x)$
$\widetilde{f}_\lambda(x)$	Expected value of $\bar{f}^{(m)}_\lambda(x)$ over $\Pi^{(m)}_\lambda \sim \mathsf{MP}(\lambda,[0,1]^d)$
$\mathcal{N}(\mathcal{T}), \mathcal{N}^\circ(\mathcal{T}), \mathcal{L}(\mathcal{T})$	Nodes, interior nodes and leaves of a tree $\mathcal{T}$
$\Sigma = (\sigma_\mathbf{v})_{\mathbf{v}\in\mathcal{N}^\circ(\mathcal{T})}$	Set of splits for all nodes in the tree $\mathcal{T}$
$\sigma_\mathbf{v} = (j_\mathbf{v}, s_\mathbf{v})$	A split at node $\mathbf{v}$ characterized by its split dimension $j_\mathbf{v} \in \{1,\dots,d\}$ and threshold $s_\mathbf{v} \in [0,1]$
$\tau_\mathbf{v}$	Birth time of a node $\mathbf{v}$

Table 2.1: Notations and definitions used in this section

2.8 Remaining proofs

In this section, we gather several proofs and technical details and definitions that were omitted in the rest of the chapter. Namely, we start with a glossary of notations (Table 2.1), then give extra definitions and notations for trees and nested trees partitions in Section 2.8.1. Then, we provide proofs that were omitted in the chapter by order of appearance, namely the proofs of Proposition 2.2, Theorem 2.1, Proposition 2.3, Proposition 2.4 and Lemma 2.1.

2.8.1 Specific notations

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

2.8.1.1 Trees and nested tree partitions

A decision tree $(\mathcal{T}, \Sigma)$ is composed of the following components:

- A finite rooted ordered binary tree $\mathcal{T}$, with nodes $\mathcal{N}(\mathcal{T})$, interior nodes $\mathcal{N}^\circ(\mathcal{T})$ and leaves $\mathcal{L}(\mathcal{T})$ (so that $\mathcal{N}(\mathcal{T})$ is the disjoint union of $\mathcal{N}^\circ(\mathcal{T})$ and $\mathcal{L}(\mathcal{T})$). The nodes $\mathbf{v} \in \mathcal{N}(\mathcal{T})$ 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 $\mathcal{T}$ 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* $\Sigma = (\sigma_\mathbf{v})_{\mathbf{v}\in\mathcal{N}^\circ(\mathcal{T})}$ 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 $\Pi = (\mathcal{T}, \Sigma)$ a partition $(C_\mathbf{v})_{\mathbf{v}\in\mathcal{L}(\mathcal{T})}$ of the unit cube $[0,1]^d$, called a *tree partition* (or *guillotine partition*). For each node $\mathbf{v} \in \mathcal{N}(\mathcal{T})$, we define a hyper-rectangular region $C_\mathbf{v}$ recursively:

- The cell associated to the root of $\mathcal{T}$ is $[0,1]^d$;
- For each $\mathbf{v} \in \mathcal{N}^\circ(\mathcal{T})$, 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 $(C_\mathbf{v})_{\mathbf{v}\in\mathcal{L}(\mathcal{T})}$ form a partition of $[0,1]^d$ by construction. In what follows, we will identify a tree with splits $(\mathcal{T}, \Sigma)$ with its associated tree partition, and a node $\mathbf{v} \in \mathcal{N}(\mathcal{T})$ 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 2.2 (Nested tree partitions)**.** A tree partition $\Pi' = (\mathcal{T}', \Sigma')$ is a *refinement* of the tree partition $\Pi = (\mathcal{T}, \Sigma)$ if $\mathcal{T}$ is a subtree of $\mathcal{T}'$ and, for every $\mathbf{v} \in \mathcal{N}(\mathcal{T}) \subseteq \mathcal{N}(\mathcal{T}')$, $\sigma_\mathbf{v} = \sigma'_\mathbf{v}$. 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 $\mathcal{N}(\mathbf{T}) = \bigcup_{t\geqslant 0} \mathcal{N}(\mathcal{T}_t) \subseteq \{0,1\}^*$. For each $\mathbf{v} \in \mathcal{N}(\mathcal{T})$, let $\tau_\mathbf{v} = \inf\{t \geqslant 0 \mid \mathbf{v} \in \mathcal{N}(\mathcal{T}_t)\} < \infty$ 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 $\mathfrak{T} = (\tau_\mathbf{v})_{\mathbf{v}\in\mathcal{N}(\mathbf{T})}$.

2.8.1.2 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:

$$j_\mathbf{v} = \operatorname*{arg\,min}_{j=1,\dots,d} \frac{e^j_\mathbf{v}}{b^j_\mathbf{v} - a^j_\mathbf{v}}, \qquad s_\mathbf{v} = a^{j_\mathbf{v}}_\mathbf{v} + (b^{j_\mathbf{v}}_\mathbf{v} - a^{j_\mathbf{v}}_\mathbf{v}) \cdot u^{j_\mathbf{v}}_\mathbf{v}, \tag{2.48}$$

where we break ties in the choice of $j_\mathbf{v}$ e.g., by choosing the smallest index j in the arg min. 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 2.3 (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 $\Phi_{\lambda,C}((E^j_{\mathbf{v}}, U^j_{\mathbf{v}})_{\mathbf{v},j})$.

2.8.2 Proof of Proposition 2.2

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 2.3) 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 $j = 1, \dots, d$, 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:

$$\widetilde{T}_{\mathbf{v}} = \min_{1\leqslant j\leqslant d} T'^j_{\mathbf{v}_j(\mathbf{v})}, \qquad \widetilde{J}_{\mathbf{v}} := \operatorname*{arg\,min}_{1\leqslant j\leqslant d} T'^j_{\mathbf{v}_j(\mathbf{v})}, \qquad \widetilde{S}_{\mathbf{v}} = S'^j_{\mathbf{v}_j(\mathbf{v})}, \tag{2.49}$$

$$\mathbf{v}_j(\mathbf{v}a) = \begin{cases} \mathbf{v}_j(\mathbf{v})a & \text{if } j = \widetilde{J}_{\mathbf{v}} \\ \mathbf{v}_j(\mathbf{v}) & \text{else.} \end{cases} \tag{2.50}$$

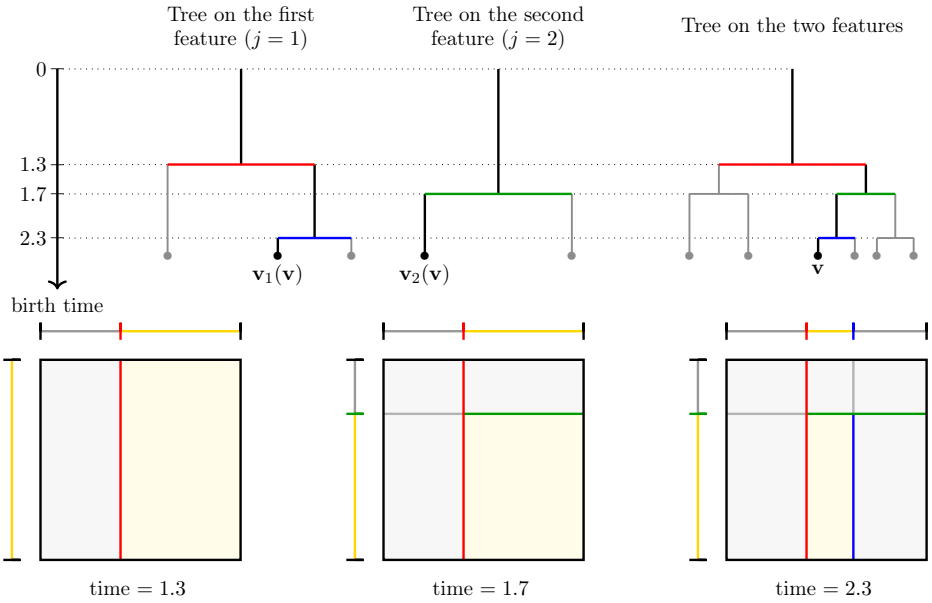

Figure 2.3: Modified construction in dimension two. At the top, from left to right: trees associated to partitions Π'^1, Π'^2 and $\widetilde{\Pi}$ respectively. At the bottom, from left to right: successive splits in $\widetilde{\Pi}$ leading to the leaf $\mathbf{v}$ (depicted in yellow).

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 2.3.

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 (2.49) 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 (2.49) 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 2.1, 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:

$$\mathbb{E}[\widetilde{K}_\lambda] = \prod_{j=1}^{d} (1 + \lambda|C^j|)\,. \tag{2.51}$$

Equality of $\mathbb{E}[K_\lambda]$ and $\mathbb{E}[\widetilde{K}_\lambda]$. In order to establish Proposition 2.2, 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,

$$\mathbb{E}[K_\lambda] = 1 + \sum_{\mathbf{v} \in \{0,1\}^*} \mathbb{P}(T_{\mathbf{v}} \leqslant \lambda) \tag{2.52}$$

$$\mathbb{E}[\widetilde{K}_\lambda] = 1 + \sum_{\mathbf{v} \in \{0,1\}^*} \mathbb{P}(\widetilde{T}_{\mathbf{v}} \leqslant \lambda)\,. \tag{2.53}$$

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 $\widetilde{\mathcal{F}}_{\mathbf{v}} = \sigma((\widetilde{T}_{\mathbf{v}'}, \widetilde{J}_{\mathbf{v}'}, \widetilde{S}_{\mathbf{v}'}), \mathbf{v}' \sqsubset \mathbf{v})$ has the same form as the conditional distribution of $(T_{\mathbf{v}}, J_{\mathbf{v}}, S_{\mathbf{v}})$ given $\mathcal{F}_{\mathbf{v}} = \sigma((T_{\mathbf{v}'}, J_{\mathbf{v}'}, S_{\mathbf{v}'}), \mathbf{v}' \sqsubset \mathbf{v})$, 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 $\mathcal{F}_{\mathbf{v}}$ 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 $\widetilde{\mathcal{F}}_{\mathbf{v}}$ 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 $\mathcal{F}_{\mathbf{v}}$. Hence, conditionally on $\mathcal{F}_{\mathbf{v}}$, $E^j_{\mathbf{v}}, U^j_{\mathbf{v}}$, $1 \leqslant j \leqslant d$ 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 $T_1, \ldots, T_d$ are independent exponential random variables of intensities $\lambda_1, \ldots, \lambda_d$, and if $T = \min_{1 \leqslant j \leqslant d} T_j$ and $J = \arg\min_{1 \leqslant j \leqslant d} T_j$, 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 $\mathcal{F}_{\mathbf{v}}$, $T_{\mathbf{v}} - \tau_{\mathbf{v}} = \min_{1 \leqslant j \leqslant d} E^j_{\mathbf{v}}/|C^j_{\mathbf{v}}| \sim \mathsf{Exp}(\sum_{j=1}^d |C^j_{\mathbf{v}}|) = \mathsf{Exp}(|C_{\mathbf{v}}|)$, $J_{\mathbf{v}} := \arg\min_{1 \leqslant j \leqslant d} E^j_{\mathbf{v}}/|C^j_{\mathbf{v}}|$ 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 $\widetilde{\mathcal{F}}_{\mathbf{v}}$. 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 $0 \leqslant v \leqslant \mathrm{depth}(\mathbf{v})$. 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 2.2. *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 $a_1, \ldots, a_d > 0$. For $1 \leqslant j \leqslant d$, define the sequence $(T^j_v, L^j_v)_{v \in \mathbf{N}}$ as follows*:

- $L_0^j = a_j$, $T_0^j = \frac{E_0^j}{a_j}$;
- *for* $v \in \mathbf{N}$, $L_{v+1}^j = U_v^j L_v^j$, $T_{v+1}^j = T_v^j + \frac{E_{v+1}^j}{L_{v+1}^j}$.

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

- $\widetilde{V}_0^j = 0$ *for* $j = 1, \dots, d$.
- *for* $v \in \mathbf{N}$, *given* $\widetilde{V}_v^j$ $(1 \leqslant j \leqslant d)$, *denoting* $\widetilde{T}_v^j = T_{\widetilde{V}_v^j}^j$ *and* $\widetilde{U}_v^j = U_{\widetilde{V}_v^j}^j$, *set*

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

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

- $\widetilde{J}_v, \widetilde{T}_v, \widetilde{U}_v$ *are independent*,
- $\mathbb{P}(\widetilde{J}_v = j | \mathcal{F}_v) = \widetilde{L}_v^j / (\sum_{j'=1}^d \widetilde{L}_v^{j'})$,
- $\widetilde{T}_v - \widetilde{T}_{v-1} \sim \mathsf{Exp}(\sum_{j=1}^d \widetilde{L}_v^j)$ (*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 2.2, 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 $\mathcal{F}_v = \mathcal{F}_{\mathbf{v}_v}$. Applying Lemma 2.2 for $v = \mathrm{depth}(\mathbf{v})$ (so that $\mathbf{v}_v = \mathbf{v}$) therefore gives the following: conditionally on $\mathcal{F}_{\mathbf{v}}$, the variables $\widetilde{T}_{\mathbf{v}}, \widetilde{J}_{\mathbf{v}}, \widetilde{U}_{\mathbf{v}}$ are independent, $\widetilde{T}_{\mathbf{v}} - \widetilde{\tau}_{\mathbf{v}} \sim \mathsf{Exp}(|\widetilde{C}_{\mathbf{v}}^j|)$, $\mathbb{P}(\widetilde{J}_{\mathbf{v}} = j | \mathcal{F}_{\mathbf{v}}) = |\widetilde{C}_{\mathbf{v}}^j| / \big(\sum_{j'=1}^d |\widetilde{C}_{\mathbf{v}}^j|\big)$ and $\widetilde{U}_{\mathbf{v}} \sim \mathcal{U}([0,1])$, so that $(\widetilde{S}_{\mathbf{v}} | \mathcal{F}_{\mathbf{v}}, \widetilde{T}_{\mathbf{v}}, \widetilde{J}_{\mathbf{v}}) \sim \mathcal{U}(\widetilde{C}_{\mathbf{v}}^{\widetilde{J}_{\mathbf{v}}})$. Hence, we have proven that, for every $\mathbf{v}$, the conditional distribution of $(T_{\mathbf{v}}, J_{\mathbf{v}}, S_{\mathbf{v}})$ given $\mathcal{F}_{\mathbf{v}}$ is the same as that of $(\widetilde{T}_{\mathbf{v}}, \widetilde{J}_{\mathbf{v}}, \widetilde{S}_{\mathbf{v}})$ given $\widetilde{\mathcal{F}}_{\mathbf{v}}$. By induction on $\mathbf{v}$, since $\mathcal{F}_\epsilon = \widetilde{\mathcal{F}}_\epsilon$ 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 (2.52) and (2.53) and combining it with (2.51) completes the proof of Proposition 2.2. □

Proof of Lemma 2.2. We show by induction on $v \in \mathbf{N}$ the following property: conditionally on $\mathcal{F}_{\mathbf{v}}$, $(\widetilde{T}_v^j, \widetilde{U}_v^j)_{1 \leqslant j \leqslant d}$ 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 $\mathcal{F}_0$ 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 $\mathcal{F}_{v+1}$, *i.e.* on $\mathcal{F}_v, \widetilde{T}_v, \widetilde{J}_v, \widetilde{U}_v$, 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 $\mathcal{F}_v$, 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 $\widetilde{U}^j_{v+1} = \widetilde{U}^j_v$ are independent $\mathcal{U}([0,1])$ random variables (conditionally on $\mathcal{F}_v$), conditioned on the independent variables $\widetilde{T}_v, \widetilde{J}_v, \widetilde{U}_v$, so they remain independent $\mathcal{U}([0,1])$ random variables.
- $(\widetilde{T}^{\widetilde{J}_v}_{v+1} - \widetilde{T}_v, \widetilde{U}^{\widetilde{J}_v}_{v+1}) = (E^{\widetilde{J}_v}_{\widetilde{V}^{\widetilde{J}_v}_{v+1}} / \widetilde{L}^{\widetilde{J}_v}_{v+1}, U^{\widetilde{J}_v}_{\widetilde{V}^{\widetilde{J}_v}_{v+1}})$ is distributed, conditionally on $\mathcal{F}_{v+1}$, *i.e.* on $\widetilde{J}_v, \widetilde{T}_v, \widetilde{V}^{\widetilde{J}_v}_{v+1}, \widetilde{L}^{\widetilde{J}_v}_{v+1}$, as $\mathsf{Exp}(\widetilde{L}^{\widetilde{J}_v}_{v+1}) \otimes \mathcal{U}([0,1])$, and independent of $(\widetilde{T}^j_{v+1}, \widetilde{U}^j_{v+1})_{j \neq \widetilde{J}_v}$.

This completes the proof by induction.

Let $v \in \mathbf{N}$. We have established that, conditionally on $\mathcal{F}_v$, the variables $(\widetilde{T}^j_v, \widetilde{U}^j_v)_{1 \leqslant j \leqslant d}$ are independent, with $\widetilde{T}^j_v - \widetilde{T}_{v-1} \sim \mathsf{Exp}(\widetilde{L}^j_v)$ and $\widetilde{U}^j_v \sim \mathcal{U}([0,1])$. In particular, conditionally on $\mathcal{F}_v$, $\widetilde{U}_v$ 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 $\widetilde{T}_v$, $\widetilde{T}_v \sim \mathsf{Exp}(\sum_{j=1}^d \widetilde{L}^j_v)$ and $\mathbb{P}(\widetilde{J}_v = j | \mathcal{F}_v) = \widetilde{L}^j_v / (\sum_{j'=1}^d \widetilde{L}^{j'}_v)$. This concludes the proof of Lemma 2.2. □

2.8.3 Proof of Theorem 2.1

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

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

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

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

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 Györfi et al. (2002), which states that a Mondrian tree estimate is consistent if, as $n \to \infty$,

(*i*) $D_\lambda(X) \to 0$ in probability, and

(*ii*) $K_\lambda / n \to 0$ in probability,

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 Györfi et al. (2002) contains deterministic convergence, but can be relaxed to convergences in probability by a close inspection of the proof. Hence, in order to conclude the proof, it suffices to establish (*i*) and (*ii*). The first condition follows from the fact that, by Corollary 2.1,

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

as well as the assumption that $\lambda_n \to \infty$. Condition (ii) follows from Proposition 2.2 and the assumption $\lambda_n^d/n \to 0$. This concludes the proof. □

2.8.4 Proof of Proposition 2.3

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 $\widetilde{f}_\lambda(x) = \mathbb{E}[\bar{f}_\lambda^{(1)}(x)]$ (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 (2.20) for Mondrian trees:

$$R(\widehat{f}_{\lambda,n}^{(1)}) = \mathbb{E}\big[(f(X) - \bar{f}_\lambda(X))^2\big] + \mathbb{E}\big[(\bar{f}_\lambda(X) - \widehat{f}_{\lambda,n}^{(1)}(X))^2\big]\,. \tag{2.54}$$

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

$$\mathbb{E}\left[(\bar{f}_\lambda(x) - f(x))^2\right] = \mathrm{Var}(\bar{f}_\lambda(x)) + (\widetilde{f}_\lambda(x) - f(x))^2\,. \tag{2.55}$$

By Proposition 2.1, 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

$$\begin{aligned}
\mathbb{E}[E \wedge a] &= \int_0^a u e^{-u}\mathrm{d}u + a\mathbb{P}(E \geqslant a) = 1 - e^{-a} \\
\mathbb{E}[(E \wedge a)^2] &= \int_0^a u^2 e^{-u}\mathrm{d}u + a^2\mathbb{P}(E \geqslant a) = 2\left(1 - (a+1)e^{-a}\right),
\end{aligned} \tag{2.56}$$

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:

$$\begin{aligned}\mathrm{Var}(R_\lambda(x)) &= \lambda^{-2}\big(1 - 2\lambda(1-x)e^{-\lambda(1-x)} - e^{-2\lambda(1-x)}\big)\\ \mathrm{Var}(L_\lambda(x)) &= \lambda^{-2}\big(1 - 2\lambda x e^{-\lambda x} - e^{-2\lambda x}\big)\,,\end{aligned}$$

and thus

$$\mathrm{Var}(\bar{f}_\lambda(x)) = \frac{1}{4\lambda^2}\big(2 - 2\lambda x e^{-\lambda x} - 2\lambda(1-x)e^{-\lambda(1-x)} - e^{-2\lambda x} - e^{-2\lambda(1-x)}\big)\,. \tag{2.57}$$

In addition, the formula (2.56) yields

$$\begin{aligned}\mathbb{E}[R_\lambda(x)] &= x + \lambda^{-1}\big(1 - e^{-\lambda(1-x)}\big)\\ \mathbb{E}[L_\lambda(x)] &= x - \lambda^{-1}\big(1 - e^{-\lambda x}\big)\,,\end{aligned}$$

and thus

$$\widetilde{f}_\lambda(x) = 1 + \frac{\mathbb{E}[L_\lambda(x)] + \mathbb{E}[R_\lambda(x)]}{2} = 1 + x + \frac{1}{2\lambda}\big(e^{-\lambda x} - e^{-\lambda(1-x)}\big)\,. \tag{2.58}$$

Combining (2.57) and (2.58) with the decomposition (2.55) gives

$$\mathbb{E}\big[\big(\bar{f}_\lambda(x) - f(x)\big)^2\big] = \frac{1}{2\lambda^2}\Big(1 - \lambda x e^{-\lambda x} - \lambda(1-x)e^{-\lambda(1-x)} - e^{-\lambda}\Big)\,. \tag{2.59}$$

Integrating over X, we obtain

$$\begin{aligned}\mathbb{E}\left[(\bar{f}_\lambda(X) - f(X))^2\right] &= \frac{1}{2\lambda^2}\left(1 - \int_0^1 \lambda x e^{-\lambda x}\mathrm{d}x - \int_0^1 \lambda(1-x)e^{-\lambda(1-x)}\mathrm{d}x - e^{-\lambda}\right)\\ &= \frac{1}{2\lambda^2}\left(1 - 2\times\frac{1}{\lambda}\big(1 - (\lambda+1)e^{-\lambda}\big) - e^{-\lambda}\right)\\ &= \frac{1}{2\lambda^2}\left(1 - \frac{2}{\lambda} + e^{-\lambda} + \frac{2}{\lambda}e^{-\lambda}\right)\,.\end{aligned} \tag{2.60}$$

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 (2.60) 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 (2.60) shows that for $\lambda \geqslant 6$, we have

$$\mathbb{E}\left[(\bar{f}_\lambda(X) - f(X))^2\right] \geqslant \frac{1}{2\lambda^2}\left(1 - \frac{2}{6}\right) = \frac{1}{3\lambda^2}\,. \tag{2.61}$$

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 (Arlot and Genuer, 2014, 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):

$$\mathbb{E}\left[\left(\widehat{f}_{\lambda,n}(X) - \widetilde{f}_\lambda(X)\right)^2\right] \geqslant \frac{\sigma^2}{n}\left(\mathbb{E}\left[K_\lambda\right] - 2\mathbb{E}_{\Pi_\lambda}\left[\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \exp(-n\mathbb{P}(X \in C_\mathbf{v}))\right]\right). \tag{2.62}$$

Now, recall that the splits defining Π_λ form a Poisson point process on $[0,1]$ of intensity $\lambda \mathrm{d}x$ (Fact 2.1). 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 $(\lambda^{-1}S_1, \dots, \lambda^{-1}S_{K_\lambda - 1})$, 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

$$\begin{aligned}
\mathbb{E}\Bigg[\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \exp(-n\mathbb{P}\left(X \in C_\mathbf{v}\right))\Bigg] &= \mathbb{E}\Bigg[\sum_{k=1}^{K_\lambda - 1} e^{-n\lambda^{-1}(S_k - S_{k-1})} + e^{-n(1-\lambda^{-1}S_{K_\lambda - 1})}\Bigg] \\
&\leqslant \mathbb{E}\Bigg[\sum_{k=1}^{\infty} \mathbf{1}(S_k \leqslant \lambda) e^{-n\lambda^{-1}E_k}\Bigg] + 1 = \sum_{k=1}^{\infty} \mathbb{E}\big[\mathbf{1}(S_k \leqslant \lambda)\big]\mathbb{E}\big[e^{-n\lambda^{-1}E_k}\big] + 1 \qquad (2.63) \\
&= \sum_{k=1}^{\infty} \mathbb{E}\big[\mathbf{1}(S_k \leqslant \lambda)\big] \cdot \int_0^\infty e^{-n\lambda^{-1}u} e^{-u}\mathrm{d}u + 1 = \frac{\lambda}{n+\lambda}\mathbb{E}\Bigg[\sum_{k=1}^{\infty} \mathbf{1}(S_k \leqslant \lambda)\Bigg] + 1 \\
&= \frac{\lambda}{n+\lambda}\mathbb{E}\left[K_\lambda\right] + 1 = \frac{\lambda}{n+\lambda}(1+\lambda) + 1 \qquad (2.64)
\end{aligned}$$

where (2.63) comes from the fact that E_k and S_{k-1} are independent. Plugging Equation (2.64) in the lower bound (2.62) yields

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

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 $6 \leqslant \lambda \leqslant \frac{n}{3}$,

$$\mathbb{E}\left[\left(\widehat{f}_{\lambda,n}(X) - \widetilde{f}_\lambda(X)\right)^2\right] \geqslant \frac{\sigma^2\lambda}{4n}. \tag{2.65}$$

Second lower bound on the variance. The lower bound (2.65) 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 $N_{\lambda,n}(x) = N_n(C_\lambda(x))$:

$$\begin{aligned}
&\mathbb{E}\big[(\widehat{f}_{\lambda,n}(X) - \bar{f}_\lambda(X))^2\big] \geqslant \mathbb{P}\big(N_{\lambda,n}(X) = 0\big) \\
&= \mathbb{E}\bigg[\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \mathbb{P}(X \in C_\mathbf{v})\,\mathbb{P}(N_n(C_\mathbf{v}) = 0)\bigg] = \mathbb{E}\bigg[\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \mathbb{P}(X \in C_\mathbf{v})\big(1 - \mathbb{P}(X \in C_\mathbf{v})\big)^n\bigg] \\
&\geqslant \mathbb{E}\bigg[\Big(\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \mathbb{P}(X \in C_\mathbf{v})\big(1 - \mathbb{P}(X \in C_\mathbf{v})\big)\Big)^n\bigg] \qquad (2.66) \\
&\geqslant \mathbb{E}\bigg[\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \mathbb{P}(X \in C_\mathbf{v})\big(1 - \mathbb{P}(X \in C_\mathbf{v})\big)\bigg]^n \qquad (2.67) \\
&= \bigg(1 - \mathbb{E}\bigg[\sum_{\mathbf{v}\in\mathcal{L}(\Pi_\lambda)} \mathbb{P}(X \in C_\mathbf{v})^2\bigg]\bigg)^n \qquad (2.68)
\end{aligned}$$

where (2.66) and (2.67) come from Jensen's inequality applied to the convex function $x \mapsto x^n$. Now, using the notations defined above, we have

$$\begin{aligned}
&\mathbb{E}\bigg[\sum_{\mathbf{v}\in\Pi_\lambda} \mathbb{P}(X \in C_\mathbf{v})^2\bigg] \leqslant \mathbb{E}\bigg[\sum_{k=1}^{K_\lambda} (\lambda^{-1}E_k)^2\bigg] \\
&= \lambda^{-2}\,\mathbb{E}\bigg[\sum_{k=1}^{\infty} \mathbf{1}(S_{k-1} \leqslant \lambda) E_k^2\bigg] = \lambda^{-2}\,\mathbb{E}\bigg[\sum_{k=1}^{\infty} \mathbf{1}(S_{k-1} \leqslant \lambda)\mathbb{E}\big[E_k^2 | S_{k-1}\big]\bigg] \\
&= 2\lambda^{-2}\,\mathbb{E}\bigg[\sum_{k=1}^{\infty} \mathbf{1}(S_{k-1} \leqslant \lambda)\bigg] \qquad (2.69) \\
&= 2\lambda^{-2}\,\mathbb{E}\left[K_\lambda\right] = \frac{2(\lambda+1)}{\lambda^2}\,, \qquad (2.70)
\end{aligned}$$

where the equality $\mathbb{E}[E_k^2|S_{k-1}] = 2$ (used in Equation (2.69)) comes from the fact that $E_k \sim \mathsf{Exp}(1)$ is independent of S_{k-1}.

The bounds (2.68) and (2.70) imply that, if $2(\lambda+1)/\lambda^2 \leqslant 1$, then

$$\mathbb{E}\big[(\widehat{f}_{\lambda,n}(X) - \bar{f}_\lambda(X))^2\big] \geqslant \left(1 - \frac{2(\lambda+1)}{\lambda^2}\right)^n. \qquad (2.71)$$

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 (2.71) gives, letting $C_2 := 1/8^8$,

$$\mathbb{E}\big[(\widehat{f}_{\lambda,n}(X) - \bar{f}_\lambda(X))^2\big] \geqslant C_2\,. \qquad (2.72)$$

Summing up. Assume that $n \geqslant 18$. Recall the bias-variance decomposition (2.54) 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 (2.72) implies that the variance (and hence the risk) is larger than C_2;
- If $6 \leqslant \lambda \leqslant \frac{n}{3}$, Equations (2.61) (bias term) and (2.65) (variance term) imply that

$$R(\widehat{f}_{\lambda,n}) \geqslant \frac{1}{3\lambda^2} + \frac{\sigma^2 \lambda}{4n}\,.$$

In particular, letting $C_0 = C_1 \wedge C_2$, we conclude that

$$\inf_{\lambda \in \mathbf{R}^+} R(\widehat{f}_{\lambda,n}) \geqslant C_1 \wedge C_2 \wedge \inf_{\lambda \in \mathbf{R}^+} \left(\frac{1}{3\lambda^2} + \frac{\sigma^2 \lambda}{4n} \right) = C_0 \wedge \frac{1}{4} \left(\frac{3\sigma^2}{n} \right)^{2/3}. \tag{2.73}$$

2.8.5 Proof of Proposition 2.4

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 $c_1 = p_0(1-2\varepsilon)^d/4$,

$$\mathbb{P}\big(N_\varepsilon \leqslant c_1 n\big) \leqslant \exp(-c_1 n/4)\,. \tag{2.74}$$

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 Audibert (2008) that the estimator $\widehat{g}_n$ defined by (2.12) satisfies, with probability $1-\delta$ over the random sample $\mathscr{D}'$ conditionally on N_ε,

$$\begin{aligned} \mathbb{E}_{(X,Y)}\big[(\widehat{g}_n(X) - Y)^2 | X \in B_\varepsilon\big] - \min_{0 \leqslant \alpha \leqslant A} \mathbb{E}_{(X,Y)}\big[(\widehat{f}_\alpha(X) - Y)^2 | X \in B_\varepsilon\big] \\ \leqslant \frac{CB^2 \log[(A+1)\delta^{-1}]}{N_\varepsilon} \end{aligned} \tag{2.75}$$

for every $\delta \in (0,1)$, where $C = 600$ and the expectation is taken with respect to an independent sample (X,Y) (the bound (2.75) 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 (2.75) 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

$$\begin{aligned}&\mathbb{E}\big[(\widehat{g}_n(X)-f(X))^2|X\in B_\varepsilon,N_\varepsilon>c_1n\big]\\&\leqslant\min_{0\leqslant\alpha\leqslant A}\mathbb{E}\big[(\widehat{f}_\alpha(X)-f(X))^2|X\in B_\varepsilon\big]+\frac{CB^2[\log(1+\log_2 n)+1]}{c_1n}.\end{aligned}\tag{2.76}$$

Finally, combining the bounds (2.74) and (2.76) yields

$$\begin{aligned}&\mathbb{E}\big[(\widehat{g}_n(X)-f(X))^2|X\in B_\varepsilon\big]\\&\leqslant\mathbb{P}\left(N_\varepsilon\leqslant c_1n\right)\cdot4B^2+\mathbb{E}\big[(\widehat{g}_n(X)-f(X))^2|X\in B_\varepsilon,N_\varepsilon>c_1n\big]\\&\leqslant4B^2e^{-c_1n/4}+\min_{0\leqslant\alpha\leqslant A}\mathbb{E}\big[(\widehat{f}_\alpha(X)-f(X))^2|X\in B_\varepsilon\big]+\frac{CB^2[\log(1+\log_2 n)+1]}{c_1n},\end{aligned}\tag{2.77}$$

which is precisely inequality (2.13).

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 2.2 (with the same adaptation as in the proof of Theorem 2.3 to bound the variance term conditionally on $X \in B_\varepsilon$) that, for every $\lambda > 0$,

$$\mathbb{E}\big[(\widehat{f}_{\lambda,n_0,M}(X)-f(X))^2|X\in B_\varepsilon\big]\leqslant\frac{(4d)^sL^2}{\lambda^{2s}}+\frac{11B^2(1+\lambda)^d}{p_0(1-2\varepsilon)^dn_0}$$

(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,

$$\begin{aligned}\min_{0\leqslant\alpha\leqslant A}\mathbb{E}\big[(\widehat{f}_\alpha(X)-f(X))^2|X\in B_\varepsilon\big]&\leqslant\min_{0\leqslant\alpha\leqslant A}\left[\frac{C_1L^2}{(2^\alpha)^{2s}}+\frac{C_2(1+2^\alpha)^d}{n}\right]\\&\leqslant4\min_{\lambda\in[1,n^{1/d}]}\left[\frac{C_1L^2}{\lambda^{2s}}+\frac{C_2(1+\lambda)^d}{n}\right],\end{aligned}\tag{2.78}$$

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 (2.77) and (2.78) that

$$\begin{aligned}\mathbb{E}\big[(\widehat{g}_n(X)-f(X))^2|X\in B_\varepsilon\big]&=O\Big(\min_{0\leqslant\lambda\leqslant n^{1/d}}\left[\frac{C_1L^2}{\lambda^{2s}}+\frac{C_2(1+\lambda)^d}{n}\right]+\frac{\log\log n}{n}\Big)\\&=O\left(L^{2d/(d+2s)}n^{-2s/(d+2s)}\right)\end{aligned}$$

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 2.3 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 $e^{-\lambda\varepsilon}/\lambda^3 = O(1/\lambda^{2s})$)

$$\mathbb{E}\big[(\widehat{f}_{\lambda,n,M}(X) - f(X))^2 | X \in B_\varepsilon\big] \leqslant \frac{C_3 L^2}{\lambda^{2s}} + \frac{C_4(1+\lambda)^d}{n}\,. \tag{2.79}$$

From the same argument as in the case $0 < s \leqslant 1$, combining inequalities (2.79) and (2.77) 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 2.4. □

2.8.6 Proof of Lemma 2.1

According to Equation (2.31) from the main text, we have

$$F_\lambda(x,z) = \lambda^d \exp(-\lambda\|x-z\|_1) \prod_{1\leqslant j\leqslant d} G_\lambda(x_j, z_j) \tag{2.80}$$

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

$$A_j = \Phi_\lambda^1(x_j) \qquad \text{and} \qquad B = \sum_{j=1}^d \Phi_\lambda^2(x_j) \tag{2.81}$$

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

$$\Phi_\lambda^k(u) = \int_0^1 \lambda\exp(-\lambda|u-v|)G_\lambda(u,v)\frac{(v-u)^k}{k!}\mathrm{d}v\,. \tag{2.82}$$

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

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

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

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

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

$$\left| \lambda \Phi_\lambda^1(u) - \underbrace{\int_0^\beta v e^{-v} \left[H(v, \lambda u, \lambda(1-u) - v) - H(v, \lambda u - v, \lambda(1-u)) \right] \mathrm{d}v}_{:= I_3} \right| \leqslant e^{-\beta} \,. \tag{2.83}$$

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

$$H(a, b_1, b_2) - H(a) \leqslant a^{-1}(e^{-b_1} + e^{-b_2}), \tag{2.84}$$

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

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

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

$$\begin{aligned}
|I_3| &\leqslant \int_0^\beta v e^{-v} \left|H(v, \lambda u, \lambda(1-u) - v) - H(v, \lambda u - v, \lambda(1-u))\right| \mathrm{d}v \\
&\leqslant \int_0^\beta v e^{-v} 2v^{-1} e^{-\beta} \mathrm{d}v \\
&\leqslant 2e^{-\beta} \int_0^\infty e^{-v} \mathrm{d}v \\
&= 2e^{-\beta}
\end{aligned} \tag{2.85}$$

Combining Equations (2.83) and (2.85) yields:

$$|\Phi_\lambda^1(u)| \leqslant \frac{3}{\lambda} e^{-\lambda[u \wedge (1-u)]/2} \tag{2.86}$$

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 2.1. □