The Annals of Statistics
2020, Vol. 48, No. 4, 2253–2276
https://doi.org/10.1214/19-AOS1886

MINIMAX OPTIMAL RATES FOR MONDRIAN TREES AND FORESTS

BY JAOUAD MOURTADA[1,*], STÉPHANE GAÏFFAS[2] AND ERWAN SCORNET[1,**]

[1]*CMAP, École polytechnique,* [*]*jaouad.mourtada@polytechnique.edu;* [**]*erwan.scornet@polytechnique.edu*
[2]*LPMA—Université Paris Diderot, stephane.gaiffas@lpsm.paris*

Introduced by Breiman (*Mach. Learn.* **45** (2001) 5–32), 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* (In *Adv. Neural Inf. Process. Syst.* (2014) 3140–3148; In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)* (2016)), whose trees are built using the so-called Mondrian process, therefore allowing to easily update their construction in a streaming fashion. In this paper 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.

1. Introduction. Introduced by Breiman [8], *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 [25], survival analysis [21] and ranking [11]; improvements of original RF are provided in literature, to cite but a few, better sampling strategies [19], new splitting methods [27] or Bayesian alternatives [10]. 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 [7]). 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 [3, 12, 26, 35, 38, 39] 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) [2, 5, 6, 9, 18, 22] grow the individual trees independently of the sample, and are thus particularly amenable to theoretical analysis. The consistency of such

Received March 2018; revised July 2019.
MSC2020 subject classifications. Primary 62G05; secondary 62G08, 62C20, 62H30.
Key words and phrases. Random forests, minimax rates, nonparametric estimation, supervised learning.

algorithms (as well as other idealized RF procedures) was first obtained by [6], 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 [2, 18]. In particular, [18] 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 in [2], 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 multidimensional setting, where trees exhibit an intricate recursive structure, only suboptimal rates are derived. As shown by lower bounds from [22], 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 [8] and the Extra-Trees algorithm [19], 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 [13, 14, 16, 34, 37].

Of particular interest in this article is the *Mondrian forest* (MF) algorithm, an efficient and accurate online random forest classifier introduced by [23]; see also [24]. This algorithm is based on the Mondrian process [31–33], 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 [23] 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 [23] 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 paper 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 [2]. 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 paper is organized as follows. In Section 2 we describe the considered setting and set the notation for trees and forests. Section 3 defines the Mondrian process introduced by [33] and describes the MF algorithm. Section 4 provides new sharp properties for Mondrian partitions, cells distribution in Proposition 1 and a control of the cells diameter in Corollary 1 while the expected number of cells is provided in Proposition 2. Building on these properties, we provide, in Section 5, statistical guarantees for MF. Theorem 1 proves consistency, while Theorems 2 and 3 provide minimax rates for $s \in (0, 1]$ and $s \in (1, 2]$, respectively. Finally, Proposition 4 proves that a combination of MF with a model aggregation algorithm adapts to the unknown smoothness $s \in (0, 2]$.

2. Setting and notation. We first describe the setting of the paper 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 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. $[0,1]^d \times \mathbb{R}$-valued random variables. We assume throughout the paper 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 $f(X) = \mathbb{E}[Y \mid X]$ 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* $\widehat{f}_n(\cdot, Z, \mathscr{D}_n) : [0,1]^d \to \mathbb{R}$ 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

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

where the expectation is taken with respect to $(X, Z, \mathscr{D}_n)$. We say that a sequence $(\widehat{f}_n)_{n\geq 1}$ 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 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

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

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.

Algorithm 1 `SampleMondrian`(C, τ, λ): samples a Mondrian partition of C, starting from time τ and until time λ.

Inputs: A cell $C = \prod_{1 \le j \le d} [a_j, b_j]$, starting time τ and lifetime parameter λ.
Sample a random variable $E_C \sim \mathsf{Exp}(|C|)$
if $\tau + E_C \le \lambda$ **then**
    Sample a split dimension $J \in \{1, \dots, d\}$, 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 $C_0 = \{x \in C : x_J \le S_J\}$ and $C_1 = C \setminus C_0$
    return `SampleMondrian`$(C_0, \tau + E_C, \lambda) \cup$ `SampleMondrian`$(C_1, \tau + E_C, \lambda)$
else
    return $\{C\}$ (i.e., do not split C).
end if

3. The Mondrian forest algorithm. Given a rectangular box $C = \prod_{j=1}^{d} [a_j, b_j] \subseteq \mathbb{R}^d$, 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 [33]; see also [32] 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 `SampleMondrian`$(C, \tau = 0, \lambda)$ described in Algorithm 1. Figure 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 $\lambda > 0$.

REMARK 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 $\lambda \le \lambda'$.

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 $\Pi_\lambda^{(1)}, \dots, \Pi_\lambda^{(M)}$, fits each one with the dataset $\mathscr{D}_n$ by averaging the labels falling into each

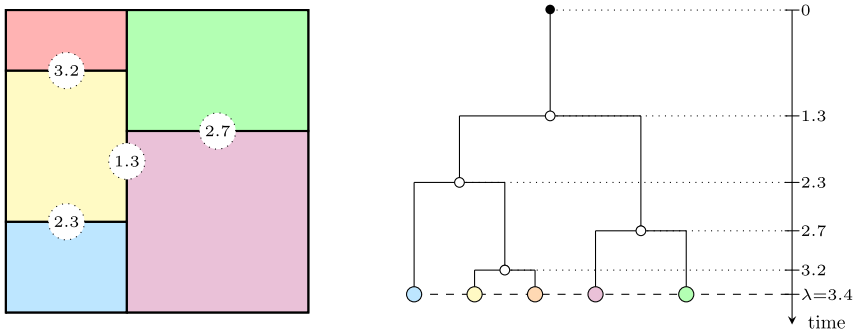

FIG. 1. *A Mondrian partition* (*left*) *with corresponding tree structure* (*right*) *which shows the evolution of the tree over time. The split times are indicated on the vertical axis while the splits are denoted with bullets* (∘).

leaf, then combines the resulting Mondrian tree estimates by averaging their predictions. In accordance with equation (2.3), we let

$$\widehat{f}_{\lambda,n,M}(x, \Pi_{\lambda,M}) = \frac{1}{M}\sum_{m=1}^{M} \widehat{f}^{(m)}_{\lambda,n}(x, \Pi^{(m)}_{\lambda}) \tag{3.1}$$

be the Mondrian forest estimate described above 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 $\widehat{f}^{(m)}_{\lambda,n}(x, \Pi^{(m)}_{\lambda})$. 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 [23] 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 paper we simply compute the average of the observations falling into a given cell.

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 [33]:

FACT 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 (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, that is, the partition on C induced by the partition Π_λ of $[0, 1]^d$. Then, $\Pi_\lambda|_C \sim \mathsf{MP}(\lambda, C)$.

Fact 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 is fundamental, and enables one to precisely characterize the behavior of the Mondrian partitions.

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

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

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*;
(ii) *For each $j = 1, \ldots, d$, $L_{j,\lambda}(x)$ 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 1 is given in Section 7. Figure 2 is a graphical representation of Proposition 1. A consequence of Proposition 1 is the next Corollary 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 5 below.

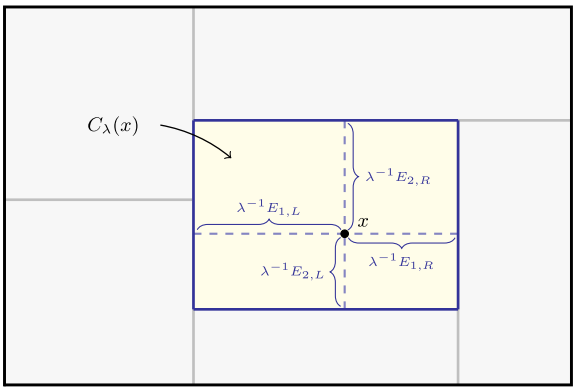

FIG. 2. *Cell distribution in a Mondrian partition* (*Proposition* 1).

COROLLARY 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*

$$\mathbb{P}\big(D_\lambda(x) \geq \delta\big) \leq d\left(1 + \frac{\lambda\delta}{\sqrt{d}}\right)\exp\left(-\frac{\lambda\delta}{\sqrt{d}}\right) \tag{4.1}$$

and

$$\mathbb{E}\big[D_\lambda(x)^2\big] \leq \frac{4d}{\lambda^2}. \tag{4.2}$$

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.

PROPOSITION 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 is given in the Supplementary Material [29] while a sketch of proof is provided in Section 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 1). A much simpler result is $\mathbb{E}[K_\lambda] \leq (e(1+\lambda))^d$ which was previously obtained in [28]. By contrast, Proposition 2 provides the *exact* value of this expectation which removes a superfluous e^d factor.

REMARK 2. Proposition 2 naturally extends (with the same proof) to the more general case of a Mondrian process with finite measures with no atoms $\nu_1, \ldots, \nu_d$ on the sides $C^1, \ldots, C^d$ of a box $C \subseteq \mathbb{R}^d$ (for a definition of the Mondrian process in this more general case, see [32]). In this case we have $\mathbb{E}[K_\lambda] = \prod_{1 \leq j \leq d}(1 + \nu_j(C^j))$.

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 so that it contains $O(\lambda^d)$ cells of diameter $O(1/\lambda)$ which is the minimal number of cells to cover $[0,1]^d$.

5. Minimax theory for Mondrian forests. This section gathers several theoretical guarantees for Mondrian trees and forests. Section 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 5.2 and with $s \in (1, 2]$ in Section 5.3, provided in both cases that λ_n is properly tuned. Note that in particular, we illustrate in Section 5.3 the fact that Mondrian forests improve over Mondrian trees when $s \in (1, 2]$. In Section 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 5.5.

5.1. *Consistency of Mondrian forests.* The consistency of the Mondrian forest estimator is established in Theorem 1 below, assuming a proper tuning of the lifetime parameter λ_n.

THEOREM 1 (Universal consistency). *Let $M \geq 1$. Consider Mondrian trees $\widehat{f}^{(m)}_{\lambda_n,n}$ (for $m = 1, \ldots, M$) and Mondrian forest $\widehat{f}_{\lambda_n,n,M}$ given by equation (3.1) for a sequence $(\lambda_n)_{n\geq 1}$ satisfying $\lambda_n \to \infty$ and $\lambda_n^d/n \to 0$. Then, under the setting described in Section 2 above, the individual trees $\widehat{f}^{(m)}_{\lambda_n,n}$ (for $m = 1, \ldots, M$) are consistent, and as a consequence, the forest $\widehat{f}_{\lambda_n,n,M}$ is consistent for any $M \geq 1$.*

The proof of Theorem 1 is given in the Supplementary Material [29]. It uses the properties of Mondrian partitions established in Section 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).

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 [7]. However, Theorem 1 leaves open the question of a theoretically optimal tuning of λ_n under additional assumptions on the regression function f which we address next.

5.2. *Mondrian trees and forests are minimax for s-Hölder functions with $s \in (0, 1]$.* The bounds obtained in Corollary 1 and Proposition 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 1. Consider (X, Y) from the setting described in Section 2, and assume also that $\mathbb{E}[\varepsilon \mid X] = 0$ and $\mathrm{Var}(\varepsilon \mid X) \leq \sigma^2 < \infty$ 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 1. Let $p \in \mathbb{N}$, $\beta \in (0, 1]$ 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 $f : [0, 1]^d \to \mathbb{R}$ such that

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

for every $x, x' \in [0, 1]^d$ and $k \in \{1, \ldots, p\}$. 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 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, [36], Chapter I.3 in [30] or Theorem 3.2 in [20]).

THEOREM 2. *Grant Assumption 1, and assume that $f \in \mathscr{C}^{0,\beta}(L)$ where $\beta \in (0, 1]$ and $L > 0$. Let $M \geq 1$. The quadratic risk of the Mondrian forest $\widehat{f}_{\lambda,n,M}$ with lifetime parameter $\lambda > 0$ satisfies*

$$\mathbb{E}\big[\big(\widehat{f}_{\lambda,n,M}(X) - f(X)\big)^2\big] \leq \frac{(4d)^\beta L^2}{\lambda^{2\beta}} + \frac{(1+\lambda)^d}{n}\big(2\sigma^2 + 9\|f\|_\infty^2\big). \tag{5.1}$$

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

$$\mathbb{E}\big[\big(\widehat{f}_{\lambda_n,n,M}(X) - f(X)\big)^2\big] = O\big(L^{2d/(d+2\beta)} n^{-2\beta/(d+2\beta)}\big) \tag{5.2}$$

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

The proof of Theorem 2 is given in Section 7. It relies on the properties about Mondrian partitions stated in Section 4. Namely, Corollary 1 allows to control the bias of Mondrian trees (first term on the right-hand side of equation (5.1)), while Proposition 2 helps in controlling the variance of Mondrian trees (second term on the right-hand side of equation (5.1)).

To the best of our knowledge, Theorem 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 [18] and [2] 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 Purely Uniformly Random Forests (PURF) 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 suboptimal results, for example, [5] exhibits 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 [22] 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 [2], who studied the Balanced Purely Random Forests (BPRF) 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 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 is valid for Mondrian forests with any number of trees and, thus, in particular for a Mondrian tree (this is also true for Theorem 1). However, it is a well-known fact that forests outperform single trees in practice [17]. Section 5.3 proposes an explanation for this phenomenon by assuming $f \in \mathscr{C}^{1,\beta}(L)$.

5.3. *Improved rates for Mondrian forests compared to a Mondrian tree.* The convergence rate stated in Theorem 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 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 suboptimal for such functions.

PROPOSITION 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*

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

for any $n \geq 18$.

The proof of Proposition 3 is given in the Supplementary Material [29]. Since the minimax rate over $\mathscr{C}^{1,1}$ in dimension 1 is $O(n^{-4/5})$, Proposition 3 proves that a single Mondrian tree is not minimax optimal over this set of functions. However, it turns out that large enough Mondrian forests, which average Mondrian trees, are minimax optimal over $\mathscr{C}^{1,1}$. Therefore, Theorem 3 below highlights the benefits of a forest compared to a single tree.

THEOREM 3. *Grant Assumption 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 (3.1). Set $\varepsilon \in (0, 1/2)$ and $B_\varepsilon = [\varepsilon, 1-\varepsilon]^d$. Then, we have*

$$\begin{aligned}
&\mathbb{E}\big[\big(\widehat{f}_{\lambda,n,M}(X) - f(X)\big)^2 | X \in B_\varepsilon\big] \\
&\quad \leq \frac{2(1+\lambda)^d}{n}\frac{2\sigma^2 + 9\|f\|_\infty^2}{p_0(1-2\varepsilon)^d} \\
&\quad\quad + \frac{144L^2 d p_1}{p_0(1-2\varepsilon)^d}\frac{e^{-\lambda\varepsilon}}{\lambda^3} + \frac{72L^2 d^3}{\lambda^4}\left(\frac{p_1 C_p}{p_0^2}\right)^2 \\
&\quad\quad + \frac{16L^2 d^{1+\beta}}{\lambda^{2(1+\beta)}}\left(\frac{p_1}{p_0}\right)^2 + \frac{8dL^2}{M\lambda^2},
\end{aligned} \tag{5.3}$$

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

$$\mathbb{E}\big[\big(\widehat{f}_{\lambda_n,n,M_n}(X) - f(X)\big)^2 \mid X \in B_\varepsilon\big] = O\big(L^{2d/(d+2s)} n^{-2s/(d+2s)}\big) \tag{5.4}$$

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 (5.4) holds if $2s \leq 3$. 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

$$\mathbb{E}\big[\big(\widehat{f}_{\lambda_n,n,M_n}(X) - f(X)\big)^2\big] = O\big(L^{2d/(d+3)} n^{-3/(d+3)}\big). \tag{5.5}$$

The proof of Theorem 3 is given in Section 7 below. It relies on an improved control of the bias, compared to the one used in Theorem 2 in the Lipschitz case. It exploits the knowledge of the distribution of the cell $C_\lambda(x)$ given in Proposition 1 instead of merely the cell diameter given in Corollary 1 (which was enough for Theorem 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 suboptimally. Such an effect was already noticed by [2] for purely random forests.

REMARK 3. While equation (5.4) 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 [2, 40]. It is however possible to set $\varepsilon = 0$ in (5.3) which leads to the sub-optimal rate stated in (5.5).

5.4. *Adaptation to the smoothness.* The minimax rates of Theorems 2 and 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 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 [30]. More specifically, we split the dataset into two parts. 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 in [4]. The appeals of this aggregation procedure are its simplicity, its optimal guarantee and the lack of parameter to tune.

Let $n_0 = \lfloor n/2 \rfloor$, $\mathscr{D}_{n_0} = \{(X_i, Y_i) : 1 \le i \le n_0\}$ and $\mathscr{D}_{n_0+1:n} = \{(X_i, Y_i) : n_0 + 1 \le i \le n\}$. 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 (3.1). The estimators $\widehat{f}_\alpha$ are computed using the sample $\mathscr{D}_{n_0}$ and the Mondrian partitions $\Pi^{(m)}_{2^\alpha}$, $1 \le m \le M$. Let

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

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

[4]. Then, the adaptive estimator is a convex combination of two Mondrian forests estimates with different lifetime parameters, given by

$$\widehat{g}_n = \operatorname*{argmin}_{g \in \widehat{\mathcal{G}}} \left\{ \frac{1}{|I_\varepsilon|} \sum_{i \in I_\varepsilon} \big(g(X_i) - Y_i\big)^2 \right\}. \tag{5.6}$$

PROPOSITION 4. *Grant Assumption 1, with $|Y| \le B$ 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 $p_0 \le p \le p_1$. Then, the estimator $\widehat{g}_n$ defined by (5.6) satisfies*

$$\begin{aligned} &\mathbb{E}\big[\big(\widehat{g}_n(X) - f(X)\big)^2 \mid X \in B_\varepsilon\big] \\ &\quad \le \min_{\alpha=0,\dots,A} \mathbb{E}\big[\big(\widehat{f}_\alpha(X) - f(X)\big)^2 \mid X \in B_\varepsilon\big] \\ &\qquad + 4B^2 e^{-c_1 n/4} + \frac{600B^2(\log(1+\log_2 n)+1)}{c_1 n}, \end{aligned} \tag{5.7}$$

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

$$\mathbb{E}\big[\big(\widehat{g}_n(X) - f(X)\big)^2 \mid X \in B_\varepsilon\big] = O\big(L^{2d/(d+2s)} n^{-2s/(d+2s)}\big), \tag{5.8}$$

where $s = p + \beta$.

The proof of Proposition 4 is to be found in the Supplementary Material [29]. Proposition 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 $s \in (0,2]$.

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

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

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

$$L(\widehat{g}_{\lambda,n,M}) = \mathbb{P}\big(\widehat{g}_{\lambda,n,M}(X) \ne Y\big), \tag{5.9}$$

where the probability is taken with respect to $(X, Y, \Pi_{\lambda,M}, \mathscr{D}_n)$ and where $\Pi_{\lambda,M}$ is the set sampled Mondrian partitions; see (3.1). Note that (5.9) is larger than the Bayes risk defined as

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

where $g^*(x) = \mathbf{1}(\eta(x) \ge 1/2)$. A general theorem [15], 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.

COROLLARY 2. *Let $M \ge 1$ 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*

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

The rate of convergence $o(n^{-1/(d+2)})$ for the error probability with a Lipschitz conditional probability η is optimal [41]. We can also extend in the same way Theorem 3 to the context of classification. This is done in the next corollary where we only consider the $\mathscr{C}^{1,1}$ case for convenience.

COROLLARY 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*

$$\mathbb{P}\big[\widehat{g}_n(X) \neq Y | X \in B_\varepsilon\big] - \mathbb{P}\big[g^*(X) \neq Y | X \in B_\varepsilon\big] = o\big(n^{-2/(d+4)}\big) \tag{5.10}$$

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.

6. Conclusion. Despite their widespread use in practice, the theoretical understanding of Random forests is still incomplete. In this work, 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 [2, 6] 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 paper 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. A direction for future work is to improve Mondrian forests using a data-driven choice of the features along which the splits are performed, reminiscent of Extra-Trees [19]. From a theoretical perspective it would be interesting to see how the minimax rates obtained here can be combined with results on the ability of forests to select informative variables (see, for instance, [35]).

7. Proofs. This Section gathers the proofs of Proposition 1 and Corollary 1 (cell distribution and cell diameter). Then, a sketch of the proof of Proposition 2 is described in this section (the full proof, which involves some technicalities, can be found in the Supplementary Material [29]). Finally, we provide the proofs of Theorem 2 and Theorem 3.

PROOF OF PROPOSITION 1. Let $0 \leq a_1, \ldots, a_n, b_1, \ldots, b_n \leq 1$ be such that $a_j \leq x_j \leq b_j$ for $1 \leq j \leq d$. Let $C := \prod_{j=1}^d [a_j, b_j]$. Note that the event

$$E_\lambda(C, x) = \big\{L_{1,\lambda}(x) \leq a_1, R_{1,\lambda}(x) \geq b_1, \ldots, L_{d,\lambda}(x) \leq a_d, R_{d,\lambda}(x) \geq b_d\big\}$$

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, that is, 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),

$\Pi_\lambda|_C$ 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

$$\mathbb{P}\big(E_\lambda(C,x)\big) = e^{-\lambda(x-a_1)}e^{-\lambda(b_1-x)} \times \cdots \times e^{-\lambda(x-a_d)}e^{-\lambda(b_d-x)}. \tag{7.1}$$

In particular, setting $a_j = b_j = x$ in (7.1), except for one a_j or b_j, and using that $L_{j,\lambda}(x) \le x$ and $R_{j,\lambda}(x) \ge x$, we obtain

$$\mathbb{P}\big(R_{j,\lambda}(x) \ge b_j\big) = e^{-\lambda(b_j-x)} \quad \text{and} \quad \mathbb{P}\big(L_{j,\lambda}(x) \le a_j\big) = e^{-\lambda(x-a_j)}. \tag{7.2}$$

Since clearly $R_{j,\lambda}(x) \le 1$ and $L_{j,\lambda}(x) \ge 0$, equation (7.2) implies (ii). Additionally, plugging Equation (7.2) back into equation (7.1) shows that $L_{1,\lambda}(x)$, $R_{1,\lambda}(x)$, $\ldots$, $L_{d,\lambda}(x)$, $R_{d,\lambda}(x)$ are independent, that is, point (i). This completes the proof. □

PROOF OF COROLLARY 1. Using Proposition 1, for $1 \le j \le d$, $D_{j,\lambda}(x) = R_{j,\lambda}(x) - x_j + x_j - L_{j,\lambda}(x)$ 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

$$\mathbb{P}\big(D_{j,\lambda}(x) \ge \delta\big) \le (1+\lambda\delta)e^{-\lambda\delta} \tag{7.3}$$

for every $\delta > 0$ (with equality if $\delta \le x_j \wedge (1-x_j)$) and $\mathbb{E}[D_{j,\lambda}(x)^2] \le \lambda^{-2}(\mathbb{E}[E_1^2]+\mathbb{E}[E_2^2]) = 4/\lambda^2$. The bound (4.1) for the diameter $D_\lambda(x) = [\sum_{j=1}^d D_{j,\lambda}(x)^2]^{1/2}$ is obtained by noting that

$$\mathbb{P}\big(D_\lambda(x) \ge \delta\big) \le \mathbb{P}\bigg(\exists j : D_{j,\lambda}(x) \ge \frac{\delta}{\sqrt{d}}\bigg) \le \sum_{j=1}^d \mathbb{P}\bigg(D_{j,\lambda}(x) \ge \frac{\delta}{\sqrt{d}}\bigg),$$

while (4.2) 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. Let us provide here an outline of the argument; a fully detailed proof is available in the Supplementary Material [29]. 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 $j = 1, \ldots, d$. 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 $J = \operatorname{argmin}_{1\le j\le d} T_j$. If $\tau + T_J > \lambda$, then C is not split (and is thus a cell of Π_λ). On the other hand, if $\tau + T_J \le \lambda$, then C is split along coordinate J at S_J (and at time $\tau + T_J$) into $C' = \{x \in C : x_J \le S_J\}$ 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 $j \ne J$ 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 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 [29]. □

PROOF OF THEOREM 2. Recall that the Mondrian forest estimate at x is given by

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

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

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

since the random trees estimators $\widehat{f}_{\lambda,n}^{(m)}$ have the same distribution for $m = 1, \dots, M$. Hence, it suffices to prove Theorem 2 for the tree estimator $\widehat{f}_{\lambda,n}^{(1)}$. We will denote for short $\widehat{f}_\lambda := \widehat{f}_{\lambda,n}^{(1)}$ 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 [18]. 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 $(\Pi_\lambda, \mathscr{D}_n)$:

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

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

$$R(\widehat{f}_\lambda) = \mathbb{E}\big[\big(f(X) - \bar{f}_\lambda(X)\big)^2\big] + \mathbb{E}\big[\big(\bar{f}_\lambda(X) - \widehat{f}_\lambda(X)\big)^2\big]. \tag{7.4}$$

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 $\mathscr{D}_n$, Π_λ).

Note that (7.4) 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 $\mathbb{E}_{\mathscr{D}_n}[\widehat{f}_\lambda(x)] \neq \bar{f}_\lambda(x)$. 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 [2, 18] which should only hold for the integrated risk.

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

$$\begin{aligned}\big|f(x) - \bar{f}_\lambda(x)\big| &= \left|\frac{1}{\mu(C_\lambda(x))}\int_{C_\lambda(x)} \big(f(x) - f(z)\big)\mu(\mathrm{d}z)\right| \\ &\le \sup_{z \in C_\lambda(x)} \big|f(x) - f(z)\big| \le L D_\lambda(x)^\beta,\end{aligned}$$

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 1, this implies

$$\mathbb{E}\big[\big(f(x) - \bar{f}_\lambda(x)\big)^2\big] \le L^2 \mathbb{E}\big[D_\lambda(x)^{2\beta}\big] \le L^2 \mathbb{E}\big[D_\lambda(x)^2\big]^\beta \le L^2 \left(\frac{4d}{\lambda^2}\right)^\beta. \tag{7.5}$$

Integrating (7.5) with respect to μ yields the following bound on the bias:

$$\mathbb{E}\big[\big(f(X) - \bar{f}_\lambda(X)\big)^2\big] \le \frac{(4d)^\beta L^2}{\lambda^{2\beta}}. \tag{7.6}$$

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

$$\mathbb{E}\big[\big(\bar{f}_\Pi(X) - \widehat{f}_\Pi(X)\big)^2\big] \le \frac{k}{n}\big(2\sigma^2 + 9\|f\|_\infty^2\big). \tag{7.7}$$

Note that Proposition 2 in [2], 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 [1]. For every $k \in \mathbb{N}^*$, applying (7.7) to the random partition $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0, 1]^d)$ conditionally on the event $\{K_\lambda = k\}$, we get

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

Using Proposition 2, we obtain an upper bound of the variance term:

$$\mathbb{E}\big[\big(\bar{f}_\lambda(X) - \widehat{f}_\lambda(X)\big)^2\big] \le \frac{(1+\lambda)^d}{n}\big(2\sigma^2 + 9\|f\|_\infty^2\big). \tag{7.8}$$

Combining (7.6) and (7.8) leads to (5.1). Finally, the bound (5.2) follows by using $\lambda = \lambda_n$ in (5.1) which concludes the proof of Theorem 2. □

PROOF OF THEOREM 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 $m = 1, \dots, M$ are based on independent partitions $\Pi_\lambda^{(m)} \sim \mathsf{MP}(\lambda, [0, 1]^d)$. Also, for x in the support of μ let

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

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

$$\begin{aligned}
\mathbb{E}\big[\big(\widehat{f}_{\lambda,M}(x) - f(x)\big)^2\big] &\le 2\mathbb{E}\big[\big(\widehat{f}_{\lambda,M}(x) - \bar{f}_{\lambda,M}(x)\big)^2\big] \\
&\quad + 2\mathbb{E}\big[\big(\bar{f}_{\lambda,M}(x) - f(x)\big)^2\big].
\end{aligned}$$

In addition, Jensen's inequality implies that

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

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

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

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

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

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

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

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

$$\begin{aligned}&\mathbb{E}\big[\big(\widehat{f}_{\lambda,M}(X) - f(X)\big)^2 | X \in B_{\varepsilon}\big] \\ &\quad\le \frac{8dL^2}{M\lambda^2} + 2\mathbb{E}\big[\big(\widehat{f}_{\lambda}^{(1)}(X) - \bar{f}_{\lambda}^{(1)}(X)\big)^2 | X \in B_{\varepsilon}\big] \\ &\qquad + 2\mathbb{E}\big[\big(\widetilde{f}_{\lambda}(X) - f(X)\big)^2 | X \in B_{\varepsilon}\big].\end{aligned} \tag{7.9}$$

Variance term. Recall that the distribution μ of X has a positive density $p : [0,1]^d \to \mathbb{R}_+^*$, 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, the variance term satisfies

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

Hence, the conditional variance in the decomposition (7.9) satisfies

$$\begin{aligned}&\mathbb{E}\big[\big(\bar{f}_{\lambda}^{(1)}(X) - \widehat{f}_{\lambda}^{(1)}(X)\big)^2 | X \in B_{\varepsilon}\big] \\ &\quad\le \mathbb{P}(X \in B_{\varepsilon})^{-1}\mathbb{E}\big[\big(\bar{f}_{\lambda}^{(1)}(X) - \widehat{f}_{\lambda}^{(1)}(X)\big)^2\big] \\ &\quad\le p_0^{-1}(1-2\varepsilon)^{-d}\frac{(1+\lambda)^d}{n}\big(2\sigma^2 + 9\|f\|_{\infty}^2\big).\end{aligned} \tag{7.10}$$

Expression of $\widetilde{f}_{\lambda}$. It remains to control the bias term in the decomposition (7.9) 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

$$\begin{aligned}\widetilde{f}_{\lambda}(x) &= \mathbb{E}\bigg[\frac{1}{\mu(C_{\lambda}(x))}\int_{[0,1]^d} f(z)p(z)\mathbf{1}\big(z \in C_{\lambda}(x)\big)\,\mathrm{d}z\bigg] \\ &= \int_{[0,1]^d} f(z)F_{p,\lambda}(x,z)\,\mathrm{d}z,\end{aligned} \tag{7.11}$$

where we defined

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

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

$$\begin{aligned}
&\big|f(z) - f(x) - \nabla f(x)^\top (z-x)\big| \\
&\quad = \Big|\int_0^1 \big[\nabla f\big(x + t(z-x)\big) - \nabla f(x)\big]^\top (z-x)\,\mathrm{d}t\Big| \\
&\quad \le \int_0^1 L\big(t\|z-x\|\big)^\beta \|z-x\|\,\mathrm{d}t \le L\|z-x\|^{1+\beta}.
\end{aligned}$$

Now, by the triangle inequality,

$$\begin{aligned}
&\Big|\Big|\int_{[0,1]^d} \big(f(z) - f(x)\big)F_{p,\lambda}(x,z)\,\mathrm{d}z\Big| - \Big|\int_{[0,1]^d} \nabla f(x)^\top (z-x) F_{p,\lambda}(x,z)\,\mathrm{d}z\Big|\Big| \\
&\quad \le \Big|\int_{[0,1]^d} \big(f(z) - f(x) - \nabla f(x)^\top (z-x)\big)F_{p,\lambda}(x,z)\,\mathrm{d}z\Big| \\
&\quad \le L\int_{[0,1]^d} \|z-x\|^{1+\beta} F_{p,\lambda}(x,z)\,\mathrm{d}z,
\end{aligned}$$

so that, using together $\int F_{p,\lambda}(x,z)\,\mathrm{d}z = 1$ and (7.11), we obtain

$$\begin{aligned}
|\widetilde{f}_\lambda(x) - f(x)| \le{}& \Big|\nabla f(x)^\top \underbrace{\int_{[0,1]^d} (z-x) F_{p,\lambda}(x,z)\,\mathrm{d}z}_{:=A}\Big| \\
&+ L\underbrace{\int_{[0,1]^d} \|z-x\|^{1+\beta} F_{p,\lambda}(x,z)\,\mathrm{d}z}_{:=B}.
\end{aligned} \tag{7.12}$$

Hence, it remains to control the two terms A, B from equation (7.12). We will start by expressing $F_{p,\lambda}$ in terms of p, using the distribution of the cell $C_\lambda(x)$ given by Proposition 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 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 $C = C(x,z) = \prod_{1\le j\le d}[x_j \wedge z_j, x_j \vee z_j] \subseteq [0,1]^d$ be the smallest box containing both x and z; also, let $a_j = x_j \wedge z_j$, $b_j = x_j \vee z_j$, $a = (a_j)_{1\le j\le d}$ and $b = (b_j)_{1\le j\le d}$. 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 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 $E_{j,L}$, $E_{j,R}$, $1 \le j \le d$ are i.i.d. $\mathsf{Exp}(1)$ random variables. Note that $b \in C_\lambda(a)$ is equivalent to $R_{\lambda,j}(a) \ge b_j$ for $j = 1,\dots,d$, *that*

is, to $E_{j,R} \geq \lambda(b_j - a_j)$. By the memory-less property of the exponential distribution, the distribution of $E_{j,R} - \lambda(b_j - a_j)$ conditionally on $E_{j,R} \geq \lambda(b_j - a_j)$ 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 $L_{\lambda,j}(a)$, $R_{\lambda,j}(a)$, $1 \leq j \leq d$, 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

$$C_\lambda(x,z) := \prod_{j=1}^{d} \big[(x_j \wedge z_j - \lambda^{-1}E_{j,L}) \vee 0, (x_j \vee z_j + \lambda^{-1}E_{j,R}) \wedge 1\big], \tag{7.13}$$

where $E_{1,L}, E_{1,R}, \ldots, E_{d,L}, E_{d,R}$ 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, this occurs with probability $\exp(-\lambda|C(x,z)|) = \exp(-\lambda\|x-z\|_1)$. Therefore,

$$\begin{aligned} F_{p,\lambda}(x,z) &= \mathbb{P}\big(z \in C_\lambda(x)\big)\mathbb{E}\bigg[\frac{p(z)}{\mu(C_\lambda(x))}\Big| z \in C_\lambda(x)\bigg] \\ &= e^{-\lambda\|x-z\|_1}\mathbb{E}\bigg[\bigg\{\int_{C_\lambda(x,z)} \frac{p(y)}{p(z)}\,\mathrm{d}y\bigg\}^{-1}\bigg], \end{aligned} \tag{7.14}$$

where $C_\lambda(x,z)$ is as in (7.13). In addition, applying (7.14) to $p \equiv 1$ yields

$$\begin{aligned} F_\lambda(x,z) = \lambda^d e^{-\lambda\|x-z\|_1} \prod_{1\leq j\leq d} \mathbb{E}\big[\{\lambda|x_j - z_j| + E_{j,L} \wedge \lambda(x_j \wedge z_j) \\ + E_{j,R} \wedge \lambda(1 - x_j \vee z_j)\}^{-1}\big]. \end{aligned} \tag{7.15}$$

The following technical lemma, whose proof is given in Supplementary Material [29], will prove useful in what follows.

LEMMA 1. *The function $F_{p,\lambda}$ given by* (7.15) *satisfies*

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

and

$$\int_{[0,1]^d} \frac{1}{2}\|z-x\|^2 F_\lambda(x,z)\,\mathrm{d}z \leq \frac{d}{\lambda^2}$$

for any $x \in [0,1]^d$.

Control of the term B in equation (7.12). It follows from (7.14) and from the bound $p(y)/p(z) \geq p_0/p_1$ that

$$F_{p,\lambda}(x,z) \leq \frac{p_1}{p_0}F_\lambda(x,z), \tag{7.16}$$

so that

$$\begin{aligned} \int_{[0,1]^d} \|z-x\|^{1+\beta} F_{p,\lambda}(x,z)\,\mathrm{d}z &\leq \frac{p_1}{p_0}\int_{[0,1]^d} \|z-x\|^{1+\beta} F_\lambda(x,z)\,\mathrm{d}z \\ &\leq \frac{p_1}{p_0}\bigg(\int_{[0,1]^d} \|z-x\|^2 F_\lambda(x,z)\,\mathrm{d}z\bigg)^{(1+\beta)/2} \qquad (7.17) \\ &\leq \frac{p_1}{p_0}\bigg(\frac{2d}{\lambda^2}\bigg)^{(1+\beta)/2}, \qquad (7.18) \end{aligned}$$

where (7.17) follows from the concavity of $x \mapsto x^{(1+\beta)/2}$ for $\beta \in (1, 2]$ while (7.18) comes from Lemma 1.

Control of the term A in equation (7.12). 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 1. However, this time the crude bound (7.16) 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

$$\left| \frac{p(y)}{p(z)} - 1 \right| = \frac{|p(y) - p(z)|}{p(z)} \le \frac{C_p}{p_0} \|y - z\| \le \frac{C_p}{p_0} \operatorname{diam} C_\lambda(x, z) \tag{7.19}$$

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

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

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

$$\begin{aligned} &\left\{ 1 + \frac{C_p}{p_0} \operatorname{diam} C_\lambda(x, z) \right\}^{-1} \operatorname{vol} C_\lambda(x, z)^{-1} \\ &\quad \le \left\{ \int_{C_\lambda(x,z)} \frac{p(y)}{p(z)} \mathrm{d}y \right\}^{-1} \\ &\quad \le \left\{ \left(1 - \frac{C_p}{p_0} \operatorname{diam} C_\lambda(x, z) \right) \vee \frac{p_0}{p_1} \right\}^{-1} \operatorname{vol} C_\lambda(x, z)^{-1}. \end{aligned} \tag{7.20}$$

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

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

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

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

Now, equation (7.20) implies that

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

Taking the expectation over $C_\lambda(x, z)$ and using (7.14) leads to

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

so that

$$|F_{p,\lambda}(x,z) - F_\lambda(x,z)| \le \frac{p_1 C_p}{p_0^2} e^{-\lambda\|x-z\|_1} \times \mathbb{E}\big[\operatorname{diam} C_\lambda(x,z) \operatorname{vol} C_\lambda(x,z)^{-1}\big]. \tag{7.21}$$

Control of $\mathbb{E}[\operatorname{diam} C_\lambda(x,z) \operatorname{vol} C_\lambda(x,z)^{-1}]$. 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 $\operatorname{diam} C_\lambda(x,z) \le \operatorname{diam}_{\ell^1} C_\lambda(x,z)$ using the triangular inequality, so that

$$\begin{aligned}
&\mathbb{E}\big[\operatorname{diam} C_\lambda(x,z) \operatorname{vol} C_\lambda(x,z)^{-1}\big] \\
&\quad \le \mathbb{E}\bigg[\sum_{j=1}^d |C_\lambda^j(x,z)| \operatorname{vol} C_\lambda(x,z)^{-1}\bigg] \\
&\quad = \sum_{j=1}^d \mathbb{E}\bigg[|C_\lambda^j(x,z)| \prod_{l=1}^d |C_\lambda^l(x,z)|^{-1}\bigg] \\
&\quad = \sum_{j=1}^d \mathbb{E}\bigg[\prod_{l\neq j} |C_\lambda^l(x,z)|^{-1}\bigg] \\
&\quad \le \sum_{j=1}^d \mathbb{E}\big[|C_\lambda^j(x,z)|\big]\mathbb{E}\big[|C_\lambda^j(x,z)|^{-1}\big]\mathbb{E}\bigg[\prod_{l\neq j} |C_\lambda^l(x,z)|^{-1}\bigg] \qquad (7.22) \\
&\quad = \sum_{j=1}^d \mathbb{E}\big[|C_\lambda^j(x,z)|\big] \times \mathbb{E}\bigg[\prod_{l=1}^d |C_\lambda^l(x,z)|^{-1}\bigg] \qquad (7.23) \\
&\quad = \mathbb{E}\big[\operatorname{diam}_{\ell^1} C_\lambda(x,z)\big] \times \exp\big(\lambda\|x-z\|_1\big) F_\lambda(x,z). \qquad (7.24)
\end{aligned}$$

Inequality (7.22) relies on the fact that $\mathbb{E}[X]\mathbb{E}[X^{-1}] \ge 1$ for any positive random variable X with $X = |C_\lambda^j(x,z)|$. Equality (7.23) comes from the independence of $|C_\lambda^1(x,z)|, \ldots, |C_\lambda^d(x,z)|$. Multiplying both sides of (7.24) by $e^{-\lambda\|x-z\|_1}$ leads to

$$\begin{aligned}
&e^{-\lambda\|x-z\|_1}\mathbb{E}\big[\operatorname{diam} C_\lambda(x,z) \operatorname{vol} C_\lambda(x,z)^{-1}\big] \\
&\quad \le \mathbb{E}\big[\operatorname{diam}_{\ell^1} C_\lambda(x,z)\big] F_\lambda(x,z).
\end{aligned} \tag{7.25}$$

In addition,

$$\begin{aligned}
\mathbb{E}\big[\operatorname{diam}_{\ell^1} C_\lambda(x,z)\big] &\le \sum_{j=1}^d \mathbb{E}\big[|x_j - z_j| + \lambda^{-1}(E_{j,R} + E_{j,L})\big] \\
&= \|x-z\|_1 + \frac{2d}{\lambda}.
\end{aligned} \tag{7.26}$$

Finally, combining equations (7.21), (7.25) and (7.26) gives

$$|F_{p,\lambda}(x,z) - F_\lambda(x,z)| \le \frac{p_1 C_p}{p_0^2}\bigg(\|x-z\|_1 + \frac{2d}{\lambda}\bigg) F_\lambda(x,z). \tag{7.27}$$

Control of A. From (7.27) 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

$$\begin{aligned}\left\|\int_{[0,1]^d}(z-x)F_{p,\lambda}(x,z)\,\mathrm{d}z-\int_{[0,1]^d}(z-x)F_{\lambda}(x,z)\,\mathrm{d}z\right\|\\ \le\int_{[0,1]^d}\|z-x\|\times\left|F_{p,\lambda}(x,z)-F_\lambda(x,z)\right|\mathrm{d}z,\end{aligned}\tag{7.28}$$

with

$$\begin{aligned}&\int_{[0,1]^d}\|z-x\|\times\left|F_{p,\lambda}(x,z)-F_\lambda(x,z)\right|\mathrm{d}z\\ &\quad\le\frac{p_1C_p}{p_0^2}\int_{[0,1]^d}\|z-x\|\left[\|x-z\|_1+\frac{2d}{\lambda}\right]F_\lambda(x,z)\,\mathrm{d}z\quad\text{(by (7.27))}\\ &\quad\le\frac{p_1C_p}{p_0^2}\left[\sqrt{d}\int_{[0,1]^d}\|z-x\|^2F_\lambda(x,z)\,\mathrm{d}z+\frac{2d}{\lambda}\int_{[0,1]^d}\|z-x\|F_\lambda(x,z)\,\mathrm{d}z\right]\\ &\quad\le\frac{p_1C_p}{p_0^2}\left[\frac{d\sqrt{d}}{\lambda^2}+\frac{2d}{\lambda}\left(\int_{[0,1]^d}\|z-x\|^2F_\lambda(x,z)\,\mathrm{d}z\right)^{1/2}\right],\end{aligned}$$

where we used the inequalities $\|v\|\le\|v\|_1\le\sqrt{d}\|v\|$ as well as the Cauchy–Schwarz inequality. Hence, using Lemma 1, we end up with

$$\begin{aligned}&\int_{[0,1]^d}\|z-x\|\times\left|F_{p,\lambda}(x,z)-F_\lambda(x,z)\right|\mathrm{d}z\\ &\quad\le\frac{p_1C_p}{p_0^2}\left[\frac{d\sqrt{d}}{\lambda^2}+\frac{2d}{\lambda}\sqrt{\frac{d}{\lambda^2}}\right]\\ &\quad=\frac{p_1C_p}{p_0^2}\frac{3d\sqrt{d}}{\lambda^2}.\end{aligned}\tag{7.29}$$

Inequalities (7.28) and (7.29) together with Lemma 1 entail that

$$\begin{aligned}&\left\|\int_{[0,1]^d}(z-x)F_{p,\lambda}(x,z)\,\mathrm{d}z\right\|^2\\ &\quad\le2\left\|\int_{[0,1]^d}(z-x)F_\lambda(x,z)\,\mathrm{d}z\right\|^2\\ &\qquad+2\left(\int_{[0,1]^d}\|z-x\|\left|F_{p,\lambda}(x,z)-F_\lambda(x,z)\right|\mathrm{d}z\right)^2\\ &\quad\le\frac{18}{\lambda^2}\sum_{j=1}^d e^{-\lambda[x_j\wedge(1-x_j)]}+2\left(\frac{p_1C_p}{p_0^2}\frac{3d\sqrt{d}}{\lambda^2}\right)^2.\end{aligned}\tag{7.30}$$

Control of the bias. The upper bound (7.12) on the bias writes

$$\left(\widetilde{f}_\lambda(x)-f(x)\right)^2\le\left(\left|\nabla f(x)^\top A\right|+LB\right)^2\le2\left(\left\|\nabla f(x)\right\|^2\times\|A\|^2+L^2B^2\right),$$

so that plugging the bounds (7.18) of B and (7.30) of $\|A\|$ gives

$$\left(\widetilde{f}_\lambda(x)-f(x)\right)^2$$

$$\leq 2L^2\left[\frac{18}{\lambda^2}\sum_{j=1}^{d} e^{-\lambda[x_j\wedge(1-x_j)]} + 2\left(\frac{p_1 C_p}{p_0^2}\frac{3d\sqrt{d}}{\lambda^2}\right)^2\right] + 2L^2\frac{p_1}{p_0}\left(\frac{2d}{\lambda^2}\right)^{(1+\beta)/2}$$

$$\leq \frac{36L^2}{\lambda^2}\sum_{j=1}^{d} e^{-\lambda[x_j\wedge(1-x_j)]} + \frac{36L^2d^3}{\lambda^4}\left(\frac{p_1 C_p}{p_0^2}\right)^2 + \frac{8L^2d^{1+\beta}}{\lambda^{2(1+\beta)}}\left(\frac{p_1}{p_0}\right)^2.$$

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

$$\mathbb{E}\big[(\widetilde{f}_\lambda(X) - f(X))^2 | X \in B_\varepsilon\big] \leq \frac{36L^2}{\lambda^2}\psi_\varepsilon(\lambda) + \frac{36L^2d^3}{\lambda^4}\left(\frac{p_1 C_p}{p_0^2}\right)^2 + \frac{8L^2d^{1+\beta}}{\lambda^{2(1+\beta)}}\left(\frac{p_1}{p_0}\right)^2, \tag{7.31}$$

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

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

Conclusion. The decomposition (7.9), together with the bounds (7.10) on the variance and (7.31) on the bias lead to inequality (5.3) from the statement of Theorem 3. In particular, if $\varepsilon \in (0, \frac{1}{2})$ is fixed, inequality (5.3) writes

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

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 3.

On the other hand, we have $e^{-\lambda\varepsilon} = 1$ whenever $\varepsilon = 0$, so that inequality (5.3) becomes in this case

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

When $2s \leq 3$ (i.e., $\beta \leq 1/2$), 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 3. □

Acknowledgment. Data Science Initiative of École polytechnique.

SUPPLEMENTARY MATERIAL

Supplement to "Minimax optimal rates for Mondrian trees and forests" (DOI: 10.1214/19-AOS1886SUPP; .pdf). Supplementary information.

REFERENCES

[1] ARLOT, S. (2008). V-fold cross-validation improved: V-fold penalization. ArXiv preprint. Available at arXiv:0802.0566.

[2] ARLOT, S. and GENUER, R. (2014). Analysis of purely random forests bias. ArXiv preprint. Available at arXiv:1407.3939.

[3] ATHEY, S., TIBSHIRANI, J. and WAGER, S. (2019). Generalized random forests. *Ann. Statist.* **47** 1148–1178. MR3909963 https://doi.org/10.1214/18-AOS1709

[4] AUDIBERT, J.-Y. (2008). Progressive mixture rules are deviation suboptimal. In *Adv. Neural Inf. Process. Syst* **20** 41–48.

[5] BIAU, G. (2012). Analysis of a random forests model. *J. Mach. Learn. Res.* **13** 1063–1095. MR2930634

[6] BIAU, G., DEVROYE, L. and LUGOSI, G. (2008). Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.* **9** 2015–2033. MR2447310

[7] BIAU, G. and SCORNET, E. (2016). A random forest guided tour. *TEST* **25** 197–227. MR3493512 https://doi.org/10.1007/s11749-016-0481-7

[8] BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

[9] BREIMAN, L. (2010). Some infinity theory for predictor ensembles. Technical Report 577, Statistics Department, Univ. California Berkeley.

[10] CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. MR2758172 https://doi.org/10.1214/09-AOAS285

[11] CLÉMENÇON, S., DEPECKER, M. and VAYATIS, N. (2013). Ranking forests. *J. Mach. Learn. Res.* **14** 39–73. MR3033325

[12] CUI, Y., ZHU, R., ZHOU, M. and KOSOROK, M. (2017). Some asymptotic results of survival tree and forest models. ArXiv preprint. Available at arXiv:1707.09631.

[13] DENIL, M., MATHESON, D. and DE FREITAS, N. (2013). Consistency of online random forests. In *Proceedings of the* 30*th Annual International Conference on Machine Learning* (*ICML*) 1256–1264.

[14] DENIL, M., MATHESON, D. and DE FREITAS, N. (2014). Narrowing the gap: Random forests in theory and in practice. In *Proceedings of the* 31*st Annual International Conference on Machine Learning* (*ICML*) 665–673.

[15] DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition. Applications of Mathematics* (*New York*) **31**. Springer, New York. MR1383093 https://doi.org/10.1007/978-1-4612-0711-5

[16] DOMINGOS, P. and HULTEN, G. (2000). Mining high-speed data streams. In *Proceedings of the Conference on Knowledge Discovery and Data Mining* (*KDD*) 71–80.

[17] FERNÁNDEZ-DELGADO, M., CERNADAS, E., BARRO, S. and AMORIM, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* **15** 3133–3181. MR3277155

[18] GENUER, R. (2012). Variance reduction in purely random forests. *J. Nonparametr. Stat.* **24** 543–562. MR2968888 https://doi.org/10.1080/10485252.2012.677843

[19] GEURTS, P., ERNST, D. and WEHENKEL, L. (2006). Extremely randomized trees. *Mach. Learn.* **63** 3–42.

[20] GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression. Springer Series in Statistics*. Springer, New York. MR1920390 https://doi.org/10.1007/b97848

[21] ISHWARAN, H., KOGALUR, U. B., BLACKSTONE, E. H. and LAUER, M. S. (2008). Random survival forests. *Ann. Appl. Stat.* **2** 841–860. MR2516796 https://doi.org/10.1214/08-AOAS169

[22] KLUSOWSKI, J. M. (2018). Complete analysis of a random forest model. ArXiv preprint. Available at arXiv:1805.02587.

[23] LAKSHMINARAYANAN, B., ROY, D. M. and TEH, Y. W. (2014). Mondrian forests: Efficient online random forests. In *Adv. Neural Inf. Process. Syst.* **27** 3140–3148.

[24] LAKSHMINARAYANAN, B., ROY, D. M. and TEH, Y. W. (2016). Mondrian forests for large-scale regression when uncertainty matters. In *Proceedings of the* 19*th International Conference on Artificial Intelligence and Statistics* (*AISTATS*).

[25] MEINSHAUSEN, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.* **7** 983–999. MR2274394

[26] MENTCH, L. and HOOKER, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.* **17** 26. MR3491120

[27] MENZE, B. H., KELM, B. M., SPLITTHOFF, D. N., KOETHE, U. and HAMPRECHT, F. A. (2011). On oblique random forests. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 453–469. Springer.

[28] MOURTADA, J., GAÏFFAS, S. and SCORNET, E. (2017). Universal consistency and minimax rates for online Mondrian forests. In *Adv. Neural Inf. Process. Syst* **30** 3759–3768.

[29] MOURTADA, J., GAÏFFAS, S. and SCORNET, E. (2020). Supplement to "Minimax optimal rates for Mondrian trees and forests." https://doi.org/10.1214/19-AOS1886SUPP.
[30] NEMIROVSKI, A. (2000). Topics in non-parametric statistics. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1998). *Lecture Notes in Math*. **1738** 85–277. Springer, Berlin. MR1775640
[31] ORBANZ, P. and ROY, D. M. (2015). Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 437–461.
[32] ROY, D. M. (2011). Computability, inference and modeling in probabilistic programming. Ph.D. thesis, Massachusetts Institute of Technology.
[33] ROY, D. M. and TEH, Y. W. (2009). The Mondrian process. In *Adv. Neural Inf. Process. Syst* **21** 1377–1384.
[34] SAFFARI, A., LEISTNER, C., SANTNER, J., GODEC, M. and BISCHOF, H. (2009). On-line random forests. In 3*rd IEEE ICCV Workshop on On-Line Computer Vision*.
[35] SCORNET, E., BIAU, G. and VERT, J.-P. (2015). Consistency of random forests. *Ann. Statist.* **43** 1716–1741. MR3357876 https://doi.org/10.1214/15-AOS1321
[36] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR0673642
[37] TADDY, M. A., GRAMACY, R. B. and POLSON, N. G. (2011). Dynamic trees for learning and design. *J. Amer. Statist. Assoc.* **106** 109–123. MR2816706 https://doi.org/10.1198/jasa.2011.ap09769
[38] WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* **113** 1228–1242. MR3862353 https://doi.org/10.1080/01621459.2017.1319839
[39] WAGER, S. and WALTHER, G. (2015). Adaptive concentration of regression trees, with application to random forests. ArXiv preprint. Available at arXiv:1503.06388.
[40] WASSERMAN, L. (2006). *All of Nonparametric Statistics. Springer Texts in Statistics*. Springer, New York. MR2172729
[41] YANG, Y. (1999). Minimax nonparametric classification. I. Rates of convergence. *IEEE Trans. Inform. Theory* **45** 2271–2284. MR1725115 https://doi.org/10.1109/18.796368