Inference with Mondrian Random Forests

Matias D. Cattaneo[1,*] Jason M. Klusowski[1] William G. Underwood[2]

April 7, 2025

Abstract

Random forests are popular methods for regression and classification analysis, and many different variants have been proposed in recent years. One interesting example is the Mondrian random forest, in which the underlying constituent trees are constructed via a Mondrian process. We give precise bias and variance characterizations, along with a Berry–Esseen-type central limit theorem, for the Mondrian random forest regression estimator. By combining these results with a carefully crafted debiasing approach and an accurate variance estimator, we present valid statistical inference methods for the unknown regression function. These methods come with explicitly characterized error bounds in terms of the sample size, tree complexity parameter, and number of trees in the forest, and include coverage error rates for feasible confidence interval estimators. Our novel debiasing procedure for the Mondrian random forest also allows it to achieve the minimax-optimal point estimation convergence rate in mean squared error for multivariate β-Hölder regression functions, for all $\beta > 0$, provided that the underlying tuning parameters are chosen appropriately. Efficient and implementable algorithms are devised for both batch and online learning settings, and we study the computational complexity of different Mondrian random forest implementations. Finally, simulations with synthetic data validate our theory and methodology, demonstrating their excellent finite-sample properties.

Keywords: Random forests, regression trees, Berry–Esseen theorem, bias correction, statistical inference, minimax estimation.

[1]Department of Operations Research and Financial Engineering, Princeton University
[2]Statistical Laboratory, University of Cambridge
[*]Corresponding author: wgu21@cam.ac.uk

Contents

1 Introduction

Random forests, first introduced by Breiman (2001), are a workhorse in modern machine learning for regression and classification tasks. Their desirable traits include computational efficiency in big data settings (via parallelization and greedy heuristics), simplicity of configuration and amenability to tuning parameter selection, ability to adapt to latent structure in high-dimensional data sets, and flexibility in handling mixed data types, among other virtues. Random forests have also achieved great empirical successes in many fields of study, including healthcare, finance, online commerce, causal inference, text analysis, bioinformatics, image classification, and ecology.

Since Breiman introduced random forests over twenty years ago, the study of their statistical properties remains an active area of research. Many fundamental questions about Breiman's random forests remain unanswered, owing in part to the subtle ingredients present in the estimation procedure which make standard analytical tools ineffective. These technical difficulties stem from the way the constituent trees greedily partition the covariate space, utilizing both the covariate and response data. This creates complicated dependencies on the data that are often exceedingly hard to untangle without overly stringent assumptions, thereby hampering theoretical progress.

To address the aforementioned technical challenges while retaining the phenomenology of Breiman's random forests, a variety of stylized versions of random forest procedures have been proposed and studied in the literature. Early proposals include centered random forests (Biau, 2012; Arnould et al., 2023) and median random forests (Duroux and Scornet, 2018; Arnould et al., 2023). Each tree in a centered random forest is constructed by first choosing a covariate uniformly at random and then splitting the cell at the midpoint along the direction of the chosen covariate. Median random forests operate in a similar way, but involve the covariate data by splitting at the empirical median along the direction of the randomly chosen covariate. Known as purely random forests, these procedures simplify Breiman's original, more data-adaptive version by growing trees that partition the covariate space in a way that is statistically independent of the response data.

Yet another variant of random forests, Mondrian random forests (Lakshminarayanan et al., 2014), have received significant attention from the statistics and machine learning communities in recent years (Ma et al., 2020; Mourtada et al., 2020, 2021; Scillitoe et al., 2021; Vicuna et al., 2021; Gao et al., 2022; O'Reilly and Tran, 2022; Baptista et al., 2024; O'Reilly and Tran, 2024; O'Reilly, 2024; Zhan et al., 2024; Osborne and O'Reilly, 2025). Like other purely random forest variants, Mondrian random forests offer a simplified modification of Breiman's original proposal in which the partition is generated independently of the data and according to a canonical stochastic process known as the Mondrian process (Roy et al., 2008). The Mondrian process takes a single tuning parameter $\lambda > 0$ known as the "lifetime" and enjoys various mathematical properties. These properties allow Mondrian random forests to be fitted in an online manner (Lakshminarayanan et al., 2014; Mourtada et al., 2021) and permit a rigorous statistical analysis, while also retaining some of the appealing features of other random forest methods. The lifetime parameter λ, in analogy with the number of refinements of a data-adaptive recursive partitioning algorithm, governs the extent to which the response data is smoothed, with a large λ resulting in a more complicated partition and therefore less smoothing. However, unlike a data-adaptive partition which modulates the smoothing intensity based on local data characteristics, the Mondrian process applies a uniform smoothing effect globally across all covariates. Thus, one can draw parallels between (axis-aligned) data-dependent partitioning schemes and more flexible—albeit, more involved—versions of the Mondrian process which employ adaptive directional smoothing, with unique lifetimes λ_j learned for each covariate, permitting tailored smoothing of the responses.

This paper studies the statistical properties of Mondrian random forests with an emphasis on *inference* techniques specific to this model. We focus on this purely random forest variant not only because of its importance in the development of random forest theory in general, but also because the Mondrian process (along with generalizations including the oblique Mondrian process and stationary random tessellation processes) is, to date, the only known randomized recursive tree mechanism for which the resulting random forest is minimax-optimal for point estimation over a class of smooth multivariate regression functions, without requiring sample splitting (Mourtada et al., 2020; O'Reilly and Tran, 2024; O'Reilly, 2024). In fact, when the covariate dimension exceeds one, the aforementioned centered and median random forests are both minimax *suboptimal*, due to their large biases, over the class of Lipschitz smooth regression functions (Klusowski, 2021). It is therefore natural to focus our study of inference for random forests on versions that at the very least exhibit competitive bias and variance, as this will have important implications for the trade-off between confidence and precision. Moreover, recent studies of generalized random forests and distributional random forests identify similar fundamental challenges, namely those of establishing Gaussian approximations and combining them with strategies for bias reduction (Näf et al., 2023).

Despite their recent popularity, relatively little is known about the formal statistical properties of Mondrian random forests. Focusing on nonparametric regression, Mourtada et al. (2020) recently showed that Mondrian forests containing just a single tree (called a Mondrian tree) can be minimax-optimal in integrated mean squared error whenever the regression function is β-Hölder continuous for some $\beta \in (0, 1]$. The authors also showed that, when appropriately tuned, large Mondrian random forests can be similarly minimax-optimal for $\beta \in (0, 2]$, while the constituent trees cannot. See also O'Reilly and Tran (2022) for analogous results on more general Mondrian tree and forest constructions. These results formally demonstrate the value of ensembling with random forests from a point estimation perspective. No results are currently available in the literature for statistical inference using Mondrian random forests.

As already mentioned, a different strand of the literature studies the statistical properties of Breiman's random forests which form ensembles of *adaptive* decision trees. In such models, each constituent tree is constructed with a greedy algorithm that recursively optimizes a goodness-of-fit metric (such as mean squared error) using both the covariates and response data; a leading example in practice is the celebrated Classification and Regression Tree (CART) methodology (Breiman et al., 1984; Breiman, 2001). The underlying complexity of the resulting procedures make their formal theoretical analysis quite difficult, and therefore only a more restricted set of results is currently available in the literature. In terms of estimation theory, Scornet et al. (2015) established consistency of adaptive random forests for additive models with a fixed number of covariates, and Chi et al. (2022), Klusowski and Tian (2024), and Cattaneo et al. (2024a) provided rates of convergence for models with a growing number of covariates, under different assumptions on the statistical and algorithmic features of the constituent decision trees. A framework for tuning tree depths via data-adaptive early stopping was developed by Miftachov and Reiß (2025). In contrast, formal inference theory is far less developed, since there are arguably no satisfactory theoretical results for fully adaptive decision tree or random forest methods. For example, Wager and Athey (2018) provide asymptotic estimation and inference results for adaptive random forests, but they employ sample splitting (i.e., the so-called "honesty" property where the partitioning and, separately, the output in the terminal cells are formed using independent subsamples), and make assumptions that rule out procedures commonly used in practice such as CART; cf., the so-called "α-regularity" condition (Cattaneo et al., 2024b). From a broad perspective, our paper connects with this distinct thread in the random forest literature by demonstrating optimal estimation and inference results for non-adaptive Mondrian random forests with explicit probability deviation guarantees, thereby also laying down the foundations for future developments of adaptive Mondrian random forest procedures.

1.1 Contributions

Our paper contributes to the literature on the foundational statistical properties of Mondrian random forest regression estimation with two main results. Firstly, we give a central limit theorem for the classical Mondrian random forest point estimator under weak conditions, and propose valid large-sample inference procedures employing a consistent standard error estimator. We establish these results by deploying a restricted moments version of the Berry–Esseen theorem for independent but not identically distributed (i.n.i.d.) random variables (Petrov, 1995, Theorem 5.7) because we need to handle delicate probabilistic features of the Mondrian random forest estimator. In particular, we deal with the existence of Mondrian cells which are "too small" and lead to a reduced effective (local) sample size for some trees in the forest. Such pathological cells are in fact typical in Mondrian random forests and complicate the probability limits of certain sample averages; in fact, small Mondrian random forests (or indeed single Mondrian trees) remain random even in the limit due to the lack of ensembling. The presence of such small cells renders inapplicable prior distributional approximation results for partitioning-based estimators in the literature (Huang, 2003; Cattaneo et al., 2020), since the commonly required quasi-uniformity assumption on the underlying partitioning scheme (cf., α-regularity in the adaptive random forest literature) is violated by partitions generated using the Mondrian process. We circumvent this technical challenge by establishing new theoretical results for Mondrian partitions and their associated Mondrian trees and forests, which may be of independent interest. Our distributional approximation does not rely on sample splitting; unlike approaches based on "honest" trees, the estimator is fit using the entire sample of covariates and responses simultaneously.

The second main contribution of our paper is to propose a debiasing approach for the Mondrian random forest point estimator. We accomplish this by first precisely characterizing the probability limit of the large sample conditional bias, and then applying a debiasing procedure based on the generalized jackknife (Schucany and Sommers, 1977). We thus exhibit a Mondrian random forest variant which is minimax-optimal in pointwise mean squared error when the regression function is β-Hölder for any $\beta > 0$. Our method works by generating an ensemble of Mondrian random forests carefully chosen to have smaller misspecification bias when extra smoothness is available, resulting in minimax optimality even for $\beta > 2$. This result complements Mourtada et al. (2020) by demonstrating the existence of a class of Mondrian random forests that can efficiently exploit the additional smoothness of the unknown regression function for minimax-optimal point estimation. Our proposed debiasing procedure is also useful when conducting statistical inference because it provides a principled method for ensuring that the bias is negligible relative to the standard deviation of the estimator. More specifically, we use our debiasing approach to construct valid confidence intervals based on robust bias correction (Calonico et al., 2018, 2022), and include an explicit bound on their coverage error probability.

For the purposes of implementation, we propose techniques for tuning parameter selection and demonstrate the practical applicability and accuracy of our methodology through empirical studies with simulated data. We also discuss applications to batch and online learning settings, presenting computationally efficient algorithms along with bounds for their average case time complexity.

1.2 Organization

Section 2 gives the assumptions on the data generating process, using a Hölder smoothness condition on the regression function to control the bias of various estimators. We also introduce the Mondrian process and use it to define the Mondrian random forest estimator, stating the assumptions on its lifetime parameter and the number of trees.

Section 3 presents our first set of main results. We begin by precisely characterizing the bias of the Mondrian random forest estimator in Lemma 1, with the aim of subsequently applying a debiasing procedure. We similarly analyze the variance of this estimator (Lemma 2), and deduce its rate of convergence in Theorem 1. Next, we present our Berry–Esseen-type central limit theorem for the centered Mondrian random forest estimator under weak conditions as Theorem 2, and discuss implications for lifetime parameter selection. To enable valid feasible statistical inference, we provide a consistent variance estimator in Lemma 3, and use it to construct confidence intervals in Theorem 3.

Section 4 introduces our proposed debiased Mondrian random forests, a family of estimators based on linear combinations of Mondrian random forests with varying lifetime parameters. These parameters are carefully chosen to annihilate leading terms in our bias characterization, yielding an estimator with superior bias properties (Lemma 4). We also study the variance of this debiased estimator (Lemma 5), and derive its rate of convergence in Theorem 4. The resulting rate is shown to be minimax-optimal in mean squared error for each Hölder parameter $\beta > 0$, under regularity conditions. Furthermore, Theorem 5 verifies that a Berry–Esseen theorem holds for the debiased Mondrian random forest. We again discuss the implications for the lifetime parameter, and provide a consistent variance estimator (Lemma 6) for constructing confidence intervals (Theorem 6).

Section 5 discusses implementation details and empirical results, beginning by presenting a data-driven approach to selecting the crucial lifetime parameter using polynomial estimation. We also give advice on choosing the number of trees, as well as other parameters associated with the debiasing procedure. Empirical simulation results are presented using synthetic data, demonstrating the practical value of our methods for optimal point estimation and feasible robust bias-corrected inference.

Section 6 considers the computational aspects of our methodology, presenting algorithmic procedures with precisely characterized average case time complexity bounds for the batch setting (Algorithm 1, Lemma 7) and for online learning regimes (Algorithm 2, Lemma 8).

Concluding remarks are given in Section 7, while the appendices contain all the mathematical proofs of our theoretical results (Appendix A), alongside additional empirical studies (Appendix B).

1.3 Notation

We write $\|\cdot\|_2$ for the usual Euclidean ℓ^2 norm on $\mathbb{R}^d$. The natural numbers are $\mathbb{N} = \{0, 1, 2, \ldots\}$. We use $a \wedge b$ for the minimum and $a \vee b$ for the maximum of two real numbers. For non-negative sequences a_n and b_n, write $a_n \lesssim b_n$ to indicate that a_n/b_n is bounded for $n \geq 1$. If $a_n \lesssim b_n \lesssim a_n$, write $a_n \asymp b_n$. For random non-negative sequences A_n and B_n, similarly write $A_n \lesssim_{\mathbb{P}} B_n$ if A_n/B_n is bounded in probability. Let $\Phi : \mathbb{R} \to \mathbb{R}$ be the cumulative distribution function of the standard normal distribution, and for $\alpha \in (0, 1)$, let q_α be the normal quantile satisfying $\Phi(q_\alpha) = \alpha$.

2 Setup

When using a Mondrian random forest, there are two sources of randomness. The first is the data, and here we consider the nonparametric regression setting with d-dimensional covariates. The second source is injected purposely from a collection of independent trees drawn from a Mondrian process using a specified lifetime parameter.

2.1 Data generation

We begin with a definition of Hölder continuity, which is used to determine a target class of regression functions, and which participates in controlling the bias of various estimators.

Definition 1 (Hölder continuity). *Take $\beta > 0$ and define $\underline{\beta} = \lceil \beta - 1 \rceil$ as the largest integer strictly less than β. We say a function $g : [0,1]^d \to \mathbb{R}$ is β-Hölder continuous and write $g \in \mathcal{H}^\beta$ if g is $\underline{\beta}$ times differentiable and $\max_{|\nu|=\underline{\beta}} |\partial^\nu g(x) - \partial^\nu g(x')| \leq C \|x - x'\|_2^{\beta - \underline{\beta}}$ for some constant $C > 0$ and all $x, x' \in [0,1]^d$. Here, $\nu \in \mathbb{N}^d$ is a multi-index with $|\nu| = \sum_{j=1}^d \nu_j$ and $\partial^\nu g(x) = \partial^{|\nu|} g(x) / \prod_{j=1}^d \partial x_j^{\nu_j}$.*

Throughout this paper, we assume that the data satisfies the following assumption.

Assumption 1 (Data generation). *Fix $d \geq 1$ and let (X_i, Y_i) be independent and identically distributed (i.i.d.) samples from a distribution on $\mathbb{R}^d \times \mathbb{R}$, writing $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$. Suppose X_i has Lebesgue density function $f(x)$ on $[0,1]^d$ which is bounded away from zero and satisfies $f \in \mathcal{H}^{\beta_f}$ for some $\beta_f > 0$. Suppose $\mathbb{E}[Y_i^2 \mid X_i]$ is bounded, let $\mu(X_i) = \mathbb{E}[Y_i \mid X_i]$, and assume $\mu \in \mathcal{H}^{\beta_\mu}$ where $\beta_\mu > 0$. Write $\varepsilon_i = Y_i - \mu(X_i)$ and assume $\sigma^2(X_i) = \mathbb{E}[\varepsilon_i^2 \mid X_i]$ is bounded away from zero with $\sigma^2 \in \mathcal{H}^{\beta_\sigma}$ for some $\beta_\sigma > 0$. Set $\beta = \beta_\mu \wedge (\beta_f + 1)$.*

Some comments are in order surrounding Assumption 1. The requirement that the covariate density $f(x)$ should be strictly positive on all of $[0,1]^d$ may seem restrictive, particularly when d is moderately large. However, since our theory is presented pointwise in the design point x, it is sufficient for this condition to hold only on some neighborhood of x. To see this, note that continuity implies the density is positive on some hypercube containing x. Upon rescaling the covariates, this hypercube can be mapped onto $[0,1]^d$. The same argument holds for the Hölder smoothness assumptions, and for the upper and lower bounds on the conditional variance function.

The definition of β is motivated as follows: firstly, we must have $\beta \leq \beta_\mu$ in order to compare our rates of convergence with classical results for β-Hölder regression functions. Secondly, due to the presence of design bias, we require in our analysis that the density function $f(x)$ should also be smooth (though not necessarily as smooth as μ), imposing $\beta \leq \beta_f + 1$. Our proofs characterize the roles of each of the smoothness parameters β_μ, β_f and β_σ precisely, and our main results depend only on β. By allowing for $\beta_f < 1$, we strictly generalize the Lipschitz density assumption of Mourtada et al. (2020, Theorem 3).

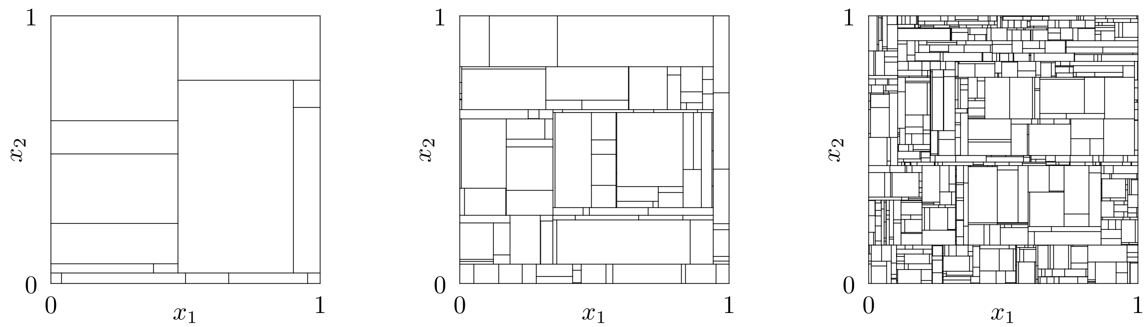
2.2 The Mondrian process

The Mondrian process was introduced by Roy et al. (2008) and offers a canonical method for generating random rectangular partitions, which can be used as the trees for a random forest (Lakshminarayanan et al., 2014). For the reader's convenience, we give a brief description of this process here; see Mourtada et al. (2020, Section 3) for a more complete construction.

For a fixed dimension d and lifetime parameter $\lambda > 0$, the Mondrian process is a stochastic process taking values in the set of finite rectangular partitions of $[0,1]^d$. For a rectangle $D = \prod_{j=1}^d [a_j, b_j] \subseteq [0,1]^d$, we denote the side aligned with dimension j by $D_j = [a_j, b_j]$, write $D_j^- = a_j$ and $D_j^+ = b_j$ for its left and right endpoints respectively, and use $|D_j| = D_j^+ - D_j^-$ for its length. The volume of D is $|D| = \prod_{j=1}^d |D_j|$ and its linear dimension (or half-perimeter) is $|D|_1 = \sum_{j=1}^d |D_j|$.

To sample a partition T from the Mondrian process $\mathcal{M}\big([0,1]^d, \lambda\big)$, start at time $t = 0$ with the trivial partition of $[0,1]^d$ which has no splits. Then repeatedly apply the following procedure to each cell D in the partition. Let t_D be the time at which the cell was formed, and sample $E_D \sim \mathrm{Exp}(|D|_1)$, where $\mathrm{Exp}(a)$ is the exponential distribution on $[0, \infty)$ with Lebesgue density

ae^{-ax}. If $t_D + E_D \leq \lambda$, then split D. This is done by first selecting a split dimension J with $\mathbb{P}(J = j) = |D_j|/|D|_1$, and then sampling a split location $S_J \sim \text{Unif}\left[D_J^-, D_J^+\right]$. The cell D splits into the two new cells $\{x \in D : x_J \leq S_J\}$ and $\{x \in D : x_J > S_J\}$, each with formation time $t_D + E_D$. The output of this sampling procedure is the partition T consisting of the cells D which were not split because $t_D + E_D > \lambda$. The cell in T containing a point $x \in [0,1]^d$ is written $T(x)$. Figure 1 shows typical realizations of $T \sim \mathcal{M}\big([0,1]^d, \lambda\big)$ for $d = 2$ and with different lifetime parameters λ.



(a) $\lambda = 3$ (b) $\lambda = 10$ (c) $\lambda = 30$

Figure 1: The Mondrian process $T \sim \mathcal{M}\big([0,1]^d, \lambda\big)$ with $d = 2$ and lifetime parameters λ.

2.3 Mondrian random forests

We define the Mondrian random forest estimator (1) as in Lakshminarayanan et al. (2014) and Mourtada et al. (2020), and will later extend it to a debiased version in Section 4. For a lifetime parameter $\lambda > 0$ and forest size $B \geq 1$, let $\mathbf{T} = (T_1, \ldots, T_B)$ be a Mondrian forest where $T_b \sim \mathcal{M}\big([0,1]^d, \lambda\big)$ are mutually independent Mondrian trees which are independent of the data. For $x \in [0,1]^d$, write $N_b(x) = \sum_{i=1}^n \mathbb{I}\{X_i \in T_b(x)\}$ for the number of samples in $T_b(x)$, with $\mathbb{I}$ denoting an indicator function. Then the Mondrian random forest estimator of $\mu(x)$ is

$$\hat{\mu}(x) = \frac{1}{B} \sum_{b=1}^{B} \frac{\sum_{i=1}^n Y_i \, \mathbb{I}\big\{X_i \in T_b(x)\big\}}{N_b(x)}. \tag{1}$$

If there are no samples X_i in $T_b(x)$ then $N_b(x) = 0$, so we define $0/0 = 0$ (see Appendix A for details). To ensure the bias and variance of the Mondrian random forest estimator converge to zero (see Section 3), and to avoid boundary issues, we impose some basic conditions on x, λ, and B in Assumption 2.

Assumption 2 (Mondrian random forest estimator)**.** *Suppose $x \in (0,1)^d$ is a fixed interior point of the support of X_i, and also that $\lambda \gtrsim (\log n)^3$, $n/\lambda^d \to \infty$, and $B \gtrsim (\log n)^d$.*

The requirement that $n/\lambda^d \to \infty$ ensures that the number of data points $N_b(x)$ falling inside a typical Mondrian cell $T_b(x)$, and hence the effective sample size of the Mondrian random forest, diverges in large samples. Assumption 2 also implies that the size of the forest B should grow with n. While our results place no upper bound on the number of trees, we suggest, for the purpose of both meeting our statistical assumptions and mitigating the computational burden, that $B \asymp \sqrt{n}$ for Mondrian random forests; selecting B for our debiased estimator requires a different set of conditions (see Sections 4 and 5). Large forests usually do not present significant computational challenges in practice as the ensemble estimator is easily parallelizable over the trees (see Section 6 for more discussion). We will emphasize explicitly where the "large forest" condition is important for our theory.

3 Inference with Mondrian random forests

Our analysis begins with a standard conditional bias–variance decomposition for the Mondrian random forest estimator:

$$\begin{aligned}
\hat{\mu}(x) - \mu(x) &= \Big(\hat{\mu}(x) - \mathbb{E}\big[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\big]\Big) + \Big(\mathbb{E}\big[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\big] - \mu(x)\Big) \\
&= \frac{1}{B} \sum_{b=1}^{B} \frac{\sum_{i=1}^{n} \big(\mu(X_i) - \mu(x)\big) \mathbb{I}\big\{X_i \in T_b(x)\big\}}{N_b(x)} \qquad (2) \\
&\quad + \frac{1}{B} \sum_{b=1}^{B} \frac{\sum_{i=1}^{n} \varepsilon_i \mathbb{I}\big\{X_i \in T_b(x)\big\}}{N_b(x)}. \qquad (3)
\end{aligned}$$

Our approach to estimation and inference is as follows. Firstly, we precisely characterize the probability limit of the "bias" term (2), and compute the second conditional moment of the "variance" term (3). This allows us to understand the bias–variance trade-off, and to derive upper bounds on the rate of convergence for the Mondrian random forest point estimator.

Secondly, we provide a central limit theorem for the "variance" term (3). By ensuring that the standard deviation dominates the conditional bias, we may conclude that a corresponding central limit theorem holds for the Mondrian random forest (1). With an appropriate estimator for the variance, we then establish procedures for valid and feasible statistical inference on the unknown regression function $\mu(x)$.

3.1 Bias and variance characterizations

We begin with (2), which captures the bias of the Mondrian random forest estimator conditional on the covariates $\mathbf{X}$ and the forest $\mathbf{T}$. The next lemma demonstrates that this conditional bias converges in L^2 at a certain rate, and provides a precise characterization of the resulting non-random limiting bias.

Lemma 1 (Bias)**.** *Suppose Assumptions 1 and 2 hold. For each $1 \leq r \leq \lfloor \beta/2 \rfloor$ there exists $B_r(x) \in \mathbb{R}$, which is a function of the derivatives of f and μ at x up to order $2r$, with*

$$\mathbb{E}\left[\left(\mathbb{E}\left[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\right] - \mu(x) - \sum_{r=1}^{\lfloor \beta/2 \rfloor} \frac{B_r(x)}{\lambda^{2r}}\right)^2\right] \lesssim \frac{1}{\lambda^{2\beta}} + \frac{1}{\lambda^{2(1 \wedge \beta)} B} + \frac{1}{\lambda^{2(1 \wedge \beta)}} \frac{\lambda^d}{n}. \qquad (4)$$

Whenever $\beta > 2$, the leading bias is the quadratic term

$$\frac{B_1(x)}{\lambda^2} = \frac{1}{2\lambda^2} \sum_{j=1}^{d} \frac{\partial^2 \mu(x)}{\partial x_j^2} + \frac{1}{2\lambda^2} \frac{1}{f(x)} \sum_{j=1}^{d} \frac{\partial \mu(x)}{\partial x_j} \frac{\partial f(x)}{\partial x_j}.$$

If $X_i \sim \mathrm{Unif}\left([0,1]^d\right)$ then $f(x) = 1$, and using multi-index notation we have

$$\frac{B_r(x)}{\lambda^{2r}} = \frac{1}{\lambda^{2r}} \sum_{|\nu| = r} \partial^{2\nu} \mu(x) \prod_{j=1}^{d} \frac{1}{\nu_j + 1}.$$

The bias characterization in Lemma 1 incorporates some high-degree polynomial terms in the lifetime parameter λ which for now may seem ignorable. The magnitude of the bias is determined by the leading term in (4), typically of order $1/\lambda^2$ whenever $\beta \geq 2$. This suffices for ensuring a negligible contribution from the bias with an appropriate choice of lifetime. However, the advantage of specifying higher-order terms will become apparent in Section 4, where we construct a debiased Mondrian random forest estimator, directly targeting and annihilating the higher-order terms in order to furnish superior estimation and inference properties. We also demonstrate numerically the detrimental role of bias in estimation and inference in Section 5.

In Lemma 1 we give some explicit examples of calculating the limiting bias when $\beta > 2$ or X_i are uniformly distributed. The general form of $B_r(x)$ is provided in Appendix A but is somewhat unwieldy except in specific situations. Nonetheless, the most important properties are that $B_r(x)$ are non-random and do not depend on the lifetime λ; these are crucial features for our debiasing procedure given in Section 4. If the forest size B does not diverge to infinity then we suffer the first-order conditional bias term $1/\big(\lambda^{1\wedge\beta}\sqrt{B}\big)$. This phenomenon was explained by Mourtada et al. (2020), who noted that it allows individual Mondrian trees ($B = 1$) to achieve minimax optimality in integrated mean squared error only when $\beta \in (0, 1]$. In contrast, large forests remove this first-order bias through ensemble averaging and as such are optimal for all $\beta \in (0, 2]$.

We now turn to (3), which captures the stochastic part of the Mondrian random forest. Lemma 2 determines the probability limit of the scaled conditional variance of this term, alongside its L^2 convergence rate. First, define

$$\tilde{\Sigma}(x) = \frac{n}{\lambda^d} \operatorname{Var}\big[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\big] \qquad \text{and} \qquad \Sigma(x) = \frac{\sigma^2(x)}{f(x)} \left(\frac{4 - 4\log 2}{3}\right)^d .$$

Lemma 2 (Variance)**.** *Suppose Assumptions 1 and 2 hold. Then*

$$\mathbb{E}\left[\big(\tilde{\Sigma}(x) - \Sigma(x)\big)^2\right] \lesssim \frac{\lambda^d}{n} + \frac{1}{B} + \frac{1}{\lambda^{2(1\wedge\beta_f\wedge\beta_\sigma)}}.$$

An upper bound on the L^2 rate of convergence of the Mondrian random forest estimator can immediately be deduced from the bias–variance decomposition, Lemma 1 and Lemma 2. This rate of convergence depends on the sequence of lifetime parameters λ; for optimal point estimation, we may balance the contributions from the bias and from the standard deviation by ensuring that $1/\lambda^{2\wedge\beta} + 1/\big(\lambda^{1\wedge\beta}\sqrt{B}\big) \asymp \sqrt{\lambda^d/n}$, or equivalently if $\lambda \asymp n^{\frac{1}{d+2(2\wedge\beta)}}$ and $B \gtrsim n^{\frac{2(2\wedge\beta)-2(1\wedge\beta)}{d+2(2\wedge\beta)}}$. We formalize these deductions in Theorem 1 and note that they imply that the Mondrian random forest is rate-minimax-optimal (Stone, 1982) in pointwise mean squared error for β-Hölder functions with $\beta \in (0, 2]$; a corresponding result for integrated mean squared error was provided by Mourtada et al. (2020, Theorem 2).

Theorem 1 (Mean squared error)**.** *Suppose Assumptions 1 and 2 hold. Then*

$$\mathbb{E}\left[\big(\hat{\mu}(x) - \mu(x)\big)^2\right] \lesssim \frac{\lambda^d}{n} + \frac{1}{\lambda^{2(2\wedge\beta)}} + \frac{1}{\lambda^{2(1\wedge\beta)}B}.$$

If further $\lambda \asymp n^{\frac{1}{d+2(2\wedge\beta)}}$ *and* $B \gtrsim n^{\frac{2(2\wedge\beta)-2(1\wedge\beta)}{d+2(2\wedge\beta)}}$*, then*

$$\mathbb{E}\left[\big(\hat{\mu}(x) - \mu(x)\big)^2\right] \lesssim n^{-\frac{2(2\wedge\beta)}{d+2(2\wedge\beta)}}.$$

We take this opportunity to contrast Mondrian random forests with classical nonparametric local smoothing methods. For example, the lifetime λ plays a similar role to the inverse bandwidth for kernel smoothing as it determines both the effective sample size n/λ^d and the scale of localization

$1/\lambda$, and thus also the associated rate of convergence. Likewise, $1/\lambda$ controls the diameter of a typical cell in Mondrian partition-based smoothing. However, due to the Mondrian process construction, some cells are typically "too small" (equivalent to an insufficiently large bandwidth) to give an appropriate effective sample size. In the same manner, classical methods based on non-random partitioning such as spline estimators typically impose a quasi-uniformity assumption to ensure all the cells are of comparable size (Huang, 2003; Cattaneo et al., 2020), a property which does not hold for the Mondrian process (not even with high probability).

3.2 Central limit theorem

Having discussed the point estimation properties of the Mondrian random forest estimator, we present a central limit theorem which forms the core of our methodology for performing statistical inference. As well as establishing asymptotic normality of the appropriately centered and scaled estimator, we also provide a rate of convergence in terms of a Berry–Esseen-style bound on the Kolmogorov–Smirnov distance from the normal distribution. In addition to precisely quantifying the quality of the Gaussian distributional approximation, this allows us to obtain explicit bounds on the coverage error of feasible confidence intervals.

Before stating the theorem, we highlight some of the challenges involved in establishing such a result. At first glance, the summands in (3) appear independent over $1 \le i \le n$, conditional on the forest $\mathbf{T}$, depending only on X_i and ε_i. However, the $N_b(x)$ appearing in the denominator depends on all X_i simultaneously, violating this independence assumption and rendering classical central limit theorems inapplicable. A natural preliminary attempt to resolve this issue is to observe that

$$N_b(x) = \sum_{i=1}^{n} \mathbb{I}\big\{X_i \in T_b(x)\big\} \approx n\,\mathbb{P}\big(X_i \in T_b(x) \mid T_b\big) \approx n f(x) |T_b(x)|$$

with high probability. One could attempt to use this by approximating the estimator with an average of i.i.d. random variables, or by employing a central limit theorem conditional on $\mathbf{X}$ and $\mathbf{T}$. However, such an approach fails because $\mathbb{E}\left[1/|T_b(x)|^2\right] = \infty$; the possible existence of small cells causes the law of the inverse cell volume to have heavy tails. For similar reasons, attempts to directly establish a central limit theorem based on $2+\delta$ moments, such as the classical Lyapunov central limit theorem, are ineffective.

We circumvent these problems by directly analyzing $\mathbb{I}\{N_b(x) \ge 1\}/N_b(x)$. We establish concentration properties for this non-linear function of X_i via the Efron–Stein inequality (Boucheron et al., 2013, Section 3.1) along with a sequence of delicate preliminary lemmas regarding inverse moments of truncated (conditional) binomial random variables. In particular, we show that $\mathbb{E}\left[\mathbb{I}\{N_b(x) \ge 1\}/N_b(x)\right] \lesssim \lambda^d/n$ and $\mathbb{E}\left[\mathbb{I}\{N_b(x) \ge 1\}/N_b(x)^2\right] \lesssim \lambda^{2d}(\log n)^d/n^2$. Asymptotic normality is then established by a careful application of a Berry–Esseen theorem (Petrov, 1995) conditional on $(\mathbf{X}, \mathbf{T})$. Appendix A provides all the technical details.

The following theorem gives our Berry–Esseen-type central limit theorem for the centered (zero mean conditional on the covariates and the trees) "variance" term from (3), scaled and standardized by its conditional variance $\tilde{\Sigma}(x)$. Note that on the event $\tilde{\Sigma}(x) = 0$, we also have $\hat{\mu}(x) = 0$ and $\mathbb{E}\left[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\right] = 0$, so continue to define $0/0 = 0$.

Theorem 2 (Central limit theorem)**.** *If Assumptions 1 and 2 hold, and $\mathbb{E}[|Y_i|^{2+\delta} \mid X_i]$ is bounded almost surely with $\delta > 0$, then*

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\sqrt{\frac{n}{\lambda^d}} \frac{\hat{\mu}(x) - \mathbb{E}\left[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\right]}{\sqrt{\tilde{\Sigma}(x)}} \le t \right) - \Phi(t) \right| \lesssim \left(\frac{\lambda^d}{n} \right)^{\frac{1 \wedge \delta}{2}} + \frac{1}{B}. \tag{5}$$

We make some remarks on Theorem 2. Firstly, since n/λ^d is the effective sample size and Y_i has only $2+\delta$ finite moments, the first term in (5) is likely to be unimprovable (Ibragimov and Linnik, 1975, Theorem 3.4.1). In particular, we attain the classical Berry–Esseen rate when $\mathbb{E}[|Y_i|^3 \mid \mathbf{X}]$ is bounded and $B \gtrsim \sqrt{n/\lambda^d}$.

The condition of $B \gtrsim (\log n)^d$ is central to our proof of Theorem 2, ensuring sufficient "mixing" of different Mondrian cells to escape the heavy-tailed phenomenon detailed in the preceding discussion. For concreteness, the large forest condition allows us to deal with expressions such as $\mathbb{E}\left[1/(|T_b(x)||T_{b'}(x)|)\right] = \mathbb{E}\left[1/|T_b(x)|\right]\mathbb{E}\left[1/|T_{b'}(x)|\right] \approx \lambda^{2d} < \infty$ where $b \neq b'$, by independence of the trees, rather than the "no ensembling" single tree analog $\mathbb{E}\left[1/|T_b(x)|^2\right] = \infty$.

Nonetheless, it is not clear whether the $1/B$ term is strictly necessary in (5) or if it is an artifact of the proof. When B is bounded, $\tilde{\Sigma}(x)$ remains random in the limit, and in fact it is not difficult to show that in this regime we have that $\mathbb{E}\big[(\tilde{\Sigma}(x))^2\big] \geq (\log n)^d$, which diverges (cf., Lemma 2). While these mildly pathological properties may not necessarily render the central limit theorem invalid, they certainly highlight some issues associated with inference based on a single tree or a small forest.

Theorem 2 applies only to the centered Mondrian random forest estimator; in order for it to be useful in a feasible inference setting, we must combine it with methods for controlling the conditional bias (see Lemma 1). In Section 4 we will show how the estimator can be debiased, giving weaker lifetime conditions for inference, improved rates of convergence, and superior coverage guarantees, whenever additional smoothness is available.

3.3 Confidence intervals

We demonstrate how to use our previous results to construct valid confidence intervals for the regression function $\mu(x)$. To do this, there are two preliminary issues which must be resolved. Firstly, the Berry–Esseen central limit theorem presented in Theorem 2 is stated for the Mondrian random forest estimator $\hat{\mu}(x)$ centered at its conditional expectation $\mathbb{E}\left[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\right]$, rather than at the true value $\mu(x)$. As such, we use Lemma 1 to ensure that the bias $\mathbb{E}\left[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\right] - \mu(x)$ is taken into account when establishing procedures for inference. Specifically, the bias should shrink faster than the standard deviation; this requires $1/\lambda^{2\wedge\beta} + 1/\big(\lambda^{1\wedge\beta}\sqrt{B}\big) \ll \sqrt{\lambda^d/n}$, which is satisfied by imposing the restrictions $\lambda \gg n^{\frac{1}{d+2(2\wedge\beta)}}$ and $B \gg n^{\frac{2(2\wedge\beta)-2(1\wedge\beta)}{d+2(2\wedge\beta)}}$ on the lifetime λ and forest size B.

The second issue is that the variances $\tilde{\Sigma}(x)$ and $\Sigma(x)$ depend on the unknown quantities $\sigma^2(x)$ and $f(x)$. To conduct feasible inference, we must therefore provide a consistent variance estimator. To this end, define

$$\hat{\sigma}^2(x) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}\frac{\big(Y_i - \hat{\mu}(x)\big)^2 \mathbb{I}\{X_i \in T_b(x)\}}{N_b(x)}, \tag{6}$$

$$\hat{\Sigma}(x) = \hat{\sigma}^2(x)\frac{n}{\lambda^d}\sum_{i=1}^{n}\left(\frac{1}{B}\sum_{b=1}^{B}\frac{\mathbb{I}\{X_i \in T_b(x)\}}{N_b(x)}\right)^2.$$

Lemma 3 (Variance estimation)**.** *If Assumptions 1 and 2 hold, and $\mathbb{E}[|Y_i|^{2+\delta} \mid X_i]$ is bounded almost surely with $\delta > 0$, then*

$$\left(\mathbb{E}\left[\left|\hat{\Sigma}(x) - \Sigma(x)\right|^{\frac{2-\mathbb{I}\{\delta<2\}}{2}}\right]\right)^{\frac{2}{2-\mathbb{I}\{\delta<2\}}} \lesssim \left(\frac{\lambda^d}{n}\right)^{\frac{1}{2}-\frac{\mathbb{I}\{\delta<2\}}{2+\delta}} + \frac{1}{\sqrt{B}} + \frac{1}{\lambda^{1\wedge\beta_\mu\wedge\beta_f\wedge\beta_\sigma}}.$$

For a confidence level $\alpha \in (0,1)$, Theorem 3 shows how to construct an asymptotically valid $100(1-\alpha)\%$ confidence interval for the regression function $\mu(x)$. The restrictions on the lifetime λ

and forest size B are the same as those previously discussed, and an explicit upper bound on the coverage error rate is provided. Define the interval estimator

$$\mathrm{CI}(x) = \left[\hat{\mu}(x) - \sqrt{\frac{\lambda^d}{n}} \hat{\Sigma}(x)^{1/2} q_{1-\alpha/2}, \ \hat{\mu}(x) - \sqrt{\frac{\lambda^d}{n}} \hat{\Sigma}(x)^{1/2} q_{\alpha/2}\right].$$

Theorem 3 (Confidence intervals)**.** *If Assumptions 1 and 2 hold, and $\mathbb{E}[|Y_i|^{2+\delta} \mid X_i]$ is bounded almost surely with $\delta > 0$, then*

$$\begin{aligned} &\left|\mathbb{P}\big(\mu(x) \in \mathrm{CI}(x)\big) - (1-\alpha)\right| \\ &\lesssim \frac{n}{\lambda^d} \frac{1}{\lambda^{2(2 \wedge \beta)}} + \left(\left(\frac{\lambda^d}{n}\right)^{1 - \frac{2\,\mathbb{I}\{\delta < 2\}}{2+\delta}} + \frac{1}{B} + \frac{1}{\lambda^{2(1 \wedge \beta_\mu \wedge \beta_f \wedge \beta_\sigma)}} + \frac{n}{\lambda^d} \frac{1}{\lambda^{2(1\wedge\beta)} B} \right)^{\frac{1}{5 + 2\,\mathbb{I}\{\delta<2\}}}. \end{aligned}$$

When coupled with an appropriate lifetime selection method (see Section 5), Theorem 3 gives a feasible procedure for uncertainty quantification in Mondrian random forests. Our procedure requires no adjustment of the original Mondrian random forest estimator beyond ensuring that the bias is negligible, and in particular does not rely on sample splitting. The construction of confidence intervals is just one corollary of the result given in Theorem 2; other applications include hypothesis testing based on the value of $\mu(x)$ at a design point x by inversion of the confidence interval, as well as specification testing by comparison with a $\sqrt{n}$-consistent parametric regression estimator. The construction of simultaneous confidence intervals for finitely many points $x_1, \ldots, x_D$ can be accomplished either using standard multiple testing corrections or by first establishing a multivariate central limit theorem using the Cramér–Wold device and formulating a consistent variance matrix estimator.

4 Debiased Mondrian random forests

We give our next main contribution: a novel variant of the Mondrian random forest estimator that corrects for higher-order bias with an approach based on generalized jackknifing (Schucany and Sommers, 1977). This estimator retains the basic form of a Mondrian random forest in the sense that it is a linear combination of Mondrian tree estimators, but in this section we allow for non-identical linear coefficients, some of which may be negative, and for differing lifetime parameters across the trees. Since the basic Mondrian random forest estimator is a special case of this more general debiased version, we will discuss only the latter throughout the rest of the paper.

We use the explicit form of the bias given in Lemma 1 to construct the debiased Mondrian forest estimator as follows, letting $J \geq 0$ be the bias correction order. With $J = 0$ we preserve the original Mondrian random forest, with $J = 1$ we remove second-order bias, and with $J = \lfloor \underline{\beta}/2 \rfloor$ we remove bias terms up to and including order $2\lfloor \underline{\beta}/2 \rfloor$, giving the maximum possible bias reduction achievable in the Hölder class $\mathcal{H}^\beta$ (Stone, 1982). As such, only bias terms of order $1/\lambda^\beta$ will remain.

For $0 \leq r \leq J$, let $\hat{\mu}_r(x)$ be a Mondrian forest estimator based on the trees $T_{br} \sim \mathcal{M}\big([0,1]^d, \lambda_r\big)$ for $1 \leq b \leq B$, where $\lambda_r = a_r \lambda$ for some $a_r > 0$ and $\lambda > 0$. Write $\mathbf{T}$ to denote the collection of all the trees, and suppose they are mutually independent. We find values of a_r along with coefficients $\omega_r \in \mathbb{R}$ which annihilate the leading J bias terms of the debiased Mondrian random forest estimator

$$\hat{\mu}_\mathrm{d}(x) = \sum_{r=0}^{J} \omega_r \hat{\mu}_r(x) = \sum_{r=0}^{J} \omega_r \frac{1}{B} \sum_{b=1}^{B} \frac{\sum_{i=1}^n Y_i \,\mathbb{I}\big\{X_i \in T_{br}(x)\big\}}{N_{br}(x)}. \tag{7}$$

This ensemble estimator retains the "forest" structure of the original estimators, but with varying lifetime parameters λ_r and coefficients ω_r. Thus, referring to (4), we desire

$$\sum_{r=0}^{J} \omega_r \left(\mu(x) + \sum_{s=1}^{J} \frac{B_s(x)}{a_r^{2s} \lambda^{2s}} \right) = \mu(x)$$

for all λ, or equivalently the system of linear equations $\sum_{r=0}^{J} \omega_r = 1$ and $\sum_{r=0}^{J} \omega_r a_r^{-2s} = 0$ for each $1 \leq s \leq J$. We solve these as follows. Define the $(J+1) \times (J+1)$ Vandermonde matrix $A_{rs} = a_{r-1}^{2-2s}$, let $\omega = (\omega_0, \ldots, \omega_J)^\mathsf{T} \in \mathbb{R}^{J+1}$ and take $e_0 = (1, 0, \ldots, 0)^\mathsf{T} \in \mathbb{R}^{J+1}$. Then a solution for the debiasing coefficients is given by $\omega = A^{-1} e_0$ whenever A is non-singular. In practice we can take a_r to be a fixed geometric or arithmetic sequence to ensure this is the case, appealing to the Vandermonde determinant formula: $\det A = \prod_{0 \leq r < s \leq J} (a_r^{-2} - a_s^{-2}) \neq 0$ whenever a_r are distinct. For example, one could set $a_r = (1+\gamma)^r$ or $a_r = 1 + \gamma r$ for some $\gamma > 0$. Because we assume β, and therefore the choice of J, do not depend on n, there is no need to quantify the invertibility of A by, for example, bounding its eigenvalues away from zero as a function of J and the choice of a_r.

The debiased Mondrian random forest estimator defined in (7) is a linear combination of standard Mondrian random forests, and as such contains both a sum over $0 \leq r \leq J$, representing the debiasing procedure, and a sum over $1 \leq b \leq B$, representing the forest averaging. We have been interpreting this estimator as a debiased version of the standard Mondrian random forest given in (1), but it is equally valid to swap the order of these sums. This gives rise to an alternative point of view: we replace each Mondrian random tree with a "debiased" version, and then take a forest of such modified trees. This perspective is perhaps more in line with existing techniques for constructing randomized ensembles, where the outermost operation represents a B-fold average of randomized base learners (not necessarily locally constant decision trees), each of which has a small bias component (Caruana et al., 2004; Zhou and Feng, 2019).

4.1 Bias and variance characterizations

In Lemma 4 we verify that this debiasing procedure does indeed annihilate the desired bias terms; it is a direct consequence of Lemma 1 and the construction of the debiased Mondrian random forest estimator $\hat{\mu}_\mathrm{d}(x)$.

Lemma 4 (Bias of the debiased estimator)**.** *Suppose Assumptions 1 and 2 hold. Then in the notation of Lemma 1 and with* $\bar{\omega} = \sum_{r=0}^{J} \omega_r a_r^{-2J-2}$,

$$\mathbb{E}\left[\left(\mathbb{E}\big[\hat{\mu}_\mathrm{d}(x) \mid \mathbf{X}, \mathbf{T}\big] - \mu(x) - \mathbb{I}\{2J+2 < \beta\} \frac{\bar{\omega} B_{J+1}(x)}{\lambda^{2J+2}} \right)^2 \right] \lesssim \frac{1}{\lambda^{2((2J+4) \wedge \beta)}} + \frac{1}{\lambda^{2(1 \wedge \beta)} B} + \frac{1}{\lambda^{2(1 \wedge \beta)}} \frac{\lambda^d}{n}.$$

Lemma 4 has the following consequence: the leading bias term is characterized in terms of $B_{J+1}(x)$ whenever $J < \beta/2 - 1$, or equivalently $J < \lfloor \beta/2 \rfloor$, that is, the debiasing order J does not exhaust the Hölder smoothness β. If this condition does not hold, then the estimator is fully debiased; the resulting leading bias term is bounded above by $1/\lambda^\beta$ up to constants but its form is left unspecified.

The following lemma controls the variance of the debiased Mondrian random forest estimator. With $\ell_{rr'} = 2a_r(1 - a_r \log(1 + a_{r'}/a_r)/a_{r'})/3$, define

$$\tilde{\Sigma}_\mathrm{d}(x) = \frac{n}{\lambda^d} \operatorname{Var}\big[\hat{\mu}_\mathrm{d}(x) \mid \mathbf{X}, \mathbf{T}\big] \qquad \text{and} \qquad \Sigma_\mathrm{d}(x) = \frac{\sigma^2(x)}{f(x)} \sum_{r=0}^{J} \sum_{r'=0}^{J} \omega_r \omega_{r'} \left(\ell_{rr'} + \ell_{r'r}\right)^d.$$

Lemma 5 (Variance of the debiased estimator)**.** *Supposing Assumptions 1 and 2 hold,*

$$\mathbb{E}\left[\left(\tilde{\Sigma}_{\mathrm{d}}(x) - \Sigma_{\mathrm{d}}(x)\right)^2\right] \lesssim \frac{\lambda^d}{n} + \frac{1}{B} + \frac{1}{\lambda^{2(1 \wedge \beta_f \wedge \beta_\sigma)}}.$$

4.2 Minimax optimality

Our next main result, Theorem 4, shows that when using an appropriate sequence of lifetime parameters λ, the debiased Mondrian random forest estimator achieves, up to constants, the minimax-optimal rate of convergence for pointwise mean squared error estimation of a d-dimensional regression function $\mu \in \mathcal{H}^\beta$ (Stone, 1982). This result holds for all $d \geq 1$ and all $\beta > 0$, complementing a previous result (see Theorem 1) established only for $\beta \in (0, 2]$ and in integrated mean squared error by Mourtada et al. (2020).

Theorem 4 (Mean squared error of the debiased estimator)**.** *Grant Assumptions 1 and 2. Then*

$$\mathbb{E}\left[\left(\hat{\mu}_{\mathrm{d}}(x) - \mu(x)\right)^2\right] \lesssim \frac{\lambda^d}{n} + \frac{1}{\lambda^{2((2J+2) \wedge \beta)}} + \frac{1}{\lambda^{2(1 \wedge \beta)} B}.$$

Thus with $J \geq \lfloor \beta/2 \rfloor$, $\lambda \asymp n^{\frac{1}{d+2\beta}}$ and $B \gtrsim n^{\frac{2\beta - 2(1 \wedge \beta)}{d+2\beta}}$, we have

$$\mathbb{E}\left[\left(\hat{\mu}_{\mathrm{d}}(x) - \mu(x)\right)^2\right] \lesssim n^{-\frac{2\beta}{d+2\beta}}.$$

The sequence of lifetime parameters λ required in Theorem 4 is chosen to balance the bias and standard deviation bounds implied by Lemmas 4 and 5 respectively, in order to minimize the pointwise mean squared error. While selecting an optimal debiasing order J needs only knowledge of an upper bound on the smoothness β, choosing an optimal sequence of λ values does assume that β is known a priori. The problem of adapting to β from data is beyond the scope of this paper; we provide some practical advice for tuning parameter selection in Section 5.

Theorem 4 complements the minimaxity results proven by Mourtada et al. (2020) for Mondrian trees (with $\beta \leq 1$) and for Mondrian random forests (with $\beta \leq 2$), with one modification: our version is stated in pointwise rather than integrated mean squared error. This is because our debiasing procedure is designed to handle interior smoothing bias and as such does not provide any correction for boundary bias. We leave the development of such boundary corrections to future work, but constructions similar to higher-order boundary-correcting kernels should be possible. If the region of integration is a compact set in the interior of $[0, 1]^d$ then we do obtain an optimal integrated mean squared error bound: if $a \in (0, 1/2)$ is fixed then under the same conditions as Theorem 4, with appropriate tuning of λ and B,

$$\mathbb{E}\left[\int_{[a,1-a]^d} \left(\hat{\mu}_{\mathrm{d}}(x) - \mu(x)\right)^2 \mathrm{d}x\right] \lesssim \frac{\lambda^d}{n} + \frac{1}{\lambda^{2\beta}} + \frac{1}{\lambda^{2(1 \wedge \beta)} B} \lesssim n^{-\frac{2\beta}{d+2\beta}}.$$

4.3 Central limit theorem

In Theorem 5, we verify that a central limit theorem holds for the debiased random forest estimator $\hat{\mu}_{\mathrm{d}}(x)$. The strategy and challenges associated with proving Theorem 5 are identical to those discussed earlier surrounding Theorem 2. In fact in Appendix A we provide a direct proof only for Theorem 5 and deduce Theorem 2 as a special case. Again on the event $\tilde{\Sigma}_{\mathrm{d}}(x) = 0$, we also have $\hat{\mu}_{\mathrm{d}}(x) = 0$ and $\mathbb{E}\left[\hat{\mu}_{\mathrm{d}}(x) \mid \mathbf{X}, \mathbf{T}\right] = 0$, so we take $0/0 = 0$.

Theorem 5 (Central limit theorem with debiasing)**.** *Suppose Assumptions 1 and 2 hold, and* $\mathbb{E}[|Y_i|^{2+\delta} \mid X_i]$ *is bounded almost surely with* $\delta > 0$*. Then*

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\sqrt{\frac{n}{\lambda^d}} \frac{\hat{\mu}_{\mathrm{d}}(x) - \mathbb{E}\left[\hat{\mu}_{\mathrm{d}}(x) \mid \mathbf{X}, \mathbf{T}\right]}{\sqrt{\tilde{\Sigma}_{\mathrm{d}}(x)}} \leq t \right) - \Phi(t) \right| \lesssim \left(\frac{\lambda^d}{n}\right)^{\frac{1 \wedge \delta}{2}} + \frac{1}{B}.$$

4.4 Confidence intervals

As before, to conduct valid feasible inference we must ensure that the bias (now significantly reduced due to our debiasing procedure) is negligible when compared to the standard deviation (which is of the same order as before). We treat here the general "partial debiasing" setting where either the debiasing order J or the Hölder smoothness β may determine the magnitude of the bias, which is $1/\lambda^{(2J+2)\wedge\beta}$. For optimal results, one should take $J \geq \lfloor \beta/2 \rfloor$ to ensure total debiasing, as in Theorem 4. We thus require $1/\lambda^{(2J+2)\wedge\beta} + 1/\left(\lambda^{1\wedge\beta}\sqrt{B}\right) \ll \sqrt{\lambda^d/n}$, satisfied by imposing $\lambda \gg n^{\frac{1}{d+2((2J+2)\wedge\beta)}}$ and $B \gg n^{\frac{2((2J+2)\wedge\beta)-2(1\wedge\beta)}{d+2((2J+2)\wedge\beta)}}$ on the lifetime parameter λ and forest size B.

Once again, we propose a variance estimator and show that it is consistent. With $\hat{\sigma}^2(x)$ as in (6) in Section 3, define

$$\hat{\Sigma}_{\mathrm{d}}(x) = \hat{\sigma}^2(x) \frac{n}{\lambda^d} \sum_{i=1}^{n} \left(\sum_{r=0}^{J} \omega_r \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{I}\{X_i \in T_{br}(x)\}}{N_{br}(x)} \right)^2. \tag{8}$$

Lemma 6 (Variance estimation for the debiased estimator)**.** *Suppose Assumptions 1 and 2 hold, and* $\mathbb{E}[|Y_i|^{2+\delta} \mid X_i]$ *is bounded almost surely with* $\delta > 0$*. Then*

$$\left(\mathbb{E}\left[\left| \hat{\Sigma}_{\mathrm{d}}(x) - \Sigma_{\mathrm{d}}(x) \right|^{\frac{2 - \mathbb{I}\{\delta<2\}}{2}} \right] \right)^{\frac{2}{2-\mathbb{I}\{\delta<2\}}} \lesssim \left(\frac{\lambda^d}{n}\right)^{\frac{1}{2} - \frac{\mathbb{I}\{\delta<2\}}{2+\delta}} + \frac{1}{\sqrt{B}} + \frac{1}{\lambda^{1\wedge\beta_\mu\wedge\beta_f\wedge\beta_\sigma}}.$$

In analogy to Section 3, we now demonstrate the construction of feasible valid confidence intervals using the debiased Mondrian random forest estimator in Theorem 6. Consider the debiased $100(1-\alpha)\%$ confidence interval estimator

$$\mathrm{CI}_{\mathrm{d}}(x) = \left[\hat{\mu}_{\mathrm{d}}(x) - \sqrt{\frac{\lambda^d}{n}} \hat{\Sigma}_{\mathrm{d}}(x)^{1/2} \, q_{1-\alpha/2}, \ \hat{\mu}_{\mathrm{d}}(x) - \sqrt{\frac{\lambda^d}{n}} \hat{\Sigma}_{\mathrm{d}}(x)^{1/2} \, q_{\alpha/2} \right]. \tag{9}$$

Theorem 6 (Confidence intervals with debiasing)**.** *Suppose Assumptions 1 and 2 hold, and* $\mathbb{E}[|Y_i|^{2+\delta} \mid X_i]$ *is bounded almost surely with* $\delta > 0$*. Then*

$$\begin{aligned} &\left| \mathbb{P}\left(\mu(x) \in \mathrm{CI}_{\mathrm{d}}(x)\right) - (1-\alpha) \right| \\ &\lesssim \frac{n}{\lambda^d} \frac{1}{\lambda^{2((2J+2)\wedge\beta)}} + \left(\left(\frac{\lambda^d}{n}\right)^{1 - \frac{2\,\mathbb{I}\{\delta<2\}}{2+\delta}} + \frac{1}{B} + \frac{1}{\lambda^{2(1\wedge\beta_\mu\wedge\beta_f\wedge\beta_\sigma)}} + \frac{n}{\lambda^d} \frac{1}{\lambda^{2(1\wedge\beta)} B} \right)^{\frac{1}{5+2\,\mathbb{I}\{\delta<2\}}}. \end{aligned}$$

One important benefit of our debiasing technique is made clear in Theorem 6: the restrictions imposed on the lifetime parameter λ are substantially relaxed, especially in smooth classes with large β. As well as the high-level of benefit of relaxed conditions, this is also useful for practical selection of appropriate lifetimes for estimation and inference respectively; see Section 5 for more details. Nonetheless, such improvements do not come without concession. The limiting variance of the debiased estimator is typically larger than that of the unbiased version in small samples (the extent of this increase depends on the choice of the debiasing parameters a_r), leading to wider confidence intervals and larger estimation error, despite the theoretical asymptotic improvements. Nonetheless, the empirical results in Section 5 demonstrate that the debiasing effect can overcome the increased variance with moderate sample sizes. Because we employ symmetric confidence intervals, the coverage error depends on the squared bias $1/\lambda^{2((2J+2)\wedge\beta)}$, whereas the corresponding Berry–Esseen rate would depend on the (larger) linear bias $1/\lambda^{(2J+2)\wedge\beta}$.

5 Implementation and empirical results

We discuss procedures for selecting the parameters involved in fitting a debiased Mondrian random forest; namely the base lifetime parameter λ, the number of trees in each forest B, the order of the bias correction J, and the debiasing scale parameters a_r for $0 \leq r \leq J$. We then provide empirical results with simulated data to demonstrate the effectiveness of our methods.

5.1 Tuning parameter selection

The most important parameter is the base Mondrian lifetime λ, which plays the role of a complexity parameter and thus governs the overall bias–variance trade-off of the estimator. Correct tuning of λ is especially important in two main respects: firstly, in order to use the central limit theorem established in Theorem 5, we must have that the bias converges to zero, requiring $\lambda \gg n^{\frac{1}{d+2((2J+2)\wedge\beta)}}$. Secondly, the minimax optimality result of Theorem 4 is valid only in the regime $\lambda \asymp n^{\frac{1}{d+2\beta}}$, and so λ requires careful determination in practice. For clarity, in this section we use the notation $\hat{\mu}_{\mathrm{d}}(x;\lambda,J)$ for the debiased Mondrian random forest estimator implemented with lifetime λ and debiasing order J, as in (7). Similarly write $\hat{\Sigma}_{\mathrm{d}}(x;\lambda,J)$ for the associated variance estimator (8).

For minimax-optimal point estimation when β is known, choose any sequence $\lambda \asymp n^{\frac{1}{d+2\beta}}$ and use $\hat{\mu}_{\mathrm{d}}(x;\lambda,J)$ with $J = \lfloor \underline{\beta}/2 \rfloor$, following the theory given in Theorem 4. For an explicit example of how to choose the lifetime, one can instead use $\hat{\mu}_{\mathrm{d}}\big(x;\hat{\lambda}_{J-1},J-1\big)$ so that the leading bias is explicitly characterized by Lemma 4, and with $\hat{\lambda}_{J-1}$ as defined below. This estimator is however not minimax-optimal as the debiasing order of $J-1<J$ does not satisfy the conditions of Theorem 4.

For performing inference, a more careful procedure is required; we suggest the following, when $\beta > 2$ is known. Set $J = \lfloor \underline{\beta}/2 \rfloor$ as before, and use $\hat{\mu}_{\mathrm{d}}\big(x;\hat{\lambda}_{J-1},J\big)$ and $\hat{\Sigma}_{\mathrm{d}}\big(x;\hat{\lambda}_{J-1},J\big)$ to construct a confidence interval (9), so that one selects a lifetime tailored for a more biased estimator than that which is actually used. This results in an inflated lifetime estimate, guaranteeing the resulting bias is negligible when it is plugged into the fully debiased estimator. This approach to tuning parameter selection and debiasing for nonparametric inference corresponds to an application of robust bias correction (Calonico et al., 2018, 2022), where the point estimator is bias-corrected and the robust standard error estimator incorporates the additional variability introduced by the correction. This gives a refined distributional approximation but may not exhaust the underlying smoothness of the regression function. An alternative approach based on Lepski's method (Lepski, 1992; Birgé, 2001) could be developed with the latter goal in mind.

It remains to propose a concrete method for computing $\hat{\lambda}_J$ in finite samples; we suggest a procedure based on minimizing the asymptotic mean squared error (AMSE) using plug-in selection with polynomial estimation, building on classical ideas from the nonparametric smoothing literature. Expressions for the AMSE are available as direct consequences of Lemmas 4 and 5, provided that $J < \lfloor \underline{\beta}/2 \rfloor$ so the Hölder smoothness is not fully exhausted.

Selecting the lifetime parameter λ with polynomial estimation

For implementation, we propose a simple rule-of-thumb approach. Suppose that $X_i \sim \mathrm{Unif}\left([0,1]^d\right)$ and that the leading bias of $\hat{\mu}_{\mathrm{d}}(x)$ is well approximated by an additively separable function so that, writing $\partial_j^{2J+2}\mu(x)$ for $\partial_j^{2J+2}\mu(x)/\partial x_j^{2J+2}$, the asymptotic bias is

$$\mathrm{ABias}(x;\lambda,J) = \frac{\bar{\omega} B_{J+1}(x)}{\lambda^{2J+2}} = \frac{1}{\lambda^{2J+2}} \frac{\bar{\omega}}{J+2} \sum_{j=1}^{d} \partial_j^{2J+2}\mu(x).$$

Suppose that the model is homoscedastic so $\sigma^2(x) = \sigma^2$ and the asymptotic variance of $\hat{\mu}_{\mathrm{d}}$ is

$$\mathrm{AVar}(x;\lambda,J) = \frac{\lambda^d}{n}\Sigma_{\mathrm{d}}(x) = \frac{\lambda^d \sigma^2}{n} \sum_{r=0}^{J}\sum_{r'=0}^{J} \omega_r \omega_{r'} \left(\ell_{rr'} + \ell_{r'r}\right)^d.$$

The asymptotic mean squared error is therefore

$$\mathrm{AMSE}(x;\lambda,J) = \frac{1}{\lambda^{4J+4}} \frac{\bar{\omega}^2}{(J+2)^2} \left(\sum_{j=1}^{d} \partial_j^{2J+2}\mu(x)\right)^2 + \frac{\lambda^d\sigma^2}{n} \sum_{r=0}^{J}\sum_{r'=0}^{J} \omega_r \omega_{r'} \left(\ell_{rr'} + \ell_{r'r}\right)^d.$$

Minimizing over $\lambda > 0$ yields the AMSE-optimal lifetime parameter

$$\lambda_J = \left(\frac{\frac{(4J+4)\bar{\omega}^2}{(J+2)^2} n \left(\sum_{j=1}^{d} \partial_j^{2J+2}\mu(x)\right)^2}{d\sigma^2 \sum_{r=0}^{J}\sum_{r'=0}^{J} \omega_r\omega_{r'} \left(\ell_{rr'} + \ell_{r'r}\right)^d} \right)^{\frac{1}{4J+4+d}}.$$

An estimator of λ_J is given by the plug-in procedure

$$\hat{\lambda}_J = \left(\frac{\frac{(4J+4)\bar{\omega}^2}{(J+2)^2} n \left(\sum_{j=1}^{d} \partial_j^{2J+2}\hat{\mu}(x)\right)^2}{d\hat{\sigma}^2 \sum_{r=0}^{J}\sum_{r'=0}^{J} \omega_r\omega_{r'} \left(\ell_{rr'} + \ell_{r'r}\right)^d} \right)^{\frac{1}{4J+4+d}}$$

for some preliminary estimators $\partial_j^{2J+2}\hat{\mu}(x)$ and $\hat{\sigma}^2$. These can be obtained by fitting a global polynomial regression to the data $(\mathbf{X}, \mathbf{Y})$ of order $2J+4$ without interaction terms. To do this, define the $n \times ((2J+4)d+1)$ design matrix $\mathbf{P}$ with rows given by

$$P(X_i) = \left(1, X_{i1}, X_{i1}^2, \ldots, X_{i1}^{2J+4}, X_{i2}, X_{i2}^2, \ldots, X_{i2}^{2J+4}, \ldots, X_{id}, X_{id}^2, \ldots, X_{id}^{2J+4}\right).$$

Then the derivative estimator is

$$\begin{aligned}
\partial_j^{2J+2}\hat{\mu}(x) &= \partial_j^{2J+2} P(x) \big(\mathbf{P}^\mathsf{T}\mathbf{P}\big)^{-1}\mathbf{P}^\mathsf{T}\mathbf{Y} \\
&= (2J+2)! \left(0_{1+(j-1)(2J+4)+(2J+1)}, 1, x_j, x_j^2/2, 0_{(d-j)(2J+4)}\right) \left(\mathbf{P}^\mathsf{T}\mathbf{P}\right)^{-1}\mathbf{P}^\mathsf{T}\mathbf{Y},
\end{aligned}$$

and the variance estimator $\hat{\sigma}^2$ is the based on the residual sum of squared errors of this model:

$$\hat{\sigma}^2 = \frac{1}{n - (2J+4)d - 1}\big(\mathbf{Y}^\mathsf{T}\mathbf{Y} - \mathbf{Y}^\mathsf{T}\mathbf{P}\big(\mathbf{P}^\mathsf{T}\mathbf{P}\big)^{-1}\mathbf{P}^\mathsf{T}\mathbf{Y}\big).$$

Choosing the number B of trees in each forest

The next parameter to choose is the number of trees in each forest. If no debiasing is applied, we suggest taking $B \asymp \sqrt{n}$ to ensure the coverage error in Theorem 3 converges to zero. If debiasing is used then we recommend setting $B \asymp n^{\frac{2J-1}{2J}}$, consistent with Theorem 4 and Theorem 6.

Setting the debiasing order J

Deciding how many orders of bias to remove requires knowledge of the Hölder smoothness of μ and f, which is in practice very difficult to estimate statistically. As such we recommend removing only the first one or two bias terms, taking $J \in \{0, 1, 2\}$ to avoid inflating the variance of the estimator.

Selecting the debiasing scalars a_r

As in Section 4, take a fixed geometric or arithmetic sequence. For example, $a_r = (1+\gamma)^r$ or $a_r = 1 + \gamma r$ where $\gamma > 0$; we suggest $a_r = (3/2)^r$.

5.2 Empirical results

To demonstrate the empirical properties of our proposed estimation and inference methodology, we present results with simulated data. Throughout this section we use the data generating process given by uniform covariates $X_i \sim \mathrm{Unif}[0,1]^d$ for $d \in \{1, 2\}$, a sinusoidal regression function $\mu(x) = \sum_{j=1}^d \sin(\pi x_j)$, and homoscedastic normal errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 3/10$. We focus on estimation at the design point $x = (1/2, \ldots, 1/2) \in \mathbb{R}^d$, and use $n = 1000$ samples and $B = 800$ trees in each forest. We demonstrate our procedures both with and without debiasing by setting $J = 1$ and $J = 0$ respectively, and when $J = 1$ we use the debiasing scalars $(a_1, a_2) = (1, 3/2)$ suggested in Section 4, yielding debiasing coefficients of $(\omega_0, \omega_1) = (-4/5, 9/5)$. For lifetime selection (LS), we first show our estimator $\hat{\lambda}_J$ based on polynomial regression (Section 5.1), and then present the infeasible oracle lifetime λ_J which exactly minimizes the asymptotic mean squared error. To illustrate robustness with respect to this tuning parameter, we repeat the same experiments but rescaling λ_J by a lifetime multiplier $\mathrm{LM} \in \{1 \pm \ell/5 : 0 \le \ell \le 2\}$. We further exhibit the robust bias correction (BC) approach discussed in Section 5 by using a debiased estimator ($J = 1$) with the AMSE-optimal lifetime parameter λ_0.

For each such estimator, we present the empirical root mean squared error (RMSE), bias, standard deviation (SD), and absolute bias/SD ratio, based on 3000 repeats. We also show the estimated standard deviation $\widehat{\mathrm{SD}} = \sqrt{\lambda^d/n}\, \hat{\Sigma}_{\mathrm{d}}(x)$, as well as the estimated variance of the errors $\hat{\sigma}^2(x)$. Since oracle properties available, we give the asymptotic bias (ABias) and asymptotic standard deviation (ASD). Finally, we present the empirical coverage rate (CR) of nominal 95% confidence intervals along with their empirical average widths (CIW).

We now derive the asymptotic oracle properties of our estimators. Firstly, by Lemma 2, the asymptotic variance of the estimator without debiasing is

$$\mathrm{AVar} = \frac{\lambda^d}{n} \Sigma(x) = \frac{\lambda^d}{n} \frac{\sigma^2(x)}{f(x)} \left(\frac{4 - 4\log 2}{3} \right)^d \approx \frac{\lambda^d \sigma^2}{n} 0.4091^d.$$

By Lemma 5, $\ell_{00} = \frac{2}{3}(1 - \log 2)$, $\ell_{01} = \frac{2}{3}\left(1 - \frac{2}{3}\log\frac{5}{2}\right)$, $\ell_{10} = 1 - \frac{3}{2}\log\frac{5}{3}$, and $\ell_{11} = 1 - \log 2$, so the asymptotic variance of its debiased counterpart is

$$\mathrm{AVar_d} = \frac{\lambda^d}{n}\Sigma_\mathrm{d}(x) = \frac{\lambda^d}{n}\frac{\sigma^2(x)}{f(x)}\sum_{r=0}^{J}\sum_{r'=0}^{J}\omega_r\omega_{r'}\left(\ell_{rr'}+\ell_{r'r}\right)^d$$
$$\approx \frac{\lambda^d\sigma^2}{n}\left(0.64\cdot 0.4091^d - 2.88\cdot 0.4932^d + 3.24\cdot 0.6137^d\right).$$

We similarly establish the asymptotic biases. Without debiasing, by Lemma 1,

$$\mathrm{ABias} = \frac{1}{\lambda^2}\sum_{|\nu|=1}\partial^{2\nu}\mu(x)\prod_{j=1}^{d}\frac{1}{\nu_j+1} = \frac{1}{2\lambda^2}\sum_{j=1}^{d}\frac{\partial^2\mu(x)}{\partial x_j^2} = -\frac{\pi^2}{2\lambda^2}\sum_{j=1}^{d}\sin(\pi x_j) = -\frac{\pi^2 d}{2\lambda^2}.$$

For the debiased estimator, with $\bar{\omega} = \omega_0 + \omega_1 a_1^{-4} = -4/9$, we recover

$$\mathrm{ABias_d} = \frac{\bar{\omega}B_2(x)}{\lambda^4} = -\frac{4}{9\lambda^4}\sum_{|\nu|=2}\partial^{2\nu}\mu(x)\prod_{j=1}^{d}\frac{1}{\nu_j+1} = -\frac{4}{27\lambda^4}\sum_{j=1}^{d}\frac{\partial^4\mu(x)}{\partial^4 x_j} = -\frac{4\pi^4 d}{27\lambda^4}.$$

Table 1 gives results in the one-dimensional setting ($d=1$). Firstly, observe that the polynomial lifetime estimator appears to be moderately accurate, displaying some oversmoothing when fitting a polynomial of order 4 (for $\hat{\lambda}_0$) and some undersmoothing with a polynomial of order 6 (with $\hat{\lambda}_1$). The effects of debiasing on RMSE are clear, with the appropriately tuned debiased Mondrian forest ($J=1$, λ_1) providing the best results (Theorem 4). Likewise, the effect of debiasing is apparent when using an undersmoothed lifetime ($J=1$, λ_0), with the bias being significantly reduced (see Lemma 4) at the expense of a larger standard deviation. The variance estimator performs well, with $\widehat{\mathrm{SD}}$ a good approximation for the finite-sample SD, and $\hat{\sigma}^2$ similarly sits close to $\sigma^2 = 0.09$. The value of robust bias correction ($J=1$, λ_0) for statistical inference is clear, with the coverage rates clustering around the nominal 95% even with perturbed lifetime values (see Theorem 6). In contrast, the no-debiasing estimator ($J=0$, λ_0) fails to attain correct coverage, while its fully debiased counterpart ($J=1$, λ_1) lacks robustness, reaching the nominal level only with larger lifetime values. Accurate coverage is at the expense of wider confidence intervals, but the differences are not large.

	J	LS	LM	λ	RMSE	Bias	SD	Bias/SD	$\widehat{\text{SD}}$	$\hat{\sigma}^2$	ARMSE	ABias	ASD	CR	CIW
No debiasing	0	$\hat{\lambda}_0$	1.0	14.73	0.0351	-0.0250	0.0247	1.0123	0.0236	0.0931	0.0369	-0.0270	0.0232	82.5%	0.093
		λ_0	1.2	23.10	0.0307	-0.0092	0.0293	0.3130	0.0292	0.0894	0.0306	-0.0092	0.0292	93.6%	0.114
			1.1	21.18	0.0300	-0.0109	0.0280	0.3888	0.0280	0.0897	0.0300	-0.0110	0.0279	93.4%	0.110
			1.0	19.25	0.0297	-0.0131	0.0267	0.4909	0.0267	0.0901	0.0298	-0.0133	0.0266	92.9%	0.105
			0.9	17.33	0.0300	-0.0160	0.0253	0.6326	0.0254	0.0907	0.0301	-0.0164	0.0253	90.7%	0.100
			0.8	15.40	0.0312	-0.0201	0.0238	0.8438	0.0241	0.0916	0.0316	-0.0208	0.0238	87.5%	0.095
Debiasing	1	$\hat{\lambda}_1$	1.0	11.14	0.0301	-0.0031	0.0300	0.1036	0.0302	0.1002	0.0296	-0.0026	0.0287	95.0%	0.119
		λ_1	1.2	7.86	0.0255	-0.0070	0.0246	0.2835	0.0269	0.1103	0.0245	-0.0038	0.0242	95.9%	0.106
			1.1	7.21	0.0255	-0.0095	0.0236	0.4031	0.0263	0.1147	0.0238	-0.0053	0.0232	95.2%	0.103
			1.0	6.55	0.0264	-0.0135	0.0227	0.5950	0.0256	0.1198	0.0235	-0.0078	0.0221	94.3%	0.100
			0.9	5.90	0.0288	-0.0191	0.0216	0.8817	0.0249	0.1259	0.0241	-0.0119	0.0210	90.6%	0.097
			0.8	5.24	0.0343	-0.0274	0.0206	1.3346	0.0240	0.1329	0.0275	-0.0191	0.0198	82.0%	0.094
Robust BC	1	$\hat{\lambda}_0$	1.0	14.73	0.0334	-0.0014	0.0333	0.0405	0.0339	0.0940	0.0336	-0.0011	0.0330	95.3%	0.133
		λ_0	1.2	23.10	0.0420	-0.0004	0.0420	0.0105	0.0419	0.0898	0.0415	-0.0001	0.0415	94.8%	0.164
			1.1	21.18	0.0401	-0.0003	0.0401	0.0078	0.0402	0.0901	0.0398	-0.0001	0.0398	95.0%	0.158
			1.0	19.25	0.0381	-0.0004	0.0381	0.0115	0.0383	0.0905	0.0379	-0.0001	0.0379	94.7%	0.150
			0.9	17.33	0.0362	-0.0003	0.0362	0.0084	0.0365	0.0912	0.0360	-0.0002	0.0360	95.0%	0.143
			0.8	15.40	0.0341	-0.0005	0.0341	0.0139	0.0346	0.0922	0.0339	-0.0003	0.0339	95.3%	0.136

Table 1: Simulation results with $d = 1$, $n = 1000$, and $B = 800$, over 3000 repeats

Table 2 presents analogous results in the two-dimensional setting ($d = 2$). The debiased estimator ($J = 1$, λ_1) again achieves the best RMSE, and the undersmoothed estimator ($J = 1$, λ_0) similarly displays the smallest bias/SD ratio. Coverage rates are generally worse than in Table 1, mostly due to the increased difficulty posed by the curse of dimensionality and a reduced effective sample size. Nonetheless, inference based on robust bias correction continues to exhibit a pronounced improvement in coverage when compared to standard non-debiased methods, and again shows a moderate increase in confidence interval widths.

	J	LS	LM	λ	RMSE	Bias	SD	Bias/SD	$\widehat{\text{SD}}$	$\hat{\sigma}^2$	ARMSE	ABias	ASD	CR	CIW
No debiasing	0	$\hat{\lambda}_0$	1.0	12.35	0.0805	-0.0646	0.0481	1.3432	0.0481	0.0989	0.0828	-0.0666	0.0479	71.1%	0.189
		λ_0	1.2	18.39	0.0758	-0.0310	0.0692	0.4481	0.0627	0.0882	0.0771	-0.0292	0.0714	88.2%	0.246
			1.1	16.85	0.0735	-0.0361	0.0640	0.5650	0.0593	0.0898	0.0741	-0.0347	0.0654	87.0%	0.233
			1.0	15.32	0.0726	-0.0427	0.0587	0.7280	0.0558	0.0919	0.0728	-0.0420	0.0595	84.9%	0.219
			0.9	13.79	0.0743	-0.0518	0.0532	0.9740	0.0520	0.0947	0.0746	-0.0519	0.0535	80.8%	0.204
			0.8	12.26	0.0796	-0.0637	0.0477	1.3346	0.0478	0.0985	0.0811	-0.0657	0.0476	71.6%	0.188
Debiasing	1	$\hat{\lambda}_1$	1.0	9.20	0.0726	-0.0144	0.0712	0.2020	0.0746	0.1277	0.0723	-0.0086	0.0691	95.1%	0.292
		λ_1	1.2	7.18	0.0584	-0.0217	0.0542	0.3999	0.0672	0.1490	0.0550	-0.0108	0.0540	96.3%	0.263
			1.1	6.58	0.0577	-0.0283	0.0503	0.5620	0.0644	0.1602	0.0518	-0.0154	0.0495	95.8%	0.252
			1.0	5.99	0.0596	-0.0381	0.0459	0.8299	0.0613	0.1733	0.0503	-0.0225	0.0450	94.0%	0.240
			0.9	5.39	0.0664	-0.0516	0.0418	1.2332	0.0578	0.1879	0.0530	-0.0343	0.0405	90.4%	0.227
			0.8	4.79	0.0797	-0.0704	0.0373	1.8873	0.0538	0.2044	0.0656	-0.0549	0.0360	79.6%	0.211
Robust BC	1	$\hat{\lambda}_0$	1.0	12.35	0.0889	-0.0053	0.0888	0.0598	0.0854	0.1047	0.0928	-0.0014	0.0927	94.9%	0.335
		λ_0	1.2	18.39	0.1208	-0.0032	0.1208	0.0265	0.0971	0.0925	0.1381	-0.0003	0.1381	89.4%	0.380
			1.1	16.85	0.1135	-0.0040	0.1134	0.0351	0.0953	0.0941	0.1266	-0.0004	0.1266	90.8%	0.373
			1.0	15.32	0.1056	-0.0039	0.1055	0.0367	0.0927	0.0964	0.1151	-0.0005	0.1151	92.6%	0.363
			0.9	13.79	0.0974	-0.0042	0.0973	0.0427	0.0893	0.0994	0.1036	-0.0008	0.1036	93.8%	0.350
			0.8	12.26	0.0883	-0.0047	0.0882	0.0532	0.0853	0.1041	0.0921	-0.0013	0.0921	94.9%	0.334

Table 2: Simulation results with $d = 2$, $n = 1000$, and $B = 800$, over 3000 repeats

6 Computational complexity and application to online learning

We discuss some computational aspects of (debiased) Mondrian random forests. We firstly consider the batch setting, where all of the data is available simultaneously, and secondly investigate the online regime, where data arrives sequentially and the model must be incrementally updated (Lakshminarayanan et al., 2014). Mondrian random forests have several properties that make them well suited for online learning: (i) in Mourtada et al. (2021) it was shown that some online Mondrian forest variants maintain statistical consistency, achieving the same asymptotic error rates as their batch counterparts under certain conditions; (ii) as we will demonstrate (Lemma 8), online Mondrian forest algorithms exploiting the Markov property of the Mondrian process are computationally efficient, therefore scaling to large streaming datasets; and (iii) the random nature of splits in Mondrian trees allows the forest to naturally adapt to changes in the underlying data distribution over time (concept drift), without requiring explicit drift detection or model reset mechanisms.

Some potential applications of online Mondrian forests with uncertainty quantification include real-time prediction and monitoring in industrial processes (Gomes et al., 2017), adaptive pricing and recommendation systems (Krauss et al., 2017; Li et al., 2018), online anomaly detection with confidence levels (Martindale et al., 2020), and streaming data analysis for the natural sciences (Abdulsalam et al., 2010).

The inference procedures developed in this paper extend to the online setting, allowing for uncertainty quantification in streaming data applications. However, care must be taken in situations where the underlying distribution may change over time, or where validity of the inferential procedures is required to hold uniformly over the data arrival times. Developing rigorous statistical inference tools for online Mondrian forests in those more complicated time-dependent regimes is an interesting direction for future work.

The core of our computational approaches for batch and online learning comprise several main ideas; these enable substantial improvements over naive algorithms based on the equations presented in previous sections. The first of these is to keep track of which data points are "local" to the evaluation point x, according to the forest $(T_{br}(x) : 1 \leq b \leq B, 0 \leq r \leq J)$. Define the *union cell* $U(x) \subseteq [0,1]^d$ and *active indices* $I(x) \subseteq \{1, \ldots, n\}$ by

$$U(x) = \prod_{j=1}^{d} \bigcup_{b=1}^{B} \bigcup_{r=0}^{J} T_{br}(x)_j \qquad \text{and} \qquad I(x) = \{1 \leq i \leq n : X_i \in U(x)\} \tag{10}$$

respectively, noting that any data point contributing to $\hat{\mu}_{\mathrm{d}}(x)$ or $\hat{\Sigma}_{\mathrm{d}}(x)$ satisfies $X_i \in U(x)$ and $i \in I(x)$. As the lifetime parameter λ grows, the volume of $U(x)$ and the proportion of contributing samples $|I(x)|/n$ both converge to zero in expectation, lowering the effective sample size and significantly decreasing the amount of computation necessary. Further, $U(x)$ can be efficiently computed with a divide-and-conquer approach whenever multiple parallel processors are available.

The second main idea is to observe that the estimators $\hat{\mu}_{\mathrm{d}}(x)$ and $\hat{\sigma}^2(x)$ can be expressed as ratios of sums. More precisely, firstly define

$$N_{br}(x) = \sum_{i \in I(x)} \mathbb{I}\{X_i \in T_{br}(x)\}, \qquad S_{br}(x) = \sum_{i \in I(x)} Y_i \, \mathbb{I}\{X_i \in T_{br}(x)\},$$
$$V_{br}(x) = \sum_{i \in I(x)} Y_i^2 \, \mathbb{I}\{X_i \in T_{br}(x)\}, \tag{11}$$

which are efficient to update as new samples arrive; furthermore, they can be computed separately for each b and r in parallel. Then one can write

$$\hat{\mu}_{\mathrm{d}}(x) = \sum_{r=0}^{J} \omega_r \frac{1}{B} \sum_{b=1}^{B} \frac{S_{br}(x)}{N_{br}(x)} \qquad \text{and} \qquad \hat{\sigma}^2(x) = \frac{1}{B} \sum_{b=1}^{B} \frac{V_{b0}(x)}{N_{b0}(x)}. \tag{12}$$

The third observation is that the estimators depend on the trees only through the cell $T_{br}(x)$. Since Mourtada et al. (2020) characterize the exact distribution of this quantity, it can be sampled without needing to grow an entire Mondrian tree. Further, the memoryless property of the exponential distribution (and thus also of the Mondrian process) means that in the online setting, only a small fraction of the cells typically need to be updated.

The fourth and final concept is to avoid fitting the relatively computationally expensive $\hat{\Sigma}_{\mathrm{d}}(x)$ too often. This estimator does not readily admit a "ratio of sums" formulation, and hence is not efficient to update incrementally. Our recommendation is to instead only update this term after $K \geq 1$ new data points have arrived on average. Note however that using the active indices $I(x)$ still permits an improvement over the naive approach, since

$$\hat{\Sigma}_{\mathrm{d}}(x) = \hat{\sigma}^2(x) \frac{n}{\lambda^d} \sum_{i \in I(x)} \left(\sum_{r=0}^{J} \omega_r \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{I}\{X_i \in T_{br}(x)\}}{N_{br}(x)} \right)^2. \tag{13}$$

Before discussing the online learning setting in more detail, we present our efficient procedure for batch estimation and inference in Algorithm 1.

Algorithm 1: Batch learning with Mondrian random forests

Input: Data (X_i, Y_i) for $1 \leq i \leq n$, forest size $B \geq 1$, debiasing order $J \geq 0$.

1. Select λ using one of the methods from Section 5.1.
2. Construct the union cell $U(x)$ and active indices $I(x)$ as in (10).
3. Calculate $N_{br}(x)$, $S_{br}(x)$ and $V_{br}(x)$ for each $1 \leq b \leq B$ and $0 \leq r \leq J$ as in (11).
4. Compute $\hat{\mu}_{\mathrm{d}}(x)$ and $\hat{\sigma}^2(x)$ with (12).
5. Calculate $\hat{\Sigma}_{\mathrm{d}}(x)$ and $\mathrm{CI}_{\mathrm{d}}(x)$ using (13) and (9) respectively.

The following lemma bounds the average case time complexity of our batch learning procedure (Algorithm 1), under the same assumptions made throughout the paper.

Lemma 7 (Computational complexity of batch learning)**.** *Suppose Assumptions 1 and 2 hold. Then the average case time complexity of Algorithm 1 is*

$$\mathbb{E}\left[\mathcal{T}_{\mathrm{b}}\right] \lesssim d(J+1)\big(nd(J+1) + B\big) + \frac{nBd(J+1)\log(2B(J+1))^d}{\lambda^d}.$$

We now turn to the online learning setting, making the following assumptions. Firstly, suppose that a (debiased) Mondrian random forest with B trees has already been fitted to a data set with n samples, using a lifetime of λ, and that this original data set is still available. Assume that the union cell $U(x)$, the index set $I(x)$, and the point estimates $\hat{\mu}_{\mathrm{d}}(x)$ and $\hat{\Sigma}_{\mathrm{d}}(x)$ have been computed, as well as the trees $T_{br}(x)$ and the quantities $S_{br}(x)$, $N_{br}(x)$ and $V_{br}(x)$ for $1 \leq b \leq B$ and $0 \leq r \leq J$. A new data set with k samples then arrives, where $1 \leq k \leq n$, and we must produce updated estimates $\hat{\mu}_{\mathrm{d}}^*(x)$, $\hat{\Sigma}_{\mathrm{d}}^*(x)$ and $\mathrm{CI}_{\mathrm{d}}^*(x)$ based on all $n+k$ samples. Our randomized procedure for doing this is described below, using a star to indicate updated quantities, and summarized in Algorithm 2.

The new sample size is $n^* = n+k$, so the first step is to update B. As recommended in Section 5, we take $B \asymp n^\xi$ for some $\xi \in (0,1)$; therefore set $B^* = \lfloor (n+k)^\xi B / n^\xi \rfloor$. Next, we update the lifetime parameter λ. To avoid excessive computation, we suggest the following: with probability $1 \wedge (k/K)$, use the methods from Section 5.1 to compute a new lifetime parameter $\lambda^* \geq \lambda$ using all of the data. Otherwise, note that $\lambda \asymp n^\zeta$ for some $\zeta \in (0, 1/d)$ (for example $\zeta = 1/(d+2\beta)$ under the conditions of Theorem 4) and set $\lambda^* = (n+k)^\zeta \lambda / n^\zeta$. Next, to update the trees $T_{br}(x)$, sample E_{brj1} and E_{brj2} i.i.d. Exp(1), and set

$$T^*_{br}(x)^-_j = T_{br}(x)^-_j \vee \left(x_j - \frac{E_{brj1}}{\lambda^* - \lambda} \right), \qquad T^*_{br}(x)^+_j = T_{br}(x)^+_j \wedge \left(x_j + \frac{E_{brj2}}{\lambda^* - \lambda} \right). \tag{14}$$

Since $B^* \geq B$, we also generate new trees $T^*_{br}(x)$ for $B+1 \leq b \leq B^*$ and $0 \leq r \leq J$ using

$$T^*_{br}(x)^-_j = 0 \vee \left(x_j - \frac{E_{brj1}}{a_r \lambda} \right), \qquad T^*_{br}(x)^+_j = 1 \wedge \left(x_j + \frac{E_{brj2}}{a_r \lambda} \right). \tag{15}$$

Computing $U^*(x)$ is simple, applying (10) to $T^*_{br}(x)$. To update $I(x)$, set

$$I^*(x) = \begin{cases} \{i \in I(x) \cup \{n+1, \ldots, n+k\} : X_i \in U^*(x)\} & \text{if } U^*(x) \subseteq U(x), \\ \{1 \leq i \leq n+k : X_i \in U^*(x)\} & \text{otherwise.} \end{cases} \tag{16}$$

For $N_{br}(x)$, and analogously for $S_{br}(x)$ and $V_{br}(x)$, apply the following method:

$$N^*_{br}(x) = \begin{cases} N_{br}(x) + \sum_{i \in V^*(x), i > n} \mathbb{I}\{X_i \in T_b(x)\} & \text{if } b \leq B \text{ and } T_{br}(x) = T^*_{br}(x) \\ \sum_{i \in I^*(x)} \mathbb{I}\{X_i \in T^*_{br}(x)\} & \text{otherwise.} \end{cases} \tag{17}$$

Finally, $\hat{\mu}^*_{\mathrm{d}}(x)$ and $\hat{\sigma}^{2*}(x)$ are computed using (12). With probability $1 \wedge (k/K)$, recalculate $\hat{\Sigma}^*_{\mathrm{d}}(x)$ with (13); otherwise set $\hat{\Sigma}^*_{\mathrm{d}}(x) = \hat{\Sigma}_{\mathrm{d}}(x)$. The confidence interval $\mathrm{CI}^*_{\mathrm{d}}(x)$ can then be constructed with (9). The following algorithm summarizes our online methodology.

Algorithm 2: Online learning with Mondrian random forests

Input: Data (X_i, Y_i) for $1 \leq i \leq n$, forest size $B \geq 1$, debiasing order $J \geq 0$, lifetime λ, forest exponent $\xi \in (0,1)$, lifetime exponent $\zeta \in (0, 1/d)$, active region $U(x)$, active indices $I(x)$, trees $T_{br}(x)$ and $N_{br}(x)$, $S_{br}(x)$, $V_{br}(x)$ for $1 \leq b \leq B$ and $0 \leq r \leq J$, new data (X_i, Y_i) for $n+1 \leq i \leq n+k$, recalculation gap $K \geq 1$.

1. Get the updated number of trees $B^* = \lfloor (n+k)^\xi B / n^\xi \rfloor$.
2. With probability $1 \wedge (k/K)$, select λ^* as in Section 5.1; otherwise, set $\lambda^* = (n+k)^\zeta \lambda / n^\zeta$.
3. Generate the incrementally updated forest $T^*_{br}(x)$ as in (14) and (15).
4. Construct the updated union cell $U^*(x)$ and active indices $I^*(x)$ as in (16).
5. Calculate $N^*_{br}(x)$, $S^*_{br}(x)$, and $V^*_{br}(x)$ as in (17) and derive $\hat{\mu}^*_{\mathrm{d}}(x)$ and $\sigma^{2*}(x)$ from (11).
6. With probability $1 \wedge (k/K)$, recalculate $\hat{\Sigma}^*_{\mathrm{d}}(x)$ using (13); otherwise, set $\hat{\Sigma}^*_{\mathrm{d}}(x) = \hat{\Sigma}_{\mathrm{d}}(x)$.
7. Compute $\mathrm{CI}^*_{\mathrm{d}}(x)$ using (9) with $\hat{\mu}^*_{\mathrm{d}}(x)$ and $\hat{\Sigma}^*_{\mathrm{d}}(x)$.

Lemma 8 bounds the average case time complexity of our online computational procedure presented in Algorithm 2.

Lemma 8 (Computational complexity of online learning)**.** *Suppose Assumptions 1 and 2 hold. Then the average case time complexity of Algorithm 2 is*

$$\mathbb{E}[\mathcal{T}_{\mathrm{o}}] \lesssim d(J+1) \left(\frac{knd(J+1)}{K} + kd + B \right) + \frac{d(J+1) \log(2B(J+1))^d}{\lambda^d} \left(n + Bk + \frac{nB}{K} \right).$$

Lemma 7 already demonstrated that Algorithm 1 is more efficient than the naive approach of computing $\hat{\mu}_{\mathrm{d}}(x)$ and $\hat{\Sigma}_{\mathrm{d}}(x)$ directly with (7) and (8), respectively, which each have a time complexity of $n(J+1)B$. The reason for this is that by first constructing the active indices $I(x)$, we avoid iterating over the entire sample for each tree in Algorithm 1. Lemma 8 formalizes the improvement achieved by Algorithm 2 in online settings, relative to the batch estimation approach of Algorithm 1. Most importantly, the terms involving the product nB are reduced to $n + Bk + nB/K$, offering a substantial speed-up in large forests when the new sample size k is much smaller than that of the existing data n, and when K is large to avoid regularly estimating the lifetime λ and variance $\hat{\Sigma}_{\mathrm{d}}(x)$.

7 Conclusion

We presented a Berry–Esseen theorem under mild conditions for the Mondrian random forest estimator, and showed how it can be used to perform statistical inference on an unknown nonparametric regression function. We introduced debiased versions of Mondrian random forests, exploiting higher-order smoothness, and demonstrated their advantages for statistical inference and their minimax optimality properties. We discussed tuning parameter selection, enabling fully feasible and practical estimation and inference procedures, and demonstrated the empirical performance of our proposed methodology. Finally, we developed efficient algorithms for batch and online settings.

There are several potential avenues for future work on inference with Mondrian random forests. The development of data-adaptive partitioning schemes is one such important direction, and could be implemented perhaps by allowing the lifetime parameter λ to vary across different covariates, yielding the d-dimensional parameter $(\lambda_1, \ldots, \lambda_d)$. One approach to designing such methodology might involve adapting sparse, greedy algorithms for non-parametric regression, similar to those described by Lafferty and Wasserman (2008), to the context of axis-aligned partitioning estimators. Specifically, by examining how changes in each λ_j affect the Mondrian forest estimator, e.g., via an estimate of $\frac{\partial}{\partial \lambda_j} \mathbb{E}[\hat{\mu}(x)]$, these parameters can be dynamically adjusted to more effectively learn low-dimensional structure in the regression function. Alternatively, one might formulate a Goldenshluger–Lepski-type procedure (Goldenshluger and Lepski, 2008) for multiple tuning parameter selection. Another potential line of research would consist of proposing further strategies for debiasing Mondrian random forests (and related estimators); an approach based on within-cell local polynomial smoothing, for example, may serve to eliminate both design bias and boundary bias, as well as allowing for less restrictive conditions on the covariate density function and the regression function.

8 Acknowledgments and funding

The authors would like to thank Rajita Chandak, Jianqing Fan, Henry Reeve, Richard Samworth, and Sanjeev Kulkarni for comments and insightful discussions. Klusowski would like to thank the participants at the 11th Bernoulli-IMS World Congress in Probability and Statistics in Bochum, Germany for valuable feedback. Underwood would like to thank the attendees at the University of Pittsburgh, University of Illinois Urbana-Champaign, and University of Michigan Statistics Seminars, as well as the members of the University of Cambridge Statistics Laboratory.

Cattaneo was supported in part by the National Science Foundation through DMS-2210561 and SES-2241575. Klusowski was supported in part by the National Science Foundation through CAREER DMS-2239448, DMS-2054808, and HDR TRIPODS CCF-1934924.

A Proofs and technical results

In this section we present the full proofs of all our results, and also state some useful technical preliminary and intermediate lemmas. We use the following simplified notation for convenience, whenever it is appropriate: write $\mathbb{I}_{ib}(x) = \mathbb{I}\{X_i \in T_b(x)\}$ and $N_b(x) = \sum_{i=1}^n \mathbb{I}_{ib}(x)$, as well as $\mathbb{I}_b(x) = \mathbb{I}\{N_b(x) \geq 1\}$. We use C to denote a positive constant whose value may change from line to line, and write $a_n = O(b_n)$ for $a_n \lesssim b_n$. We begin with an overview of the main proof strategies and a discussion of the challenges involved in Section A.1. We then give some preliminary lemmas in Section A.2, and present the proofs for Section 3 (including Lemma 1, Lemma 2, Theorem 1, Theorem 2, Lemma 3, and Theorem 3) in Section A.3; the proofs for Section 4 (including Lemma 4, Lemma 5, Theorem 4, Theorem 5, Lemma 6, and Theorem 6) in Section A.4; and the proofs for Section 6 (including Lemma 7 and Lemma 8) in Section A.5.

A.1 Overview of proof strategies

This section provides some insight into the general approach we use to establish our main results. We focus on the technical innovations forming the core of our arguments, and refer the reader to the upcoming sections for full proofs.

A.1.1 Preliminary technical results

The starting point for our proofs is a result characterizing the distribution of the shape of a Mondrian cell $T(x)$. This property is a consequence of the fact that the restriction of a Mondrian process to a subcell remains a Mondrian process (Mourtada et al., 2020). We have

$$|T(x)_j| = \left(\frac{E_{j1}}{\lambda} \wedge x_j\right) + \left(\frac{E_{j2}}{\lambda} \wedge (1 - x_j)\right)$$

for all $1 \leq j \leq d$, recalling that $T(x)_j$ is the side of the cell $T(x)$ aligned with axis j, and where E_{j1} and E_{j2} are mutually independent Exp(1) random variables. Our assumptions that $x \in (0,1)$ and $\lambda \to \infty$ mean that the "boundary terms" x_j and $1 - x_j$ are eventually ignorable and so $|T(x)_j| = (E_{j1} + E_{j2})/\lambda$ with high probability. Controlling the size of the largest cell in the forest containing x is now straightforward with a union bound, giving

$$\max_{1 \leq b \leq B} \max_{1 \leq j \leq d} |T_b(x)_j| \lesssim_{\mathbb{P}} \frac{\log B}{\lambda}.$$

This shows that, up to logarithmic terms, none of the cells in the forest at x are significantly larger than average, ensuring that the Mondrian random forest estimator is "localized" around x on the scale of $1/\lambda$, an important property for our bias characterization.

Having provided upper bounds for the sizes of Mondrian cells, we also must establish some lower bounds in order to ensure a sufficient effective sample size and to quantify the "small cells" phenomenon mentioned previously. The first step towards this is to bound the first two moments of the truncated inverse Mondrian cell volume; we show that

$$\mathbb{E}\left[1 \wedge \frac{1}{n|T(x)|}\right] \asymp \frac{\lambda^d}{n} \qquad \text{and} \qquad \mathbb{E}\left[1 \wedge \frac{1}{n^2|T(x)|^2}\right] \asymp \frac{\lambda^{2d}(\log n)^d}{n^2}.$$

These bounds are computed using the exact distribution of $|T(x)|$. Note that $\mathbb{E}\left[1/|T(x)|^2\right] = \infty$ because $1/(E_{j1} + E_{j2})$ has only $2 - \delta$ finite moments, so the truncation is crucial here. Since we have "almost two" moments, this truncation is at the expense of only a logarithmic term. Nonetheless, third and higher truncated moments will not enjoy such tight bounds, demonstrating both the fragility of this result and the inadequacy of tools such as the Lyapunov central limit theorem which require $2+\delta$ marginal moments.

To conclude this investigation into the "small cell" phenomenon, we apply the previous bounds to ensure that the empirical effective sample sizes $N_b(x) = \sum_{i=1}^n \mathbb{I}\{X_i \in T_b(x)\}$ are approximately of the order n/λ^d in an appropriate sense; we demonstrate that

$$\mathbb{E}\left[\frac{\mathbb{I}\{N_b(x) \geq 1\}}{N_b(x)}\right] \lesssim \frac{\lambda^d}{n} \qquad \text{and} \qquad \mathbb{E}\left[\frac{\mathbb{I}\{N_b(x) \geq 1\}}{N_b(x)^2}\right] \lesssim \frac{\lambda^{2d}(\log n)^d}{n^2},$$

as well as "mixed" bounds $\mathbb{E}\left[\mathbb{I}\{N_b(x) \geq 1\}\mathbb{I}\{N_{b'}(x) \geq 1\}/(N_b(x)N_{b'}(x))\right] \lesssim \lambda^{2d}/n^2$ when $b \neq b'$, which arise from covariance terms across multiple trees. The proof of this result is involved and technical, and proceeds by induction. The idea is to construct a class of subcells by taking all possible intersections of the cells in T_b and $T_{b'}$ (we show two trees here for clarity; there may be more) and noting that each $N_b(x)$ is the sum of the number of points in each such "refined cell" intersected with $T_b(x)$. We then swap out each refined cell one at a time and replace the number of data points it contains with its volume multiplied by $nf(x)$, showing that the expectation on the left hand side does not increase too much using a moment bound for inverse binomial random variables based on Bernstein's inequality. By induction and independence of the trees, eventually the problem is reduced to computing moments of truncated inverse Mondrian cell volumes, as above.

A.1.2 Bias characterization

Our first substantial result is the bias characterization given as Lemma 1, in which we precisely characterize the probability limit of the conditional bias

$$\mathbb{E}\left[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\right] - \mu(x) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}\big(\mu(X_i) - \mu(x)\big)\frac{\mathbb{I}\{X_i \in T_b(x)\}}{N_b(x)}.$$

The first step in this proof is to pass to the "infinite forest" limit by taking an expectation conditional on $\mathbf{X}$, or equivalently marginalizing over $\mathbf{T}$, applying the conditional Markov inequality to see

$$\big|\mathbb{E}\left[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\right] - \mathbb{E}\left[\hat{\mu}(x) \mid \mathbf{X}\right]\big| \lesssim_{\mathbb{P}} \frac{1}{\lambda^{1\wedge\beta_\mu}\sqrt{B}}.$$

While this may seem a crude approximation, it is already known that fixed-size Mondrian forests have suboptimal bias properties when compared to forests with a diverging number of trees. In fact, when $\beta \geq 1$, the error $1/\big(\lambda^{1\wedge\beta}\sqrt{B}\big)$ exactly accounts for the first-order bias of individual Mondrian trees (Mourtada et al., 2020).

Next we show that $\mathbb{E}\left[\hat{\mu}(x) \mid \mathbf{X}\right]$ converges in probability to its expectation, using the Efron–Stein theorem to handle this non-linear function of the i.i.d. variables X_i. The important insight here is that replacing a sample X_i with an independent copy $\tilde{X}_i$ can change the value of $N_b(x)$ by at most one. Further, this can happen only on the event $\{X_i \in T_b(x)\} \cup \{\tilde{X}_i \in T_b(x)\}$, which occurs with probability on the order $1/\lambda^d$ (the expected cell volume) for each tree $1 \leq b \leq B$. The Hölder property of μ and the upper bound on the maximum cell size then give $|\mu(X_i) - \mu(x)| \lesssim \max_{1\leq j\leq d}|T_b(x)_j|^{1\wedge\beta_\mu} \lesssim_{\mathbb{P}} 1/\lambda^{1\wedge\beta_\mu}$ whenever $X_i \in T_b(x)$, so we combine this with moment bounds for the denominator $N_b(x)$ to see

$$\left|\mathbb{E}\left[\hat{\mu}(x) \mid \mathbf{X}\right] - \mathbb{E}\left[\hat{\mu}(x)\right]\right| \lesssim_{\mathbb{P}} \frac{1}{\lambda^{1\wedge\beta_\mu}}\sqrt{\frac{\lambda^d}{n}}.$$

The next step is to approximate the resulting non-random bias $\mathbb{E}[\hat{\mu}(x)] - \mu(x)$ as a polynomial in $1/\lambda$. To this end, we firstly apply a concentration-type result for the binomial distribution to deduce that

$$\mathbb{E}\left[\frac{\mathbb{I}\{N_b(x) \geq 1\}}{N_b(x)} \;\Big|\; \mathbf{T}\right] \approx \frac{1}{n \int_{T_b(x)} f(s)\,\mathrm{d}s}$$

in an appropriate sense, and hence, by conditioning on $\mathbf{T}$ and $\mathbf{X}$ without X_i,

$$\mathbb{E}[\hat{\mu}(x)] - \mu(x) \approx \mathbb{E}\left[\frac{\int_{T_b(x)} (\mu(s) - \mu(x)) f(s)\,\mathrm{d}s}{\int_{T_b(x)} f(s)\,\mathrm{d}s}\right]. \tag{18}$$

Next we apply the multivariate version of Taylor's theorem to the integrands in both the numerator and the denominator in (18), and then apply the Maclaurin series of $1/(1+x)$ and the multinomial theorem to recover a single polynomial in $1/\lambda$. The error term is on the order of $1/\lambda^\beta$ and depends on the smoothness of μ and f, and the polynomial coefficients are given by various expectations involving exponential random variables. The final step is to verify using symmetry of Mondrian cells that all the odd monomial coefficients are zero, and to calculate some explicit examples of the form of the limiting bias.

A.1.3 Central limit theorem

To prove our second main result (Theorem 2), we apply a version of the Berry–Esseen theorem for i.n.i.d. random variables, conditional on $(\mathbf{X}, \mathbf{T})$, which only requires $2 + \delta$ moments. Define the variables

$$S_i(x) = \sqrt{\frac{n}{\lambda^d}} \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{I}\{X_i \in T_b(x)\}\varepsilon_i}{N_b(x)},$$

which are independent and zero-mean given $(\mathbf{X}, \mathbf{T})$, and further satisfy

$$\sqrt{\frac{n}{\lambda^d}} \big(\hat{\mu}(x) - \mathbb{E}[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}]\big) = \sum_{i=1}^{n} S_i(x).$$

Thus by Petrov (1995, Theorem 5.7), conditional on $(\mathbf{X}, \mathbf{T})$, taking a marginal expectation,

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\tilde{\Sigma}(x)^{-\frac{1}{2}} \sum_{i=1}^{n} S_i \leq t \right) - \Phi(t) \right| \lesssim \mathbb{E}\left[1 \wedge \left(\tilde{\Sigma}(x)^{-\frac{2+(\delta \wedge 1)}{2}} \sum_{i=1}^{n} \mathbb{E}\Big[|S_i|^{2+(\delta \wedge 1)} \mid \mathbf{X}, \mathbf{T} \Big] \right) \right].$$

Bounding the right-hand side now reduces to establishing properties of $\tilde{\Sigma}(x)$ and its large-sample limit $\Sigma(x)$. To this end, we again use the Efron–Stein theorem to bound $\mathrm{Var}\big[\tilde{\Sigma}(x)\big]$ and then apply a careful sequence of approximations to control $\mathbb{E}\big[\tilde{\Sigma}(x)\big] - \Sigma(x)$. The final task is to calculate the limiting variance $\Sigma(x)$. It is a straightforward but tedious exercise to verify that each denominator $N_b(x)$ can be replaced by $n f(x) |T_b(x)|$, yielding

$$\Sigma(x) = \frac{\sigma^2(x)}{f(x)} \lim_{\lambda \to \infty} \frac{1}{\lambda^d} \mathbb{E}\left[\frac{|T_b(x) \cap T_{b'}(x)|}{|T_b(x)|\,|T_{b'}(x)|} \right] = \frac{\sigma^2(x)}{f(x)} \left(\mathbb{E}\left[\frac{(E_1 \wedge E_1') + (E_2 \wedge E_2')}{(E_1 + E_2)(E_1' + E_2')} \right] \right)^d,$$

where E_1, E_2, E_1', and E_2' are independent $\mathrm{Exp}(1)$, by the cell shape distribution and independence of the trees. This final expectation is calculated by integration, using various incomplete gamma function identities.

A.1.4 Confidence intervals

While Theorem 2 gives a distributional approximation for the infeasible t-statistic, in order to construct confidence intervals we must instead approximate the corresponding feasible t-statistic. To do this, first observe that if τ and $\hat{\tau}$ are real-valued random variables and $\varepsilon > 0$, then the following anti-concentration inequality holds:

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\big(\hat{\tau} \leq t\big) - \Phi(t) \right| \leq \sup_{t \in \mathbb{R}} \left| \mathbb{P}\big(\tau \leq t\big) - \Phi(t) \right| + \varepsilon \sqrt{2/\pi} + \mathbb{P}\big(|\hat{\tau} - \tau| > \varepsilon\big).$$

We apply this result to

$$\hat{\tau} = \sqrt{\frac{n}{\lambda^d}} \left(\frac{\hat{\mu}(x) - \mu(x)}{\sqrt{\hat{\Sigma}(x)}} - \frac{\mathbb{E}\left[\hat{\mu}(x)\right] - \mu(x)}{\sqrt{\Sigma(x)}} \right) \qquad \text{and} \qquad \tau = \sqrt{\frac{n}{\lambda^d}} \frac{\hat{\mu} - \mathbb{E}\left[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\right]}{\sqrt{\tilde{\Sigma}(x)}},$$

bounding $\mathbb{P}\big(|\hat{\tau} - \tau| > \varepsilon\big)$ using our established results on $\hat{\mu}(x)$, $\tilde{\Sigma}(x)$ and $\hat{\Sigma}(x)$. Exploiting symmetry of the resulting confidence interval permits a quadratic dependence of the coverage error on the bias.

A.2 Preliminary lemmas

We begin by bounding the maximum size of any cell in a Mondrian forest containing x. This result is used regularly throughout many of our other proofs, and captures the "localizing" behavior of the Mondrian random forest estimator, showing that Mondrian cells have side lengths at most on the order of $1/\lambda$.

Lemma 9 (Upper bound on the largest cell in a Mondrian forest). *Let $T_1, \ldots, T_B \sim \mathcal{M}\big([0,1]^d, \lambda\big)$ and take $x \in (0,1)^d$. Then for all $t > 0$*

$$\mathbb{P}\left(\max_{1 \leq b \leq B} \max_{1 \leq j \leq d} |T_b(x)_j| \geq \frac{t}{\lambda} \right) \leq 2dBe^{-t/2}.$$

Proof (Lemma 9)**.** We use the explicit distribution of the shape of Mondrian cells given by Mourtada et al. (2020, Proposition 1). In particular, we have $|T_b(x)_j| = \left(\frac{E_{bj1}}{\lambda} \wedge x_j\right) + \left(\frac{E_{bj2}}{\lambda} \wedge (1 - x_j)\right)$ where E_{bj1} and E_{bj2} are independent $\mathrm{Exp}(1)$ random variables for $1 \leq b \leq B$ and $1 \leq j \leq d$. Thus $|T_b(x)_j| \leq \frac{E_{bj1} + E_{bj2}}{\lambda}$ and so by a union bound

$$\begin{aligned}
\mathbb{P}\left(\max_{1 \leq b \leq B} \max_{1 \leq j \leq d} |T_b(x)_j| \geq \frac{t}{\lambda} \right) &\leq \mathbb{P}\left(\max_{1 \leq b \leq B} \max_{1 \leq j \leq d} (E_{bj1} \vee E_{bj2}) \geq \frac{t}{2} \right) \\
&\leq 2dB \, \mathbb{P}\left(E_{bj1} \geq \frac{t}{2} \right) \leq 2dBe^{-t/2}.
\end{aligned}$$

□

The next result is another "localization" result, this time showing that the union over the forest of the cells $T_b(x)$ containing x do not contain "too many" samples X_i. In other words, the Mondrian random forest estimator fitted at x should only depend on n/λ^d (the effective sample size) data points up to logarithmic terms.

Lemma 10 (Upper bound on the number of active data points)**.** *Suppose that Assumptions 1 and 2 hold and define $N_\cup(x) = \sum_{i=1}^n \mathbb{I}\left\{X_i \in \bigcup_{b=1}^B T_b(x)\right\}$. Then for $t > 0$ and $n \geq \lambda^d$, with $\|f\|_\infty = \sup_{x \in [0,1]^d} f(x)$,*

$$\mathbb{P}\left(N_\cup(x) > t^{d+1} \frac{n}{\lambda^d} \|f\|_\infty \right) \leq 4dBe^{-t/4}.$$

Proof (Lemma 10)**.** Note that

$$N_\cup(x) \sim \mathrm{Bin}\left(n, \int_{\bigcup_{b=1}^B T_b(x)} f(s)\,\mathrm{d}s\right) \leq \mathrm{Bin}\left(n, 2^d \max_{1\leq b\leq B} \max_{1\leq j\leq d} |T_b(x)_j|^d \|f\|_\infty\right)$$

conditionally on $\mathbf{T}$. If $N \sim \mathrm{Bin}(n,p)$ then, by Bernstein's inequality, $\mathbb{P}\left(N \geq (1+t)np\right) \leq \exp\left(-\frac{t^2n^2p^2/2}{np(1-p)+tnp/3}\right) \leq \exp\left(-\frac{3t^2np}{6+2t}\right)$. Thus for $t \geq 2$,

$$\mathbb{P}\left(N_\cup(x) > (1+t)n\frac{2^dt^d}{\lambda^d}\|f\|_\infty \;\middle|\; \max_{1\leq b\leq B}\max_{1\leq j\leq d}|T_j(x)| \leq \frac{t}{\lambda}\right) \leq \exp\left(-\frac{2^dt^dn}{\lambda^d}\right).$$

By Lemma 9, $\mathbb{P}\left(\max_{1\leq b\leq B}\max_{1\leq j\leq d}|T_j(x)| > \frac{t}{\lambda}\right) \leq 2dBe^{-t/2}$. Hence

$$\begin{aligned}
&\mathbb{P}\left(N_\cup(x) > 2^{d+1}t^{d+1}\frac{n}{\lambda^d}\|f\|_\infty\right) \\
&\quad\leq \mathbb{P}\left(N_\cup(x) > 2tn\frac{2^dt^d}{\lambda^d}\|f\|_\infty \;\middle|\; \max_{1\leq b\leq B}\max_{1\leq j\leq d}|T_j(x)| \leq \frac{t}{\lambda}\right) + \mathbb{P}\left(\max_{1\leq b\leq B}\max_{1\leq j\leq d}|T_j(x)| > \frac{t}{\lambda}\right) \\
&\quad\leq \exp\left(-\frac{2^dt^dn}{\lambda^d}\right) + 2dBe^{-t/2}.
\end{aligned}$$

Noting the result is trivial for $t < 2$ and replacing t by $t/2$ gives that for $n \geq \lambda^d$,

$$\mathbb{P}\left(N_\cup(x) > t^{d+1}\frac{n}{\lambda^d}\|f\|_\infty\right) \leq 4dBe^{-t/4}.$$

□

Next we give a series of results culminating in a generalized moment bound for the denominator appearing in the Mondrian random forest estimator. We begin by providing a moment bound for the truncated inverse binomial distribution, which will be useful for controlling $\frac{\mathbb{I}_b(x)}{N_b(x)} \leq 1 \wedge \frac{1}{N_b(x)}$ because conditional on T_b we have $N_b(x) \sim \mathrm{Bin}\left(n, \int_{T_b(x)} f(s)\,\mathrm{d}s\right)$. Our constants could be suboptimal but they are sufficient for our applications.

Lemma 11 (An inverse moment bound for the binomial distribution)**.** *For $n \geq 1$ and $p \in [0,1]$, let $N \sim \mathrm{Bin}(n,p)$. Take $a_1,\ldots,a_k \geq 0$ and $l_1,\ldots,l_k \geq 1$. Then with $L = \sum_{j=1}^k l_j$,*

$$\mathbb{E}\left[\prod_{j=1}^k\left(1 \wedge \frac{1}{N+a_j}\right)^{l_j}\right] \leq (9L)^{2L}\prod_{j=1}^k\left(1 \wedge \frac{1}{np+a_j}\right)^{l_j}.$$

Proof (Lemma 11)**.** By Bernstein's inequality, $\mathbb{P}\left(N \leq np - t\right) \leq \exp\left(-\frac{3t^2}{6np+2t}\right)$. Therefore we have

$\mathbb{P}(N \leq np/4) \leq \exp\left(-\frac{27n^2p^2/16}{6np+3np/2}\right) = e^{-9np/40}$. Partitioning by this event gives

$$
\begin{aligned}
\mathbb{E}\left[\prod_{j=1}^{k}\left(1 \wedge \frac{1}{N+a_j}\right)^{l_j}\right] &\leq e^{-9np/40}\prod_{j=1}^{k}\frac{1}{1 \vee a_j^{l_j}} + \prod_{j=1}^{k}\frac{1}{1 \vee (\frac{np}{4}+a_j)^{l_j}} \\
&\leq \prod_{j=1}^{k}\frac{1}{1 \vee \left(\frac{9np}{40kl_j}+a_j\right)^{l_j}} + \prod_{j=1}^{k}\frac{1}{1 \vee (\frac{np}{4}+a_j)^{l_j}} \\
&\leq 2\prod_{j=1}^{k}\frac{1}{1 \vee \left(\frac{9np}{40kl_j}+a_j\right)^{l_j}} \leq 2\prod_{j=1}^{k}\frac{(40kl_j/9)^{l_j}}{1 \vee (np+a_j)^{l_j}} \\
&\leq (9L)^{2L}\prod_{j=1}^{k}\left(1 \wedge \frac{1}{np+a_j}\right)^{l_j}.
\end{aligned}
$$

□

Our next result is probably the most technically involved in the paper, allowing one to bound moments of (products of) $\frac{\mathbb{I}_b(x)}{N_b(x)}$ by the corresponding moments of (products of) $\frac{1}{n|T_b(x)|}$, again based on the heuristic that $N_b(x)$ is conditionally binomial so concentrates around its conditional expectation $n\int_{T_b(x)} f(x)\,\mathrm{d}s \asymp n|T_b(x)|$. By independence of the trees, the latter expected products then factorize since the dependence on the data X_i has been eliminated. The proof is complicated, and relies on the following induction procedure. First we consider the common refinement consisting of the subcells $\mathcal{R}$ generated by all possible intersections of $T_b(x)$ over the selected trees (say $T_b(x), T_{b'}(x), T_{b''}(x)$ though there could be arbitrarily many). Note that $N_b(x)$ is the sum of the number of samples X_i in each such subcell in $\mathcal{R}$. We then apply Lemma 11 repeatedly to each subcell in $\mathcal{R}$ in turn, replacing the number of samples X_i in that subcell with its volume multiplied by the sample size n, and controlling the error incurred at each step. We record the subcells which have been "checked" in this manner using the class $\mathcal{D} \subseteq \mathcal{R}$ and proceed by finite induction, beginning with $\mathcal{D} = \emptyset$ and ending at $\mathcal{D} = \mathcal{R}$.

Lemma 12 (Generalized moment bound for Mondrian random forest denominators)**.** *Suppose Assumptions 1 and 2 hold. Let $T_b \sim \mathcal{M}([0,1]^d, \lambda)$ be independent and $k_b \geq 1$ for $1 \leq b \leq B_0$. Then with $k = \sum_{b=1}^{B_0} k_b$, for sufficiently large n,*

$$
\mathbb{E}\left[\prod_{b=1}^{B_0}\frac{\mathbb{I}_b(x)}{N_b(x)^{k_b}}\right] \leq \left(\frac{36k}{\inf_{x\in[0,1]^d} f(x)}\right)^{2^{2k}} \prod_{b=1}^{B_0}\mathbb{E}\left[1 \wedge \frac{1}{(n|T_b(x)|)^{k_b}}\right].
$$

Proof (Lemma 12)**.** Define the common refinement of $\{T_b(x) : 1 \leq b \leq B_0\}$ as the class of sets

$$
\mathcal{R} = \left\{\bigcap_{b=1}^{B_0} D_b : D_b \in \{T_b(x), T_b(x)^{\mathrm{c}}\}\right\} \Big\backslash \left\{\emptyset, \bigcap_{b=1}^{B_0} T_b(x)^{\mathrm{c}}\right\}
$$

where $T_b(x)^{\mathrm{c}} = [0,1]^d \setminus T_b(x)$, and let $\mathcal{D} \subset \mathcal{R}$. We will proceed by induction on the elements of $\mathcal{D}$, which represents the subcells we have checked, starting from $\mathcal{D} = \emptyset$ and finishing at $\mathcal{D} = \mathcal{R}$. For $D \in \mathcal{R}$ let $\mathcal{A}(D) = \{1 \leq b \leq B_0 : D \subseteq T_b(x)\}$ be the indices of the trees which are active on subcell D, and for $1 \leq b \leq B_0$ let $\mathcal{A}(b) = \{D \in \mathcal{R} : D \subseteq T_b(x)\}$ be the subcells which are contained in $T_b(x)$, so that $b \in \mathcal{A}(D) \iff D \in \mathcal{A}(b)$. For a subcell $D \in \mathcal{R}$, write $N_b(D) = \sum_{i=1}^n \mathbb{I}\{X_i \in D\}$ so that $N_b(x) = \sum_{D \in \mathcal{A}(b)} N_b(D)$. Note that for any $D \in \mathcal{R} \setminus \mathcal{D}$,

$$\mathbb{E}\left[\prod_{b=1}^{B_0} \frac{1}{1 \vee \left(\sum_{D' \in \mathcal{A}(b) \setminus \mathcal{D}} N_b(D') + n \sum_{D' \in \mathcal{A}(b) \cap \mathcal{D}} |D'|\right)^{k_b}}\right]$$

$$= \mathbb{E}\left[\prod_{b \notin \mathcal{A}(D)} \frac{1}{1 \vee \left(\sum_{D' \in \mathcal{A}(b) \setminus \mathcal{D}} N_b(D') + n \sum_{D' \in \mathcal{A}(b) \cap \mathcal{D}} |D'|\right)^{k_b}}\right.$$

$$\left. \times \mathbb{E}\left[\prod_{b \in \mathcal{A}(D)} \frac{1}{1 \vee \left(\sum_{D' \in \mathcal{A}(b) \setminus \mathcal{D}} N_b(D') + n \sum_{D' \in \mathcal{A}(b) \cap \mathcal{D}} |D'|\right)^{k_b}} \;\middle|\; \mathbf{T}, N_b(D') : D' \in \mathcal{R} \backslash (\mathcal{D} \cup \{D\})\right]\right].$$

Now the inner conditional expectation is over $N_b(D)$ only. Since f is bounded away from zero,

$$\begin{aligned} N_b(D) &\sim \mathrm{Bin}\left(n - \sum_{D' \in \mathcal{R} \backslash (\mathcal{D} \cup \{D\})} N_b(D'),\; \frac{\int_D f(s)\, \mathrm{d}s}{1 - \int_{\bigcup (\mathcal{R} \backslash \mathcal{D}) \backslash D} f(s)\, \mathrm{d}s}\right) \\ &\geq \mathrm{Bin}\left(n - \sum_{D' \in \mathcal{R} \backslash (\mathcal{D} \cup \{D\})} N_b(D'),\; |D| \inf_{x \in [0,1]^d} f(x)\right) \end{aligned}$$

conditional on $\mathbf{T}$ and $N_b(D') : D' \in \mathcal{R} \setminus (\mathcal{D} \cup \{D\})$. Further, by Lemma 10 with $n \geq \lambda^d$,

$$\mathbb{P}\left(\sum_{D' \in \mathcal{R} \backslash (\mathcal{D} \cup \{D\})} N_b(D') > t^{d+1} \frac{n}{\lambda^d} \|f\|_\infty\right) \leq \mathbb{P}\left(N_\cup(x) > t^{d+1} \frac{n}{\lambda^d} \|f\|_\infty\right) \leq 4dB_0 e^{-t/4}.$$

Thus $N_b(D) \geq \mathrm{Bin}(n/2, |D| \inf_x f(x))$ conditional on $\{\mathbf{T}, N_b(D') : D' \in \mathcal{R} \setminus (\mathcal{D} \cup \{D\})\}$ with probability at least $1 - 4dB_0 e^{\frac{-\sqrt{\lambda}}{8\|f\|_\infty}}$. So by Lemma 11,

$$\begin{aligned} &\mathbb{E}\left[\prod_{b \in \mathcal{A}(D)} \frac{1}{1 \vee \left(\sum_{D' \in \mathcal{A}(b) \setminus \mathcal{D}} N_b(D') + n \sum_{D' \in \mathcal{A}(b) \cap \mathcal{D}} |D'|\right)^{k_b}} \;\middle|\; \mathbf{T}, N_b(D') : D' \in \mathcal{R} \setminus (\mathcal{D} \cup \{D\})\right] \\ &\quad \leq (9k)^{2k}\, \mathbb{E}\left[\prod_{b \in \mathcal{A}(D)} \frac{1}{1 \vee \left(\sum_{D' \in \mathcal{A}(b) \setminus (\mathcal{D} \cup \{D\})} N_b(D') + n|D| \inf_x f(x)/2 + n \sum_{D' \in \mathcal{A}(b) \cap \mathcal{D}} |D'|\right)^{k_b}}\right] \\ &\qquad + 4dB_0 e^{\frac{-\sqrt{\lambda}}{8\|f\|_\infty}} \\ &\quad \leq \left(\frac{18k}{\inf_x f(x)}\right)^{2k} \mathbb{E}\left[\prod_{b \in \mathcal{A}(D)} \frac{1}{1 \vee \left(\sum_{D' \in \mathcal{A}(b) \setminus (\mathcal{D} \cup \{D\})} N_b(D') + n \sum_{D' \in \mathcal{A}(b) \cap (\mathcal{D} \cup \{D\})} |D'|\right)^{k_b}}\right] \\ &\qquad + 4dB_0 e^{\frac{-\sqrt{\lambda}}{8\|f\|_\infty}}. \end{aligned}$$

Therefore plugging this back into the marginal expectation yields

$$\begin{aligned}
&\mathbb{E}\left[\prod_{b=1}^{B_0} \frac{1}{1 \vee \left(\sum_{D' \in \mathcal{A}(b) \setminus \mathcal{D}} N_b(D') + n \sum_{D' \in \mathcal{A}(b) \cap \mathcal{D}} |D'|\right)^{k_b}}\right] \\
&\quad \leq \left(\frac{18k}{\inf_x f(x)}\right)^{2k} \mathbb{E}\left[\prod_{b=1}^{B_0} \frac{1}{1 \vee \left(\sum_{D' \in \mathcal{A}(b) \setminus (\mathcal{D} \cup \{D\})} N_b(D') + n \sum_{D' \in \mathcal{A}(b) \cap (\mathcal{D} \cup \{D\})} |D'|\right)^{k_b}}\right] \\
&\qquad + 4dB_0 e^{\frac{-\sqrt{\lambda}}{8\|f\|_\infty}}.
\end{aligned}$$

Now we apply induction, starting with $\mathcal{D} = \emptyset$ and adding $D \in \mathcal{R} \setminus \mathcal{D}$ to $\mathcal{D}$ until $\mathcal{D} = \mathcal{R}$. This takes at most $|\mathcal{R}| \leq 2^k$ steps and yields

$$\begin{aligned}
\mathbb{E}\left[\prod_{b=1}^{B_0} \frac{\mathbb{I}_b(x)}{N_b(x)^{k_b}}\right] &\leq \mathbb{E}\left[\prod_{b=1}^{B_0} \frac{1}{1 \vee N_b(x)^{k_b}}\right] = \mathbb{E}\left[\prod_{b=1}^{B_0} \frac{1}{1 \vee \left(\sum_{D \in \mathcal{A}(b)} N_b(D)\right)^{k_b}}\right] \leq \cdots \\
&\leq \left(\frac{18k}{\inf_x f(x)}\right)^{2^{2k}} \left(\prod_{b=1}^{B_0} \mathbb{E}\left[\frac{1}{1 \vee (n|T_b(x)|)^{k_b}}\right] + 4dB_0 2^k e^{\frac{-\sqrt{\lambda}}{8\|f\|_\infty}}\right),
\end{aligned}$$

where the expectation factorizes due to independence of $T_b(x)$. The last step is to remove the trailing exponential term. To do this, note that by Jensen's inequality,

$$\prod_{b=1}^{B_0} \mathbb{E}\left[\frac{1}{1 \vee (n|T_b(x)|)^{k_b}}\right] \geq \prod_{b=1}^{B_0} \frac{1}{\mathbb{E}\left[1 \vee (n|T_b(x)|)^{k_b}\right]} \geq \prod_{b=1}^{B_0} \frac{1}{n^{k_b}} = n^{-k}$$

while the assumption of $\lambda \gtrsim (\log n)^3$ implies $\lambda \geq (\log n)^3 / C^2$ eventually for some $C > 0$, giving

$$4dB_0 2^k e^{\frac{-\sqrt{\lambda}}{8\|f\|_\infty}} \leq 4dB_0 2^k e^{\frac{-(\log n)^{3/2}}{8C\|f\|_\infty}} \leq 4dB_0 2^k e^{-k \log n - \log(4dB_0 2^k)} \leq n^{-k}$$

for sufficiently large n because B_0, d, and k are fixed. □

Now that moments of (products of) $\frac{\mathbb{I}_b(x)}{N_b(x)}$ have been bounded by moments of (products of) $\frac{1}{n|T_b(x)|}$, we establish further explicit bounds for these in the next result. Note that the problem has been reduced to determining properties of Mondrian cells, so once again we return to the exact cell shape distribution given by Mourtada et al. (2020), and evaluate the appropriate expectations by integration. Note that the truncation by taking the minimum with one inside the expectation is essential here, as otherwise second moment of the inverse Mondrian cell volume is not even finite. As such, there is a "penalty" of $(\log n)^d$ when bounding truncated second moments, and the upper bound for the kth moment is significantly larger than the naive assumption of $(\lambda^d/n)^k$ whenever $k > 2$. This "small cell" phenomenon in which the inverse volumes of Mondrian cells have heavy tails is a recurring challenge in our analysis.

Lemma 13 (Inverse moments of the volume of a Mondrian cell)**.** *Suppose Assumption 2 holds and let $T \sim \mathcal{M}([0,1]^d, \lambda)$. Then with $k \geq 1$, for sufficiently large n,*

$$\mathbb{E}\left[1 \wedge \frac{1}{(n|T(x)|)^k}\right] \leq \left(\frac{2}{2-k} \frac{\lambda^{dk}}{n^k}\right)^{\mathbb{I}\{k<2\}} \left(\frac{3\lambda^{2d}(\log n)^d}{n^2}\right)^{\mathbb{I}\{k \geq 2\}} \prod_{j=1}^{d} \frac{1}{x_j(1-x_j)}.$$

Proof (Lemma 13). By Mourtada et al. (2020, Proposition 1), we have

$$|T(x)| = \prod_{j=1}^{d} \left\{ \left(\frac{1}{\lambda} E_{j1}\right) \wedge x_j + \left(\frac{1}{\lambda} E_{j2}\right) \wedge (1 - x_j) \right\},$$

where E_{j1} and E_{j2} are mutually independent Exp(1) random variables. Thus for any $0 < t < 1$, using the fact that $E_{j1} + E_{j2} \sim \text{Gamma}(2, 1)$,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{1 \vee (n|T(x)|)^k}\right] &\leq \frac{1}{n^k} \mathbb{E}\left[\frac{\mathbb{I}\{\min_j (E_{j1} + E_{j2}) \geq t\}}{|T(x)|^k}\right] + \mathbb{P}\left(\min_{1 \leq j \leq d} (E_{j1} + E_{j2}) < t\right) \\
&\leq \frac{1}{n^k} \prod_{j=1}^{d} \mathbb{E}\left[\frac{\mathbb{I}\{E_{j1} + E_{j2} \geq t\}}{\left(\frac{1}{\lambda} E_{j1} \wedge x_j + \frac{1}{\lambda} E_{j2} \wedge (1 - x_j)\right)^k}\right] + d\, \mathbb{P}(E_{j1} < t) \\
&\leq \frac{\lambda^{dk}}{n^k} \prod_{j=1}^{d} \frac{1}{x_j(1 - x_j)} \mathbb{E}\left[\frac{\mathbb{I}\{E_{j1} + E_{j2} \geq t\}}{(E_{j1} + E_{j2})^k \wedge 1}\right] + d(1 - e^{-t}) \\
&\leq \frac{\lambda^{dk}}{n^k} \prod_{j=1}^{d} \frac{1}{x_j(1 - x_j)} \left(\int_t^1 \frac{e^{-s}}{s^{k-1}} \, \mathrm{d}s + \int_1^\infty s e^{-s} \, \mathrm{d}s\right) + dt \\
&\leq \frac{\lambda^{dk}}{n^k} \prod_{j=1}^{d} \frac{1}{x_j(1 - x_j)} \left(\int_t^1 s^{1-k} \, \mathrm{d}s + 1\right) + dt \\
&= dt + \frac{\lambda^{dk}}{n^k} \prod_{j=1}^{d} \frac{1}{x_j(1 - x_j)} \times \begin{cases} 1 + \frac{1}{2-k} - \frac{t^{2-k}}{2-k} & \text{if } 1 \leq k < 2, \\ 1 - \log t & \text{if } k = 2. \end{cases}
\end{aligned}$$

If $k > 2$ we simply use $\frac{1}{1 \vee (n|T(x)|)^k} \leq \frac{1}{1 \vee (n|T(x)|)^2}$. Now if $1 \leq k < 2$ we let $t \to 0$, giving

$$\mathbb{E}\left[\frac{1}{1 \vee (n|T(x)|)^k}\right] \leq \frac{2}{2 - k} \frac{\lambda^{dk}}{n^k} \prod_{j=1}^{d} \frac{1}{x_j(1 - x_j)},$$

and if $k = 2$ then we set $t = 1/n^2$ so that for sufficiently large n,

$$\mathbb{E}\left[\frac{1}{1 \vee (n|T(x)|)^2}\right] \leq \frac{d}{n^2} + \frac{\lambda^{2d}\left(1 + 2(\log n)^d\right)}{n^2} \prod_{j=1}^{d} \frac{1}{x_j(1 - x_j)} \leq \frac{3\lambda^{2d}(\log n)^d}{n^2} \prod_{j=1}^{d} \frac{1}{x_j(1 - x_j)}.$$

Lower bounds which match up to constants for $1 \leq k < 2$ are easily obtained by noting $\mathbb{E}\left[1 \wedge \frac{1}{(n|T(x)|)^k}\right] \geq \mathbb{E}\left[1 \wedge \frac{1}{n|T(x)|}\right]^k$ by Jensen's inequality and

$$\mathbb{E}\left[1 \wedge \frac{1}{n|T(x)|}\right] \geq \frac{1}{1 + n\mathbb{E}\left[|T(x)|\right]} \geq \frac{1}{1 + 2^d n / \lambda^d} \gtrsim \frac{\lambda^d}{n}.$$

To obtain a lower bound when $k = 2$, note that

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{1 \vee (n|T(x)|)^2}\right] &\geq \mathbb{E}\left[1 \wedge \frac{\lambda^d}{n^2} \prod_{j=1}^{d} \frac{1}{(E_{j1} + E_{j2})^2}\right] \geq \frac{\lambda^d}{n^2} \mathbb{E}\left[\frac{1}{(E_1 + E_2)^2 \vee n^{-1/d}}\right]^d \\
&\geq \frac{\lambda^d}{n^2} \left(\int_{n^{-\frac{1}{2d}}}^1 \frac{e^{-s}}{s} \, \mathrm{d}s\right)^d \geq \frac{\lambda^d}{n^2} \frac{1}{e} \left(\int_{n^{-\frac{1}{2d}}}^1 \frac{1}{s} \, \mathrm{d}s\right)^d \geq \frac{\lambda^d}{n^2} \frac{1}{e} \left(\frac{1}{2d} \log n\right)^d.
\end{aligned}$$

□

The ongoing endeavor to bound moments of (products of) $\frac{\mathbb{I}_b(x)}{N_b(x)}$ is concluded with the next result, chaining together the previous two lemmas to provide an explicit bound with no expectations on the right-hand side.

Lemma 14 (Simplified generalized moment bound for Mondrian random forest denominators). *Grant Assumptions 1 and 2. Let $T_b \sim \mathcal{M}\big([0,1]^d, \lambda\big)$ and $k_b \geq 1$ for $1 \leq b \leq B_0$. Then with $k = \sum_{b=1}^{B_0} k_b$, for sufficiently large n,*

$$
\begin{aligned}
&\mathbb{E}\left[\prod_{b=1}^{B_0} \frac{\mathbb{I}_b(x)}{N_b(x)^{k_b}}\right] \\
&\quad \leq \left(\frac{36k}{\inf_{x\in[0,1]^d} f(x)}\right)^{2^{2k}} \left(\prod_{j=1}^{d} \frac{1}{x_j(1-x_j)}\right)^{B_0} \prod_{b=1}^{B_0} \left(\frac{2}{2-k_b}\frac{\lambda^{dk_b}}{n^{k_b}}\right)^{\mathbb{I}\{k_b<2\}} \left(\frac{3\lambda^{2d}(\log n)^d}{n^2}\right)^{\mathbb{I}\{k_b\geq 2\}}.
\end{aligned}
$$

Proof (Lemma 14). This follows directly from Lemmas 12 and 13. □

Our final preliminary lemma is concerned with further properties of the inverse truncated binomial distribution, again with the aim of analyzing $\frac{\mathbb{I}_b(x)}{N_b(x)}$. This time, instead of merely upper bounding the moments, we aim to give convergence results for those moments, again in terms of moments of $\frac{1}{n|T_b(x)|}$. This time we only need to handle the first and second moment, so this result does not strictly generalize Lemma 11 except in simple cases. The proof is by Taylor's theorem and the Cauchy–Schwarz inequality, using explicit expressions for moments of the binomial distribution and bounds from Lemma 11.

Lemma 15 (Expectation inequalities for the binomial distribution). *Let $N \sim \mathrm{Bin}(n,p)$ and take $a, b \geq 1$. Then*

$$
\begin{aligned}
0 &\leq \mathbb{E}\left[\frac{1}{N+a}\right] - \frac{1}{np+a} \leq \frac{2^{19}}{(np+a)^2}, \\
0 &\leq \mathbb{E}\left[\frac{1}{(N+a)(N+b)}\right] - \frac{1}{(np+a)(np+b)} \leq \frac{2^{27}}{(np+a)(np+b)}\left(\frac{1}{np+a}+\frac{1}{np+b}\right).
\end{aligned}
$$

Proof (Lemma 15). For the first result, Taylor's theorem with Lagrange remainder applied to $N \mapsto \frac{1}{N+a}$ around np gives

$$
\mathbb{E}\left[\frac{1}{N+a}\right] = \mathbb{E}\left[\frac{1}{np+a} - \frac{N-np}{(np+a)^2} + \frac{(N-np)^2}{(\xi+a)^3}\right]
$$

for some ξ between np and N. The second term on the right-hand side is zero-mean, clearly showing the non-negativity part of the result, and applying the Cauchy–Schwarz inequality to the remaining term gives

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{N+a}\right] - \frac{1}{np+a} &\leq \mathbb{E}\left[\frac{(N-np)^2}{(np+a)^3} + \frac{(N-np)^2}{(N+a)^3}\right] \\
&\leq \frac{\mathbb{E}\big[(N-np)^2\big]}{(np+a)^3} + \sqrt{\mathbb{E}\big[(N-np)^4\big]\mathbb{E}\left[\frac{1}{(N+a)^6}\right]}.
\end{aligned}
$$

Now we use $\mathbb{E}\big[(N-np)^4\big] \leq np(1+3np)$ and apply Lemma 11 to see that

$$
\mathbb{E}\left[\frac{1}{N+a}\right] - \frac{1}{np+a} \leq \frac{np}{(np+a)^3} + \sqrt{\frac{54^6 np(1+3np)}{(np+a)^6}} \leq \frac{2^{19}}{(np+a)^2}.
$$

For the second result, Taylor's theorem applied to $N \mapsto \frac{1}{(N+a)(N+b)}$ around np gives

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{(N+a)(N+b)}\right] &= \mathbb{E}\left[\frac{1}{(np+a)(np+b)} - \frac{(N-np)(2np+a+b)}{(np+a)^2(np+b)^2}\right] \\
&\quad + \mathbb{E}\left[\frac{(N-np)^2}{(\xi+a)(\xi+b)}\left(\frac{1}{(\xi+a)^2} + \frac{1}{(\xi+a)(\xi+b)} + \frac{1}{(\xi+b)^2}\right)\right]
\end{aligned}
$$

for some ξ between np and N. The second term on the right-hand side is zero-mean, clearly showing the non-negativity part of the result, and applying the Cauchy–Schwarz inequality to the remaining term gives

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{(N+a)(N+b)}\right] - \frac{1}{np+a} &\leq \mathbb{E}\left[\frac{2(N-np)^2}{(N+a)(N+b)}\left(\frac{1}{(N+a)^2} + \frac{1}{(N+b)^2}\right)\right] \\
&\quad + \mathbb{E}\left[\frac{2(N-np)^2}{(np+a)(np+b)}\left(\frac{1}{(np+a)^2} + \frac{1}{(np+b)^2}\right)\right] \\
&\leq \sqrt{4\mathbb{E}\left[(N-np)^4\right]\mathbb{E}\left[\frac{1}{(N+a)^6(N+b)^2} + \frac{1}{(N+b)^6(N+a)^2}\right]} \\
&\quad + \frac{2\mathbb{E}\big[(N-np)^2\big]}{(np+a)(np+b)}\left(\frac{1}{(np+a)^2} + \frac{1}{(np+b)^2}\right).
\end{aligned}
$$

Now we use $\mathbb{E}\big[(N-np)^4\big] \leq np(1+3np)$ and apply Lemma 11 to see that

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{(N+a)(N+b)}\right] - \frac{1}{np+a} &\leq \sqrt{\frac{4np(1+3np)\cdot 72^8}{(np+a)^2(np+b)^2}\left(\frac{1}{(np+a)^4} + \frac{1}{(np+b)^4}\right)} \\
&\quad + \frac{2np}{(np+a)(np+b)}\left(\frac{1}{(np+a)^2} + \frac{1}{(np+b)^2}\right) \\
&\leq \frac{2^{27}}{(np+a)(np+b)}\left(\frac{1}{np+a} + \frac{1}{np+b}\right).
\end{aligned}
$$

□

A.3 Proofs for Section 3

We give rigorous proofs of the bias and variance characterizations, rate of convergence, central limit theorem, variance estimation, and confidence interval validity results for the Mondrian random forest estimator.

Proof (Lemma 1)**.** We begin by showing that $\mathbb{E}[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}]$ converges to $\mathbb{E}[\hat{\mu}(x) \mid \mathbf{X}]$.

Part 1: Removing the dependence on the trees

By measurability and with $\mu(X_i) = \mathbb{E}[Y_i \mid X_i]$ almost surely,

$$
\mathbb{E}[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}] - \mu(x) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}\big(\mu(X_i) - \mu(x)\big)\frac{\mathbb{I}_{ib}(x)}{N_b(x)}.
$$

Now conditional on $\mathbf{X}$, the terms in the outer sum depend only on T_b so are i.i.d. Since $\mu \in \mathcal{H}^{\beta_\mu}$,

$$
\begin{aligned}
\mathrm{Var}\left[\mathbb{E}\left[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\right] - \mu(x) \mid \mathbf{X}\right] &\leq \frac{1}{B}\mathbb{E}\left[\left(\sum_{i=1}^{n}\big(\mu(X_i) - \mu(x)\big)\frac{\mathbb{I}_{ib}(x)}{N_b(x)}\right)^2 \Bigm| \mathbf{X}\right] \\
&\lesssim \frac{1}{B}\mathbb{E}\left[\max_{1\leq i\leq n}\left\{\mathbb{I}_{ib}(x)\big\|X_i - x\big\|_2^{2(1\wedge\beta_\mu)}\right\}\left(\sum_{i=1}^{n}\frac{\mathbb{I}_{ib}(x)}{N_b(x)}\right)^2 \Bigm| \mathbf{X}\right] \\
&\lesssim \frac{1}{B}\sum_{j=1}^{d}\mathbb{E}\left[|T(x)_j|^{2(1\wedge\beta_\mu)}\right] \lesssim \frac{1}{\lambda^{2(1\wedge\beta_\mu)}B},
\end{aligned}
$$

where we used the law of $T(x)_j$ from Mourtada et al. (2020, Proposition 1). Hence

$$
\mathbb{E}\left[\big(\mathbb{E}\left[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\right] - \mathbb{E}\left[\hat{\mu}(x) \mid \mathbf{X}\right]\big)^2\right] \lesssim \frac{1}{\lambda^{2(1\wedge\beta_\mu)}B}.
$$

Part 2: Showing the conditional bias converges
Now $\mathbb{E}\left[\hat{\mu}(x) \mid \mathbf{X}\right]$ is a non-linear function of the i.i.d. random variables X_i, so we use the Efron–Stein inequality (Efron and Stein, 1981) to bound its variance. Let $\tilde{X}_{ij} = X_i$ if $i \neq j$ and be an independent copy of X_j, denoted $\tilde{X}_j$, if $i = j$. Write $\tilde{\mathbf{X}}_j = (\tilde{X}_{1j}, \ldots, \tilde{X}_{nj})$ and similarly $\tilde{\mathbb{I}}_{ijb}(x) = \mathbb{I}\big\{\tilde{X}_{ij} \in T_b(x)\big\}$ and $N_{jb}(x) = \sum_{i=1}^{n}\tilde{\mathbb{I}}_{ijb}(x)$.

$$
\begin{aligned}
&\mathrm{Var}\left[\sum_{i=1}^{n}\big(\mu(X_i) - \mu(x)\big)\mathbb{E}\left[\frac{\mathbb{I}_{ib}(x)}{N_b(x)} \Bigm| \mathbf{X}\right]\right] \\
&\leq \frac{1}{2}\sum_{j=1}^{n}\mathbb{E}\left[\left(\sum_{i=1}^{n}\big(\mu(X_i) - \mu(x)\big)\mathbb{E}\left[\frac{\mathbb{I}_{ib}(x)}{N_b(x)} \Bigm| \mathbf{X}\right] - \sum_{i=1}^{n}\Big(\mu(\tilde{X}_{ij}) - \mu(x)\Big)\mathbb{E}\left[\frac{\tilde{\mathbb{I}}_{ijb}(x)}{\tilde{N}_{jb}(x)} \Bigm| \tilde{\mathbf{X}}_j\right]\right)^2\right] \\
&\leq \frac{1}{2}\sum_{j=1}^{n}\mathbb{E}\left[\left(\sum_{i=1}^{n}\left(\big(\mu(X_i) - \mu(x)\big)\frac{\mathbb{I}_{ib}(x)}{N_b(x)} - \Big(\mu(\tilde{X}_{ij}) - \mu(x)\Big)\frac{\tilde{\mathbb{I}}_{ijb}(x)}{\tilde{N}_{jb}(x)}\right)\right)^2\right] \\
&\leq \sum_{j=1}^{n}\mathbb{E}\left[\left(\sum_{i\neq j}\big(\mu(X_i) - \mu(x)\big)\left(\frac{\mathbb{I}_{ib}(x)}{N_b(x)} - \frac{\mathbb{I}_{ib}(x)}{\tilde{N}_{jb}(x)}\right)\right)^2\right] \\
&\quad + 2\sum_{j=1}^{n}\mathbb{E}\left[\big(\mu(X_j) - \mu(x)\big)^2\frac{\mathbb{I}_{jb}(x)}{N_b(x)^2}\right].
\end{aligned} \tag{19}
$$

For the first term in (19) to be non-zero, we must have $|N_b(x) - \tilde{N}_{jb}(x)| = 1$. Writing $N_{-jb}(x) = \sum_{i\neq j}\mathbb{I}_{ib}(x)$, we may assume by symmetry that $\tilde{N}_{jb}(x) = N_{-jb}(x)$ and $N_b(x) = N_{-jb}(x) + 1$, and also that $\mathbb{I}_{jb}(x) = 1$. Hence since f is bounded and $\mu \in \mathcal{H}^{\beta_\mu}$, writing $\mathbb{I}_{-jb}(x) = \mathbb{I}\left\{N_{-jb}(x) \geq 1\right\}$, by the Hölder inequality, Lemma 9 and Lemma 14,

$$\sum_{j=1}^{n} \mathbb{E}\left[\left(\sum_{i\neq j} (\mu(X_i) - \mu(x)) \left(\frac{\mathbb{I}_{ib}(x)}{N_b(x)} - \frac{\mathbb{I}_{ib}(x)}{\tilde{N}_{jb}(x)}\right)\right)^2\right]$$

$$\lesssim \sum_{j=1}^{n} \mathbb{E}\left[\max_{1\leq l\leq d} |T_b(x)_l|^{2(1\wedge\beta_\mu)} \left(\frac{\sum_{i\neq j} \mathbb{I}_{ib}(x)\mathbb{I}_{jb}(x)}{N_{-jb}(x)(N_{-jb}(x)+1)}\right)^2\right] \lesssim \mathbb{E}\left[\max_{1\leq l\leq d} |T_b(x)_l|^{2(1\wedge\beta_\mu)} \frac{\mathbb{I}_b(x)}{N_b(x)}\right]$$

$$\lesssim \mathbb{E}\left[\max_{1\leq l\leq d} |T_b(x)_l|^{6(1\wedge\beta_\mu)}\right]^{1/3} \mathbb{E}\left[\frac{\mathbb{I}_b(x)}{N_b(x)^{3/2}}\right]^{2/3} \lesssim \frac{1}{\lambda^{2(1\wedge\beta_\mu)}} \frac{\lambda^d}{n}.$$

For the second term in (19) we again use $\mu \in \mathcal{H}^{\beta_\mu}$ to see

$$\sum_{j=1}^{n} \mathbb{E}\left[(\mu(X_j) - \mu(x))^2 \frac{\mathbb{I}_{jb}(x)}{N_b(x)^2}\right] \lesssim \mathbb{E}\left[\max_{1\leq l\leq d} |T_b(x)_l|^{2(1\wedge\beta_\mu)} \frac{\mathbb{I}_b(x)}{N_b(x)}\right] \lesssim \frac{1}{\lambda^{2(1\wedge\beta_\mu)}} \frac{\lambda^d}{n}$$

in the same manner. Hence

$$\operatorname{Var}\left[\sum_{i=1}^{n} (\mu(X_i) - \mu(x))\, \mathbb{E}\left[\frac{\mathbb{I}_{ib}(x)}{N_b(x)} \;\middle|\; \mathbf{X}\right]\right] \lesssim \frac{1}{\lambda^{2(1\wedge\beta_\mu)}} \frac{\lambda^d}{n},$$

and so by the above and the previous part,

$$\mathbb{E}\left[\left(\mathbb{E}\left[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\right] - \mathbb{E}\left[\hat{\mu}(x)\right]\right)^2\right] \lesssim \frac{1}{\lambda^{2(1\wedge\beta_\mu)} B} + \frac{1}{\lambda^{2(1\wedge\beta_\mu)}} \frac{\lambda^d}{n}.$$

Part 3: Computing the limiting bias

It remains to compute the limiting value of $\mathbb{E}[\hat{\mu}(x)] - \mu(x)$. Let $\mathbf{X}_{-i} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ and $N_{-ib}(x) = \sum_{j=1}^{n} \mathbb{I}\{j \neq i\}\mathbb{I}\{X_j \in T_b(x)\}$. Then

$$\mathbb{E}[\hat{\mu}(x)] - \mu(x) = \mathbb{E}\left[\sum_{i=1}^{n} (\mu(X_i) - \mu(x)) \frac{\mathbb{I}_{ib}(x)}{N_b(x)}\right] = \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}\left[\frac{(\mu(X_i) - \mu(x))\, \mathbb{I}_{ib}(x)}{N_{-ib}(x)+1} \;\middle|\; \mathbf{T}, \mathbf{X}_{-i}\right]\right]$$

$$= n\, \mathbb{E}\left[\frac{\int_{T_b(x)} (\mu(s) - \mu(x))\, f(s)\, \mathrm{d}s}{N_{-ib}(x)+1}\right].$$

By Lemma 15, as $N_{-ib}(x) \sim \operatorname{Bin}\left(n-1, \int_{T_b(x)} f(s)\, \mathrm{d}s\right)$ given $\mathbf{T}$ and f is bounded away from zero,

$$\left|\mathbb{E}\left[\frac{1}{N_{-ib}(x)+1} \;\middle|\; \mathbf{T}\right] - \frac{1}{(n-1)\int_{T_b(x)} f(s)\, \mathrm{d}s + 1}\right| \lesssim \frac{1}{n^2 \left(\int_{T_b(x)} f(s)\, \mathrm{d}s\right)^2} \wedge 1 \lesssim \frac{1}{n^2 |T_b(x)|^2} \wedge 1$$

and also

$$\left|\frac{1}{(n-1)\int_{T_b(x)} f(s)\, \mathrm{d}s + 1} - \frac{1}{n\int_{T_b(x)} f(s)\, \mathrm{d}s}\right| \lesssim \frac{1}{n^2 \left(\int_{T_b(x)} f(s)\, \mathrm{d}s\right)^2} \wedge 1 \lesssim \frac{1}{n^2 |T_b(x)|^2} \wedge 1.$$

So by Lemma 9 and Lemma 13, since $\mu \in \mathcal{H}^{\beta_\mu}$ and f is bounded below, using the Hölder inequality,

$$
\begin{aligned}
&\left| \mathbb{E}\left[\hat{\mu}(x)\right] - \mu(x) - \mathbb{E}\left[\frac{\int_{T_b(x)} (\mu(s) - \mu(x)) f(s)\,\mathrm{d}s}{\int_{T_b(x)} f(s)\,\mathrm{d}s}\right]\right| \\
&\quad \lesssim \mathbb{E}\left[\frac{n\int_{T_b(x)} |\mu(s) - \mu(x)|\, f(s)\,\mathrm{d}s}{n^2|T_b(x)|^2 \vee 1}\right] \lesssim \mathbb{E}\left[\frac{\max_{1\le l\le d}|T_b(x)_l|^{1\wedge\beta_\mu}}{n|T_b(x)| \vee 1}\right] \\
&\quad \lesssim \mathbb{E}\left[\max_{1\le l\le d}|T_b(x)_l|^{3(1\wedge\beta_\mu)}\right]^{1/3} \mathbb{E}\left[\frac{1}{n^{3/2}|T_b(x)|^{3/2} \vee 1}\right]^{2/3} \lesssim \frac{1}{\lambda^{1\wedge\beta_\mu}}\frac{\lambda^d}{n}.
\end{aligned}
$$

Next set $A = \frac{1}{f(x)|T_b(x)|}\int_{T_b(x)}(f(s) - f(x))\,\mathrm{d}s \ge \inf_{s\in[0,1]^d}\frac{f(s)}{f(x)} - 1 > -1$. Use the Maclaurin series of $\frac{1}{1+x}$ up to order $\underline{\beta} - 1$ to see $\frac{1}{1+A} = \sum_{k=0}^{\underline{\beta}-1}(-1)^k A^k + O\left(A^{\underline{\beta}}\right)$. Hence

$$
\begin{aligned}
\mathbb{E}\left[\frac{\int_{T_b(x)} (\mu(s) - \mu(x)) f(s)\,\mathrm{d}s}{\int_{T_b(x)} f(s)\,\mathrm{d}s}\right] &= \mathbb{E}\left[\frac{\int_{T_b(x)} (\mu(s) - \mu(x)) f(s)\,\mathrm{d}s}{f(x)|T_b(x)|}\frac{1}{1+A}\right] \\
&= \mathbb{E}\left[\frac{\int_{T_b(x)} (\mu(s) - \mu(x)) f(s)\,\mathrm{d}s}{f(x)|T_b(x)|}\left(\sum_{k=0}^{\underline{\beta}-1}(-1)^k A^k + O\left(|A|^{\underline{\beta}}\right)\right)\right].
\end{aligned}
$$

Note that since $\mu \in \mathcal{H}^{\beta_\mu}$ and $f \in \mathcal{H}^{\beta_f}$, by integrating the tail probability given in Lemma 9, the Maclaurin remainder term is bounded by

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\int_{T_b(x)} |\mu(s) - \mu(x)|\, f(s)\,\mathrm{d}s}{f(x)|T_b(x)|}|A|^{\underline{\beta}}\right] \\
&\quad = \mathbb{E}\left[\frac{\int_{T_b(x)} |\mu(s) - \mu(x)|\, f(s)\,\mathrm{d}s}{f(x)|T_b(x)|}\left(\frac{1}{f(x)|T_b(x)|}\int_{T_b(x)}(f(s) - f(x))\,\mathrm{d}s\right)^{\underline{\beta}}\right] \\
&\quad \lesssim \mathbb{E}\left[\max_{1\le l\le d}|T_b(x)_l|^{1\wedge\beta_\mu}\left(\max_{1\le l\le d}|T_b(x)_l|^{1\wedge\beta_f}\right)^{\underline{\beta}}\right] \\
&\quad \lesssim \mathbb{E}\left[\max_{1\le l\le d}|T_b(x)_l|^{1\wedge\beta_\mu + \underline{\beta}(1\wedge\beta_f)}\right] \lesssim \mathbb{E}\left[\max_{1\le l\le d}|T_b(x)_l|^{\beta}\right] \lesssim \frac{d}{\lambda^\beta}\int_0^\infty e^{-t}t^\beta\,\mathrm{d}t \lesssim \frac{1}{\lambda^\beta},
\end{aligned}
$$

where we used that $1\wedge\beta_\mu + \underline{\beta}(1\wedge\beta_f) \ge \beta$. Hence to summarize the progress so far, we have

$$
\begin{aligned}
&\left|\mathbb{E}\left[\hat{\mu}(x)\right] - \mu(x) - \sum_{k=0}^{\underline{\beta}-1}(-1)^k\,\mathbb{E}\left[\frac{\int_{T_b(x)} (\mu(s) - \mu(x)) f(s)\,\mathrm{d}s}{f(x)^{k+1}|T_b(x)|^{k+1}}\left(\int_{T_b(x)}(f(s) - f(x))\,\mathrm{d}s\right)^k\right]\right| \\
&\quad \lesssim \frac{1}{\lambda^{1\wedge\beta_\mu}}\frac{\lambda^d}{n} + \frac{1}{\lambda^\beta}.
\end{aligned}
$$

We continue to evaluate this expectation. First, by Taylor's theorem and with ν a multi-index, since $f \in \mathcal{H}^{\beta_f}$,

$$
\left(\int_{T_b(x)}(f(s) - f(x))\,\mathrm{d}s\right)^k = \left(\sum_{|\nu|=1}^{\underline{\beta}_f}\frac{\partial^\nu f(x)}{\nu!}\int_{T_b(x)}(s-x)^\nu\,\mathrm{d}s\right)^k + O\left(|T_b(x)|^k\max_{1\le l\le d}|T_b(x)_l|^{\beta_f}\right).
$$

Next, by the multinomial theorem with a multi-index u indexed by ν with $|\nu| \geq 1$,

$$\left(\sum_{|\nu|=1}^{\beta_f} \frac{\partial^\nu f(x)}{\nu!} \int_{T_b(x)} (s-x)^\nu \, \mathrm{d}s \right)^k = \sum_{|u|=k} \binom{k}{u} \left(\frac{\partial^\nu f(x)}{\nu!} \int_{T_b(x)} (s-x)^\nu \, \mathrm{d}s \right)^u$$

where $\binom{k}{u}$ is a multinomial coefficient. By Taylor's theorem with $\mu \in \mathcal{H}^{\beta_\mu}$ and $f \in \mathcal{H}^{\beta_f}$, and using $\beta_\mu \wedge (1 \wedge \beta_\mu + \beta_f) \geq \beta$,

$$\begin{aligned}
&\int_{T_b(x)} (\mu(s) - \mu(x)) \, f(s) \, \mathrm{d}s \\
&\quad = \sum_{|\nu'|=1}^{\beta_\mu} \sum_{|\nu''|=0}^{\beta_f} \frac{\partial^{\nu'} \mu(x)}{\nu'!} \frac{\partial^{\nu''} f(x)}{\nu''!} \int_{T_b(x)} (s-x)^{\nu'+\nu''} \, \mathrm{d}s + O\left(|T_b(x)| \max_{1 \leq l \leq d} |T_b(x)_l|^\beta \right).
\end{aligned}$$

Now by integrating the tail probabilities in Lemma 9, $\mathbb{E}\left[\max_{1 \leq l \leq d} |T_b(x)_l|^\beta\right] \lesssim \frac{1}{\lambda^\beta}$. Therefore by Lemma 13, writing $T_b(x)^\nu$ for $\int_{T_b(x)} (s-x)^\nu \, \mathrm{d}s$,

$$\begin{aligned}
&\sum_{k=0}^{\beta-1} (-1)^k \, \mathbb{E}\left[\frac{\int_{T_b(x)} (\mu(s) - \mu(x)) \, f(s) \, \mathrm{d}s}{f(x)^{k+1} |T_b(x)|^{k+1}} \left(\int_{T_b(x)} (f(s) - f(x)) \, \mathrm{d}s \right)^k \right] \\
&= \sum_{k=0}^{\beta-1} (-1)^k \, \mathbb{E}\left[\frac{\sum_{|\nu'|=1}^{\beta_\mu} \sum_{|\nu''|=0}^{\beta_f} \frac{\partial^{\nu'} \mu(x)}{\nu'!} \frac{\partial^{\nu''} f(x)}{\nu''!} T_b(x)^{\nu'+\nu''}}{f(x)^{k+1} |T_b(x)|^{k+1}} \sum_{|u|=k} \binom{k}{u} \left(\frac{\partial^\nu f(x)}{\nu!} T_b(x)^\nu \right)^u \right] + O\left(\frac{1}{\lambda^\beta} \right) \\
&= \sum_{|\nu'|=1}^{\beta_\mu} \sum_{|\nu''|=0}^{\beta_f} \sum_{|u|=0}^{\beta-1} \frac{\partial^{\nu'} \mu(x)}{\nu'!} \frac{\partial^{\nu''} f(x)}{\nu''!} \left(\frac{\partial^\nu f(x)}{\nu!} \right)^u \binom{|u|}{u} \frac{(-1)^{|u|}}{f(x)^{|u|+1}} \mathbb{E}\left[\frac{T_b(x)^{\nu'+\nu''} (T_b(x)^\nu)^u}{|T_b(x)|^{|u|+1}} \right] + O\left(\frac{1}{\lambda^\beta} \right).
\end{aligned}$$

Now we show this is a polynomial in λ. For $1 \leq j \leq d$, define the independent variables $E_{1j*} \sim \mathrm{Exp}(1) \wedge (\lambda x_j)$ and $E_{2j*} \sim \mathrm{Exp}(1) \wedge (\lambda(1 - x_j))$ so $T_b(x) = \prod_{j=1}^d [x_j - E_{1j*}/\lambda, x_j + E_{2j*}/\lambda]$. Then

$$\begin{aligned}
T_b(x)^\nu &= \int_{T_b(x)} (s-x)^\nu \, \mathrm{d}s = \prod_{j=1}^d \int_{x_j - E_{1j*}/\lambda}^{x_j + E_{2j*}/\lambda} (s - x_j)^{\nu_j} \, \mathrm{d}s = \prod_{j=1}^d \int_{-E_{1j*}}^{E_{2j*}} (s/\lambda)^{\nu_j} 1/\lambda \, \mathrm{d}s \\
&= \lambda^{-d-|\nu|} \prod_{j=1}^d \int_{-E_{1j*}}^{E_{2j*}} s^{\nu_j} \, \mathrm{d}s = \lambda^{-d-|\nu|} \prod_{j=1}^d \frac{E_{2j*}^{\nu_j+1} + (-1)^{\nu_j} E_{1j*}^{\nu_j+1}}{\nu_j + 1}.
\end{aligned}$$

So by independence over j,

$$\begin{aligned}
&\mathbb{E}\left[\frac{T_b(x)^{\nu'+\nu''} (T_b(x)^\nu)^u}{|T_b(x)|^{|u|+1}} \right] \\
&\quad = \lambda^{-|\nu'|-|\nu''|-|\nu| \cdot u} \prod_{j=1}^d \mathbb{E}\left[\frac{E_{2j*}^{\nu'_j+\nu''_j+1} + (-1)^{\nu'_j+\nu''_j} E_{1j*}^{\nu'_j+\nu''_j+1}}{(\nu'_j + \nu''_j + 1)(E_{2j*} + E_{1j*})} \frac{\left(E_{2j*}^{\nu_j+1} + (-1)^{\nu_j} E_{1j*}^{\nu_j+1} \right)^u}{(\nu_j + 1)^u (E_{2j*} + E_{1j*})^{|u|}} \right]. \qquad (20)
\end{aligned}$$

The final step is to replace E_{1j*} by $E_{1j} \sim \mathrm{Exp}(1)$ and similarly for E_{2j*}. Note that for a positive constant C,

$$\mathbb{P}\left(\bigcup_{j=1}^d (\{E_{1j*} \neq E_{1j}\} \cup \{E_{2j*} \neq E_{2j}\}) \right) \leq 2d \, \mathbb{P}\left(\mathrm{Exp}(1) \geq \lambda \min_{1 \leq j \leq d} (x_j \wedge (1 - x_j)) \right) \leq 2d e^{-C\lambda}.$$

Further, the quantity inside the expectation in (20) is bounded almost surely by one and so the error incurred by replacing E_{1j*} and E_{2j*} by E_{1j} and E_{2j} in (20) is at most $2de^{-C\lambda} \lesssim \lambda^{-\beta}$. Thus the limiting bias is

$$\mathbb{E}\left[\hat{\mu}(x)\right] - \mu(x) = \sum_{|\nu'|=1}^{\underline{\beta}_\mu} \sum_{|\nu''|=0}^{\underline{\beta}_f} \sum_{|u|=0}^{\underline{\beta}-1} \frac{\partial^{\nu'}\mu(x)}{\nu'!} \frac{\partial^{\nu''}f(x)}{\nu''!} \left(\frac{\partial^{\nu}f(x)}{\nu!}\right)^u \binom{|u|}{u} \frac{(-1)^{|u|}}{f(x)^{|u|+1}} \lambda^{-|\nu'|-|\nu''|-|\nu|\cdot u} \tag{21}$$

$$\times \prod_{j=1}^{d} \mathbb{E}\left[\frac{E_{2j}^{\nu_j'+\nu_j''+1} + (-1)^{\nu_j'+\nu_j''} E_{1j}^{\nu_j'+\nu_j''+1}}{(\nu_j'+\nu_j''+1)(E_{2j}+E_{1j})} \frac{\left(E_{2j}^{\nu_j+1} + (-1)^{\nu_j} E_{1j}^{\nu_j+1}\right)^u}{(\nu_j+1)^u (E_{2j}+E_{1j})^{|u|}}\right] + O\left(\frac{1}{\lambda^{1\wedge\beta_\mu}} \frac{\lambda^d}{n} + \frac{1}{\lambda^\beta}\right),$$

recalling that u is a multi-index which is indexed by the multi-index ν. This is a polynomial in λ of degree at most $\underline{\beta}$, since higher-order terms can be absorbed into $O(1/\lambda^\beta)$, which has finite coefficients depending only on the derivatives up to order $\underline{\beta} \leq \underline{\beta}_\mu$ and $\underline{\beta} - 1 \leq \underline{\beta}_f$ of μ and f respectively at x. Now we show that the odd-degree terms in this polynomial are all zero. Note that a term is of odd degree if and only if $|\nu'| + |\nu''| + |\nu| \cdot u$ is odd. This implies that there exists $1 \leq j \leq d$ such that exactly one of either $\nu_j' + \nu_j''$ is odd or $\sum_{|\nu|=1}^{\underline{\beta}-1} \nu_j u_\nu$ is odd.

If $\nu_j' + \nu_j''$ is odd, then $\sum_{|\nu|=1}^{\underline{\beta}-1} \nu_j u_\nu$ is even, so $|\{\nu : \nu_j u_\nu \text{ is odd}\}|$ is even. Consider the effect of swapping E_{1j} and E_{2j}, an operation which by independence preserves their joint law, in each of

$$\frac{E_{2j}^{\nu_j'+\nu_j''+1} + (-1)^{\nu_j'+\nu_j''} E_{1j}^{\nu_j'+\nu_j''+1}}{E_{2j}+E_{1j}} \tag{22}$$

and

$$\frac{\left(E_{2j}^{\nu_j+1} + (-1)^{\nu_j} E_{1j}^{\nu_j+1}\right)^u}{(E_{2j}+E_{1j})^{|u|}} = \prod_{\substack{|\nu|=1 \\ \nu_j u_\nu \text{ even}}}^{\underline{\beta}-1} \frac{\left(E_{2j}^{\nu_j+1} + (-1)^{\nu_j} E_{1j}^{\nu_j+1}\right)^{u_\nu}}{(E_{2j}+E_{1j})^{u_\nu}} \prod_{\substack{|\nu|=1 \\ \nu_j u_\nu \text{ odd}}}^{\underline{\beta}-1} \frac{\left(E_{2j}^{\nu_j+1} + (-1)^{\nu_j} E_{1j}^{\nu_j+1}\right)^{u_\nu}}{(E_{2j}+E_{1j})^{u_\nu}}. \tag{23}$$

Clearly $\nu_j' + \nu_j''$ being odd inverts the sign of (22). For (23), each term in the first product has either ν_j even or u_ν even, so its sign is preserved. Every term in the second product of (23) has its sign inverted due to both ν_j and u_ν being odd, but there are an even number of terms, preserving the overall sign. Therefore the expected product of (22) and (23) is zero by symmetry.

If however $\nu_j' + \nu_j''$ is even, then $\sum_{|\nu|=1}^{\underline{\beta}-1} \nu_j u_\nu$ is odd so $|\{\nu : \nu_j u_\nu \text{ is odd}\}|$ is odd. Clearly the sign of (22) is preserved. Again the sign of the first product in (23) is preserved, and the sign of every term in (23) is inverted. However there are now an odd number of terms in the second product, so its overall sign is inverted. Therefore the expected product of (22) and (23) is again zero.

Part 4: Calculating the second-order bias

Next we calculate some special cases, beginning with the form of the leading second-order bias, where the exponent in λ is $|\nu'| + |\nu''| + u \cdot |\nu| = 2$, proceeding by cases on the values of $|\nu'|$, $|\nu''|$, and $|u|$. Firstly, if $|\nu'| = 2$ then $|\nu''| = |u| = 0$. Note that if any $\nu_j' = 1$ then the expectation in (21) is zero. Hence we can assume $\nu_j' \in \{0, 2\}$, yielding

$$\frac{1}{2\lambda^2} \sum_{j=1}^{d} \frac{\partial^2 \mu(x)}{\partial x_j^2} \frac{1}{3} \mathbb{E}\left[\frac{E_{2j}^3 + E_{1j}^3}{E_{2j}+E_{1j}}\right] = \frac{1}{2\lambda^2} \sum_{j=1}^{d} \frac{\partial^2 \mu(x)}{\partial x_j^2} \frac{1}{3} \mathbb{E}\left[E_{1j}^2 + E_{2j}^2 - E_{1j}E_{2j}\right] = \frac{1}{2\lambda^2} \sum_{j=1}^{d} \frac{\partial^2 \mu(x)}{\partial x_j^2},$$

where we used that E_{1j} and E_{2j} are independent Exp(1). Next we consider $|\nu'| = 1$ and $|\nu''| = 1$, so $|u| = 0$. Note that if $\nu'_j = \nu''_{j'} = 1$ with $j \neq j'$ then the expectation in (21) is zero. So we need only consider $\nu'_j = \nu''_j = 1$, giving

$$\frac{1}{\lambda^2}\frac{1}{f(x)}\sum_{j=1}^{d}\frac{\partial \mu(x)}{\partial x_j}\frac{\partial f(x)}{\partial x_j}\frac{1}{3}\mathbb{E}\left[\frac{E_{2j}^3+E_{1j}^3}{E_{2j}+E_{1j}}\right] = \frac{1}{\lambda^2}\frac{1}{f(x)}\sum_{j=1}^{d}\frac{\partial \mu(x)}{\partial x_j}\frac{\partial f(x)}{\partial x_j}.$$

Finally we have the case where $|\nu'| = 1$, $|\nu''| = 0$ and $|u| = 1$. Then $u_\nu = 1$ for some $|\nu| = 1$ and zero otherwise. Note that if $\nu'_j = \nu_{j'} = 1$ with $j \neq j'$ then the expectation is zero. So we need only consider $\nu'_j = \nu_j = 1$, giving

$$\begin{aligned}
&-\frac{1}{\lambda^2}\frac{1}{f(x)}\sum_{j=1}^{d}\frac{\partial \mu(x)}{\partial x_j}\frac{\partial f(x)}{\partial x_j}\frac{1}{4}\mathbb{E}\left[\frac{(E_{2j}^2-E_{1j}^2)^2}{(E_{2j}+E_{1j})^2}\right] \\
&= -\frac{1}{4\lambda^2}\frac{1}{f(x)}\sum_{j=1}^{d}\frac{\partial \mu(x)}{\partial x_j}\frac{\partial f(x)}{\partial x_j}\mathbb{E}\left[E_{1j}^2+E_{2j}^2-2E_{1j}E_{2j}\right] = -\frac{1}{2\lambda^2}\frac{1}{f(x)}\sum_{j=1}^{d}\frac{\partial \mu(x)}{\partial x_j}\frac{\partial f(x)}{\partial x_j}.
\end{aligned}$$

Hence the second-order bias term is

$$\frac{1}{2\lambda^2}\sum_{j=1}^{d}\frac{\partial^2 \mu(x)}{\partial x_j^2} + \frac{1}{2\lambda^2}\frac{1}{f(x)}\sum_{j=1}^{d}\frac{\partial \mu(x)}{\partial x_j}\frac{\partial f(x)}{\partial x_j}.$$

Part 5: Calculating the bias if the data is uniformly distributed

If $X_i \sim \mathrm{Unif}\left([0,1]^d\right)$ then $f(x) = 1$ and the bias expansion from (21) becomes

$$\sum_{|\nu'|=1}^{\beta}\lambda^{-|\nu'|}\frac{\partial^{\nu'}\mu(x)}{\nu'!}\prod_{j=1}^{d}\mathbb{E}\left[\frac{E_{2j}^{\nu'_j+1}+(-1)^{\nu'_j}E_{1j}^{\nu'_j+1}}{(\nu'_j+1)(E_{2j}+E_{1j})}\right].$$

Note that this is zero if any ν'_j is odd. Therefore we can group these terms based on the exponent of λ to see

$$\frac{B_r(x)}{\lambda^{2r}} = \frac{1}{\lambda^{2r}}\sum_{|\nu|=r}\frac{\partial^{2\nu}\mu(x)}{(2\nu)!}\prod_{j=1}^{d}\frac{1}{2\nu_j+1}\mathbb{E}\left[\frac{E_{2j}^{2\nu_j+1}+E_{1j}^{2\nu_j+1}}{E_{2j}+E_{1j}}\right].$$

Since $\int_0^\infty \frac{e^{-t}}{a+t}\,\mathrm{d}t = e^a\Gamma(0,a)$ and $\int_0^\infty s^a\Gamma(0,a)\,\mathrm{d}s = \frac{a!}{a+1}$, with $\Gamma(0,a) = \int_a^\infty \frac{e^{-t}}{t}\,\mathrm{d}t$ the upper incomplete gamma function, the expectation is easily calculated as

$$\mathbb{E}\left[\frac{E_{2j}^{2\nu_j+1}+E_{1j}^{2\nu_j+1}}{E_{2j}+E_{1j}}\right] = 2\int_0^\infty s^{2\nu_j+1}e^{-s}\int_0^\infty \frac{e^{-t}}{s+t}\,\mathrm{d}t\,\mathrm{d}s = 2\int_0^\infty s^{2\nu_j+1}\Gamma(0,s)\,\mathrm{d}s = \frac{(2\nu_j+1)!}{\nu_j+1},$$

so

$$\frac{B_r(x)}{\lambda^{2r}} = \frac{1}{\lambda^{2r}}\sum_{|\nu|=r}\frac{\partial^{2\nu}\mu(x)}{(2\nu)!}\prod_{j=1}^{d}\frac{1}{2\nu_j+1}\frac{(2\nu_j+1)!}{\nu_j+1} = \frac{1}{\lambda^{2r}}\sum_{|\nu|=r}\partial^{2\nu}\mu(x)\prod_{j=1}^{d}\frac{1}{\nu_j+1}.$$

□

Proof (Lemma 2)**.** By Lemma 5 with $J = 0$, $a_0 = 1$, and $\omega_0 = 1$. □

Proof (Theorem 1)**.** By Theorem 4 with $J = 0$, $a_0 = 1$, and $\omega_0 = 1$. □

Proof (Theorem 2)**.** By Theorem 5 with $J = 0$, $a_0 = 1$, and $\omega_0 = 1$. □

Proof (Lemma 3)**.** By Lemma 6 with $J = 0$, $a_0 = 1$, and $\omega_0 = 1$. □

Proof (Theorem 3)**.** By Theorem 6 with $J = 0$, $a_0 = 1$, and $\omega_0 = 1$, replacing β by $2 \wedge \beta$. □

A.4 Proofs for Section 4

We give rigorous proofs of the bias and variance characterizations, minimax optimality, central limit theorem, variance estimation, and confidence interval validity results for the debiased Mondrian random forest estimator.

The bias characterization of Lemma 4 with debiasing is a purely algebraic consequence of the original bias characterization and the construction of the debiased Mondrian random forest estimator.

Proof (Lemma 4)**.** By the definition of the debiased estimator and Lemma 1, as J and a_r are fixed,

$$\begin{aligned}
&\mathbb{E}\left[\left(\mathbb{E}\left[\hat{\mu}_{\mathrm{d}}(x) \mid \mathbf{X}, \mathbf{T}\right] - \sum_{l=0}^{J} \omega_l \left(\mu(x) + \sum_{r=1}^{\lfloor \underline{\beta}/2 \rfloor} \frac{B_r(x)}{a_l^{2r} \lambda^{2r}}\right)\right)^2\right] \\
&\quad = \mathbb{E}\left[\left(\sum_{l=0}^{J} \omega_l \left(\mathbb{E}\left[\hat{\mu}_l(x) \mid \mathbf{X}, \mathbf{T}\right] - \mu(x) - \sum_{r=1}^{\lfloor \underline{\beta}/2 \rfloor} \frac{B_r(x)}{a_l^{2r} \lambda^{2r}}\right)\right)^2\right] \\
&\quad \lesssim \sum_{l=0}^{J} \mathbb{E}\left[\left(\mathbb{E}\left[\hat{\mu}_l(x) \mid \mathbf{X}, \mathbf{T}\right] - \mu(x) - \sum_{r=1}^{\lfloor \underline{\beta}/2 \rfloor} \frac{B_r(x)}{a_l^{2r} \lambda^{2r}}\right)^2\right] \lesssim \frac{1}{\lambda^{2\beta}} + \frac{1}{\lambda^{2(1\wedge\beta)} B} + \frac{1}{\lambda^{2(1\wedge\beta)}} \frac{\lambda^d}{n}.
\end{aligned}$$

It remains to evaluate the resulting bias. Recalling that $A_{rs} = a_{r-1}^{2-2s}$ and $A\omega = e_0$, we have

$$\begin{aligned}
\sum_{l=0}^{J} \omega_l \left(\mu(x) + \sum_{r=1}^{\lfloor \underline{\beta}/2 \rfloor} \frac{B_r(x)}{a_l^{2r} \lambda^{2r}}\right) &= \mu(x) \sum_{l=0}^{J} \omega_l + \sum_{r=1}^{\lfloor \underline{\beta}/2 \rfloor} \frac{B_r(x)}{\lambda^{2r}} \sum_{l=0}^{J} \frac{\omega_l}{a_l^{2r}} \\
&= \mu(x) (A\omega)_1 + \sum_{r=1}^{\lfloor \underline{\beta}/2 \rfloor \wedge J} \frac{B_r(x)}{\lambda^{2r}} (A\omega)_{r+1} + \sum_{r=(\lfloor \underline{\beta}/2 \rfloor \wedge J)+1}^{\lfloor \underline{\beta}/2 \rfloor} \frac{B_r(x)}{\lambda^{2r}} \sum_{l=0}^{J} \frac{\omega_l}{a_l^{2r}} \\
&= \mu(x) + \mathbb{I}\{\lfloor \underline{\beta}/2 \rfloor \geq J+1\} \frac{B_{J+1}(x)}{\lambda^{2J+2}} \sum_{l=0}^{J} \frac{\omega_l}{a_l^{2J+2}} + O\left(\frac{1}{\lambda^{2J+4}}\right) \\
&= \mu(x) + \mathbb{I}\{2J+2 < \beta\} \frac{\bar{\omega} B_{J+1}(x)}{\lambda^{2J+2}} + O\left(\frac{1}{\lambda^{2J+4}}\right).
\end{aligned}$$

□

Proof (Lemma 5)**.** Firstly, note that with $\sigma_i^2 = \sigma^2(X_i)$ for brevity,

$$\begin{aligned}
\tilde{\Sigma}_{\mathrm{d}}(x) &= \frac{n}{\lambda^d} \operatorname{Var}\left[\sum_{r=0}^{J} \omega_r \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} \frac{Y_i \mathbb{I}\{X_i \in T_{br}(x)\}}{N_{br}(x)} \;\Big|\; \mathbf{X}, \mathbf{T}\right] \\
&= \frac{n}{\lambda^d} \sum_{i=1}^{n} \sum_{r=0}^{J} \sum_{r'=0}^{J} \omega_r \omega_{r'} \frac{1}{B^2} \sum_{b=1}^{B} \sum_{b'=1}^{B} \frac{\mathbb{I}_{ibr}(x) \mathbb{I}_{ib'r'}(x) \sigma_i^2}{N_{br}(x) N_{b'r'}(x)}.
\end{aligned}$$

Part 1: bounding the variance of $\tilde{\Sigma}_{\mathrm{d}}(x)$

$$
\begin{aligned}
\mathrm{Var}\left[\tilde{\Sigma}_{\mathrm{d}}(x)\right] &= \mathrm{Var}\left[\frac{n}{\lambda^d}\sum_{i=1}^n \sum_{r=0}^J \sum_{r'=0}^J \omega_r \omega_{r'} \frac{1}{B^2}\sum_{b=1}^B \sum_{b'=1}^B \frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\sigma_i^2}{N_{br}(x)N_{b'r'}(x)}\right] \\
&\lesssim \frac{n^2}{\lambda^{2d}}\frac{1}{B^4}\,\mathrm{Var}\left[\sum_{i=1}^n \sum_{b=1}^B \sum_{b'=1}^B \frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\sigma_i^2}{N_{br}(x)N_{b'r'}(x)}\right] \\
&\lesssim \frac{n^2}{\lambda^{2d}}\frac{1}{B^4}\mathbb{E}\left[\mathrm{Var}\left[\sum_{i=1}^n \sum_{b=1}^B \sum_{b'=1}^B \frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\sigma_i^2}{N_{br}(x)N_{b'r'}(x)} \;\Big|\; \mathbf{X}\right]\right] \\
&\quad + \frac{n^2}{\lambda^{2d}}\frac{1}{B^4}\,\mathrm{Var}\left[\mathbb{E}\left[\sum_{i=1}^n \sum_{b=1}^B \sum_{b'=1}^B \frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\sigma_i^2}{N_{br}(x)N_{b'r'}(x)} \;\Big|\; \mathbf{X}\right]\right]. \qquad (24)
\end{aligned}
$$

For the first term in (24),

$$
\mathbb{E}\left[\mathrm{Var}\left[\sum_{i=1}^n \sum_{b=1}^B \sum_{b'=1}^B \frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\sigma_i^2}{N_{br}(x)N_{b'r'}(x)} \;\Big|\; \mathbf{X}\right]\right] = \sum_{i=1}^n \sum_{j=1}^n \sum_{b=1}^B \sum_{b'=1}^B \sum_{\tilde{b}=1}^B \sum_{\tilde{b}'=1}^B
$$

$$
\mathbb{E}\left[\sigma_i^2\sigma_j^2\left(\frac{\mathbb{I}_{ibr}(x)}{N_{br}(x)}\frac{\mathbb{I}_{ib'r'}(x)}{N_{b'r'}(x)} - \mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x)}{N_{br}(x)}\frac{\mathbb{I}_{ib'r'}(x)}{N_{b'r'}(x)}\Big|\mathbf{X}\right]\right)\left(\frac{\mathbb{I}_{j\tilde{b}r}(x)}{N_{\tilde{b}r}(x)}\frac{\mathbb{I}_{j\tilde{b}'r'}(x)}{N_{\tilde{b}'r'}(x)} - \mathbb{E}\left[\frac{\mathbb{I}_{j\tilde{b}r}(x)}{N_{\tilde{b}r}(x)}\frac{\mathbb{I}_{j\tilde{b}'r'}(x)}{N_{\tilde{b}'r'}(x)}\Big|\mathbf{X}\right]\right)\right].
$$

Since T_{br} is independent of $T_{b'r'}$ given $\mathbf{X}$, the summands are zero whenever $|\{b, b', \tilde{b}, \tilde{b}'\}| = 4$. Further, by the Cauchy–Schwarz inequality and Lemma 14,

$$
\begin{aligned}
&\frac{n^2}{\lambda^{2d}}\frac{1}{B^4}\mathbb{E}\left[\mathrm{Var}\left[\sum_{i=1}^n \sum_{b=1}^B \sum_{b'=1}^B \frac{\mathbb{I}_{ibr}(x)}{N_{br}(x)}\frac{\mathbb{I}_{ib'r'}(x)}{N_{b'r'}(x)}\sigma^2(X_i) \;\Big|\; \mathbf{X}\right]\right] \\
&\quad\lesssim \frac{n^2}{\lambda^{2d}}\frac{1}{B^3}\sum_{b=1}^B \sum_{b'=1}^B \mathbb{E}\left[\left(\sum_{i=1}^n \frac{\mathbb{I}_{ibr}(x)}{N_{br}(x)}\frac{\mathbb{I}_{ib'r'}(x)}{N_{b'r'}(x)}\right)^2\right] \lesssim \frac{n^2}{\lambda^{2d}}\frac{1}{B^3}\sum_{b=1}^B \sum_{b'=1}^B \mathbb{E}\left[\frac{\mathbb{I}_{br}(x)}{N_{br}(x)}\frac{\mathbb{I}_{b'r'}(x)}{N_{b'r'}(x)}\right] \\
&\quad\lesssim \frac{n^2}{\lambda^{2d}}\frac{1}{B^3}\left(B^2\frac{\lambda^{2d}}{n^2} + B\frac{\lambda^{2d}(\log n)^d}{n^2}\right) \lesssim \frac{1}{B} + \frac{(\log n)^d}{B^2} \lesssim \frac{1}{B}.
\end{aligned}
$$

For the second term in (24), the random variable inside the variance is a nonlinear function of the i.i.d. variables X_i, so we apply the Efron–Stein inequality (Efron and Stein, 1981). Let $\hat{X}_{ij} = X_i$ if $i \neq j$ and be an independent copy of X_j, denoted $\hat{X}_j$, if $i = j$, and define $\sigma_{ij}^2 = \sigma^2(\hat{X}_{ij})$. Write $\hat{\mathbb{I}}_{ijbr}(x) = \mathbb{I}\{\hat{X}_{ij} \in T_{br}(x)\}$ and $\hat{\mathbb{I}}_{jbr}(x) = \mathbb{I}\{\hat{X}_j \in T_{br}(x)\}$, and also $\hat{N}_{jbr}(x) = \sum_{i=1}^n \hat{\mathbb{I}}_{ijbr}(x)$, We use the leave-one-out notation $N_{-jbr}(x) = \sum_{i \neq j} \mathbb{I}_{ibr}(x)$ and also write $N_{-jbr \cap b'r'} = \sum_{i \neq j} \mathbb{I}_{ibr}(x) \mathbb{I}_{ib'r'}(x)$.

$$
\begin{aligned}
&\frac{n^2}{\lambda^{2d}} \frac{1}{B^4} \operatorname{Var}\left[\mathbb{E}\left[\sum_{i=1}^n \sum_{b=1}^B \sum_{b'=1}^B \frac{\mathbb{I}_{ibr}(x)}{N_{br}(x)} \frac{\mathbb{I}_{ib'r'}(x)}{N_{b'r'}(x)} \sigma_i^2 \mid \mathbf{X}\right]\right] \lesssim \frac{n^2}{\lambda^{2d}} \operatorname{Var}\left[\mathbb{E}\left[\sum_{i=1}^n \frac{\mathbb{I}_{ibr}(x)}{N_{br}(x)} \frac{\mathbb{I}_{ib'r'}(x)}{N_{b'r'}(x)} \sigma_i^2 \mid \mathbf{X}\right]\right] \\
&\quad \lesssim \frac{n^2}{\lambda^{2d}} \sum_{j=1}^n \mathbb{E}\left[\left(\sum_{i=1}^n \left(\frac{\mathbb{I}_{ibr}(x) \mathbb{I}_{ib'r'}(x) \sigma_i^2}{N_{br}(x) N_{b'r'}(x)} - \frac{\hat{\mathbb{I}}_{ijbr}(x) \hat{\mathbb{I}}_{ijb'r'}(x) \hat{\sigma}_{ij}^2}{\hat{N}_{jbr}(x) \hat{N}_{jb'r'}(x)}\right)\right)^2\right] \\
&\quad \lesssim \frac{n^2}{\lambda^{2d}} \sum_{j=1}^n \mathbb{E}\left[\left(\left|\frac{1}{N_b(x) N_{b'r'}(x)} - \frac{1}{\hat{N}_{jbr}(x) \hat{N}_{jb'r'}(x)}\right| \sum_{i \neq j} \mathbb{I}_{ibr}(x) \mathbb{I}_{ib'r'}(x) \sigma_i^2\right)^2\right] \\
&\qquad + \frac{n^2}{\lambda^{2d}} \sum_{j=1}^n \mathbb{E}\left[\left(\frac{\mathbb{I}_{jbr}(x) \mathbb{I}_{jb'r'}(x) \sigma_j^2}{N_{br}(x) N_{b'r'}(x)} - \frac{\hat{\mathbb{I}}_{jbr}(x) \hat{\mathbb{I}}_{jb'r'}(x) \hat{\sigma}_{jj}^2}{\hat{N}_{jbr}(x) \hat{N}_{jb'r'}(x)}\right)^2\right] \\
&\quad \lesssim \frac{n^2}{\lambda^{2d}} \sum_{j=1}^n \mathbb{E}\left[N_{-jbr \cap b'r'}(x)^2 \left|\frac{1}{N_{br}(x) N_{b'r'}(x)} - \frac{1}{\hat{N}_{jbr}(x) \hat{N}_{jb'r'}(x)}\right|^2 + \frac{\mathbb{I}_{jbr}(x) \mathbb{I}_{jb'r'}(x)}{N_{br}(x)^2 N_{b'r'}(x)^2}\right]. \qquad (25)
\end{aligned}
$$

For the first term in (25), note that since $|N_{br}(x) - \hat{N}_{jbr}(x)| \leq \mathbb{I}_{jbr}(x) + \hat{\mathbb{I}}_{jbr}(x)$ and similarly $|N_{b'r'}(x) - \hat{N}_{jb'r'}(x)| \leq \mathbb{I}_{jb'r'}(x) + \hat{\mathbb{I}}_{jb'r'}(x)$,

$$
\begin{aligned}
&\left|\frac{1}{N_{br}(x) N_{b'r'}(x)} - \frac{1}{\hat{N}_{jbr}(x) \hat{N}_{jb'r'}(x)}\right| \\
&\quad \leq \frac{1}{N_{br}(x)} \left|\frac{1}{N_{b'r'}(x)} - \frac{1}{\hat{N}_{jb'r'}(x)}\right| + \frac{1}{\hat{N}_{jb'r'}(x)} \left|\frac{1}{N_{br}(x)} - \frac{1}{\hat{N}_{jbr}(x)}\right| \\
&\quad \leq \frac{\mathbb{I}_{jbr}(x) + \hat{\mathbb{I}}_{jbr}(x)}{N_{-jbr}(x) N_{-jb'r'}(x)^2} + \frac{\mathbb{I}_{jb'r'}(x) + \hat{\mathbb{I}}_{jb'r'}(x)}{N_{-jb'r'}(x) N_{-jbr}(x)^2}.
\end{aligned}
$$

Therefore by Lemma 14,

$$
\begin{aligned}
&\frac{n^2}{\lambda^{2d}} \frac{1}{B^4} \operatorname{Var}\left[\mathbb{E}\left[\sum_{i=1}^n \sum_{b=1}^B \sum_{b'=1}^B \frac{\mathbb{I}_{ibr}(x) \mathbb{I}_{ib'r'}(x) \sigma^2(X_i)}{N_{br}(x) N_{b'r'}(x)} \mid \mathbf{X}\right]\right] \\
&\quad \lesssim \frac{n^2}{\lambda^{2d}} \sum_{j=1}^n \mathbb{E}\left[\frac{\mathbb{I}_{jbr}(x) \mathbb{I}_{br \cap b'r'}(x)}{N_{br}(x)^2 N_{b'r'}(x)^2}\right] \lesssim \frac{n^2}{\lambda^{2d}} \mathbb{E}\left[\frac{\mathbb{I}_{br}(x) \mathbb{I}_{b'r'}(x)}{N_{br}(x)^{3/2} N_{b'r'}(x)^{3/2}}\right] \lesssim \frac{n^2}{\lambda^{2d}} \frac{\lambda^{3d}}{n^3} \lesssim \frac{\lambda^d}{n}.
\end{aligned}
$$

We deduce that

$$
\operatorname{Var}\left[\tilde{\Sigma}_{\mathrm{d}}(x)\right] \lesssim \frac{1}{B} + \frac{\lambda^d}{n}.
$$

Part 2: controlling the expectation of $\tilde{\Sigma}_{\mathrm{d}}(x)$

$$
\mathbb{E}\left[\tilde{\Sigma}_{\mathrm{d}}(x)\right] = \frac{n}{\lambda^d} \sum_{i=1}^n \sum_{r=0}^J \sum_{r'=0}^J \omega_r \omega_{r'} \frac{1}{B^2} \sum_{b=1}^B \sum_{b'=1}^B \mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x) \mathbb{I}_{ib'r'}(x) \sigma^2(X_i)}{N_{br}(x) N_{b'r'}(x)}\right].
$$

Firstly, by Lemma 14, the diagonal terms in the forest are

$$
\left|\frac{n}{\lambda^d} \sum_{i=1}^n \sum_{r=0}^J \sum_{r'=0}^J \omega_r \omega_{r'} \frac{1}{B^2} \sum_{b=1}^B \mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x) \mathbb{I}_{ibr'}(x) \sigma^2(X_i)}{N_{br}(x) N_{br'}(x)}\right]\right| \lesssim \frac{n}{\lambda^d} \frac{1}{B} \mathbb{E}\left[\frac{\mathbb{I}_{br}(x)}{N_{br}(x)}\right] \lesssim \frac{1}{B},
$$

so it suffices to take $b \neq b'$ since

$$\mathbb{E}\left[\tilde{\Sigma}_{\mathrm{d}}(x)\right] = \frac{n^2}{\lambda^d} \sum_{r=0}^{J} \sum_{r'=0}^{J} \omega_r \omega_{r'} \, \mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\sigma^2(X_i)}{N_{br}(x)N_{b'r'}(x)}\right] + O\left(\frac{1}{B}\right).$$

Next, note that

$$\mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\sigma^2(X_i)}{N_{br}(x)N_{b'r'}(x)}\right] = \sigma^2(x)\mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)}{N_{br}(x)N_{b'r'}(x)}\right] + \mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\big(\sigma^2(X_i) - \sigma^2(x)\big)}{N_{br}(x)N_{b'r'}(x)}\right].$$

Since $\sigma^2 \in \mathcal{H}^{\beta_\sigma}$, we have by Lemma 9 and Lemma 14 that

$$\frac{n^2}{\lambda^d}\mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\big|\sigma^2(X_i) - \sigma^2(x)\big|}{N_{br}(x)N_{b'r'}(x)}\right] \lesssim \frac{n^2}{\lambda^d}\frac{1}{n}\mathbb{E}\left[\frac{\mathbb{I}_{b'r'}(x)\max_j |T_{br}(x)_j|}{N_{b'r'}(x)}\right] \lesssim \frac{1}{\lambda^{1\wedge\beta_\sigma}}.$$

Therefore

$$\frac{n^2}{\lambda^d}\mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\sigma^2(X_i)}{N_{br}(x)N_{b'r'}(x)}\right] = \sigma^2(x)\frac{n^2}{\lambda^d}\mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)}{N_{br}(x)N_{b'r'}(x)}\right] + O\left(\frac{1}{\lambda^{1\wedge\beta_\sigma}}\right).$$

Next, by conditioning on T_{br}, $T_{b'r'}$, $N_{-ibr}(x)$, and $N_{-ib'r'}(x)$,

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)}{N_{br}(x)N_{b'r'}(x)}\right] &= \mathbb{E}\left[\frac{\int_{T_{br}(x)\cap T_{b'r'}(x)} f(\xi)\,\mathrm{d}\xi}{(N_{-ibr}(x)+1)(N_{-ib'r'}(x)+1)}\right] \\
&= f(x)\,\mathbb{E}\left[\frac{|T_{br}(x)\cap T_{b'r'}(x)|}{(N_{-ibr}(x)+1)(N_{-ib'r'}(x)+1)}\right] + \mathbb{E}\left[\frac{\int_{T_{br}(x)\cap T_{b'r'}(x)} (f(\xi) - f(x))\,\mathrm{d}\xi}{(N_{-ibr}(x)+1)(N_{-ib'r'}(x)+1)}\right] \\
&= f(x)\,\mathbb{E}\left[\frac{|T_{br}(x)\cap T_{b'r'}(x)|}{(N_{-ibr}(x)+1)(N_{-ib'r'}(x)+1)}\right] + O\left(\frac{\lambda^d}{n^2}\frac{1}{\lambda^{1\wedge\beta_f}}\right)
\end{aligned}$$

by an argument based on Lemma 9, the Hölder property of $f(x)$, and the proof of Lemma 14. Hence

$$\begin{aligned}
\frac{n^2}{\lambda^d}\mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\sigma^2(X_i)}{N_{br}(x)N_{b'r'}(x)}\right] &= \sigma^2(x)f(x)\frac{n^2}{\lambda^d}\mathbb{E}\left[\frac{|T_{br}(x)\cap T_{b'r'}(x)|}{(N_{-ibr}(x)+1)(N_{-ib'r'}(x)+1)}\right] \\
&\quad + O\left(\frac{1}{\lambda^{1\wedge\beta_f\wedge\beta_\sigma}}\right).
\end{aligned}$$

Apply Lemma 15 to approximate the expectation with $N_{-ib'r'\setminus br}(x) = \sum_{j\neq i}\mathbb{I}\{X_j \in T_{b'r'}(x)\setminus T_{br}(x)\}$:

$$\begin{aligned}
&\mathbb{E}\left[\frac{|T_{br}(x)\cap T_{b'r'}(x)|}{(N_{-ibr}(x)+1)(N_{-ib'r'}(x)+1)}\right] \\
&\quad = \mathbb{E}\left[\frac{|T_{br}(x)\cap T_{b'r'}(x)|}{N_{-ibr}(x)+1}\,\mathbb{E}\left[\frac{1}{N_{-ib'r'\cap br}(x) + N_{-ib'r'\setminus br}(x) + 1} \;\Big|\; \mathbf{T}, N_{-ib'r'\cap br}(x), N_{-ibr\setminus b'r'}(x)\right]\right].
\end{aligned}$$

Now conditional on $\mathbf{T}$, $N_{-ib'r'\cap br}(x)$, and $N_{-ibr\setminus b'r'}(x)$,

$$N_{-ib'r'\setminus br}(x) \sim \mathrm{Bin}\left(n - 1 - N_{-ibr}(x),\; \frac{\int_{T_{b'r'}(x)\setminus T_{br}(x)} f(\xi)\,\mathrm{d}\xi}{1 - \int_{T_{br}(x)} f(\xi)\,\mathrm{d}\xi}\right).$$

We bound these parameters above and below. Firstly, by applying Lemma 10 with $B = 1$, we have

$$\mathbb{P}\left(N_{-ibr}(x) > t^{d+1}\frac{n}{\lambda^d}\right) \leq 4de^{-t/(4\|f\|_\infty(1+1/a_r))} \leq e^{-t/C}$$

for some $C > 0$ and sufficiently large t. Next, note if f is β_f-Hölder with constant L, by Lemma 9,

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{\int_{T_{b'r'}(x)\setminus T_{br}(x)} f(\xi)\,\mathrm{d}\xi}{1-\int_{T_{br}(x)} f(\xi)\,\mathrm{d}\xi} - f(x)|T_{b'r'}(x)\setminus T_{br}(x)|\right| > 2L\,|T_{b'r'}(x)\setminus T_{br}(x)|\sum_{j=1}^{d}|T_{b'r'}(x)_j|^{1\wedge\beta_f}\right)\\
&\leq \mathbb{P}\left(\frac{\int_{T_{b'r'}(x)\setminus T_{br}(x)} |f(\xi)-f(x)|\,\mathrm{d}\xi}{1-\int_{T_{br}(x)} f(\xi)\,\mathrm{d}\xi} > 2L\,|T_{b'r'}(x)\setminus T_{br}(x)|\sum_{j=1}^{d}|T_{b'r'}(x)_j|^{1\wedge\beta_f}\right)\\
&\leq \mathbb{P}\left(\frac{1}{1-\int_{T_{br}(x)} f(\xi)\,\mathrm{d}\xi} > 2\right) \leq \mathbb{P}\left(\int_{T_{br}(x)} f(\xi)\,\mathrm{d}\xi > \frac{1}{2}\right)\\
&\leq \mathbb{P}\left(|T_{br}(x)| > \frac{1}{2\|f\|_\infty}\right) \lesssim \mathbb{P}\left(\max_{1\leq j\leq d}|T_{br}(x)_j|^{1\wedge\beta_f} > \frac{1}{2\|f\|_\infty}\right) \lesssim 2de^{-\lambda/(4\|f\|_\infty)} \lesssim e^{-\lambda/C},
\end{aligned}$$

increasing C as necessary. Thus with probability at least $1 - e^{-t/C} - e^{-\lambda/C}$,

$$\begin{aligned}
N_{-ib'r'\setminus br}(x) &\leq \mathrm{Bin}\left(n,\ |T_{b'r'}(x)\setminus T_{br}(x)|\left(f(x) + 2L\sum_{j=1}^{d}|T_{b'r'}(x)_j|^{1\wedge\beta_f}\right)\right)\\
N_{-ib'r'\setminus br}(x) &\geq \mathrm{Bin}\left(n\left(1-\frac{t^{d+1}}{\lambda^d}-\frac{1}{n}\right),\ |T_{b'r'}(x)\setminus T_{br}(x)|\left(f(x) - 2L\sum_{j=1}^{d}|T_{b'r'}(x)_j|^{1\wedge\beta_f}\right)\right).
\end{aligned}$$

So by Lemma 15 conditionally on $\mathbf{T}$, $N_{-ib'r'\cap br}(x)$, and $N_{-ibr\setminus b'r'}(x)$, taking $t = 4C\log n$ and recalling $\lambda \gtrsim (\log n)^3$, with probability at least $1 - n^{-3}$,

$$\begin{aligned}
&\Bigg|\mathbb{E}\left[\frac{1}{N_{-ib'r'\cap br}(x) + N_{-ib'r'\setminus br}(x) + 1}\ \Big|\ \mathbf{T}, N_{-ib'r'\cap br}(x), N_{-ibr\setminus b'r'}(x)\right]\\
&\qquad - \frac{1}{N_{-ib'r'\cap br}(x) + nf(x)|T_{b'r'}(x)\setminus T_{br}(x)| + 1}\Bigg| \lesssim \frac{1 + n|T_{b'r'}(x)_j|^{1\wedge\beta_f}|T_{b'r'}(x)\setminus T_{br}(x)|}{(N_{-ib'r'\cap br}(x) + n|T_{b'r'}(x)\setminus T_{br}(x)| + 1)^2}.
\end{aligned}$$

Therefore by the same approach as the proof of Lemma 12,

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{|T_{br}(x)\cap T_{b'r'}(x)|}{(N_{-ibr}(x)+1)(N_{-ib'r'}(x)+1)} - \frac{|T_{br}(x)\cap T_{b'r'}(x)|}{(N_{-ibr}(x)+1)(N_{-ib'r'\cap br}(x) + nf(x)|T_{b'r'}(x)\setminus T_{br}(x)| + 1)}\right]\right|\\
&\lesssim \mathbb{E}\left[\frac{|T_{br}(x)\cap T_{b'r'}(x)|}{N_{-ibr}(x)+1}\frac{1 + n|T_{b'r'}(x)_j|^{1\wedge\beta_f}|T_{b'r'}(x)\setminus T_{br}(x)|}{(N_{-ib'r'\cap br}(x) + n|T_{b'r'}(x)\setminus T_{br}(x)| + 1)^2}\right] + n^{-3}\\
&\lesssim \mathbb{E}\left[\frac{|T_{br}(x)\cap T_{b'r'}(x)|}{n|T_{br}(x)|+1}\frac{1 + n|T_{b'r'}(x)_j|^{1\wedge\beta_f}|T_{b'r'}(x)\setminus T_{br}(x)|}{(n|T_{b'r'}(x)|+1)^2}\right] + n^{-3}\\
&\lesssim \mathbb{E}\left[\frac{1}{n}\frac{1}{(n|T_{br}(x)|+1)(n|T_{b'r'}(x)|+1)} + \frac{1}{n}\frac{|T_{b'r'}(x)_j|^{1\wedge\beta_f}}{n|T_{b'r'}(x)|+1}\right] + n^{-3}\\
&\lesssim \frac{\lambda^{2d}}{n^3} + \frac{1}{n\lambda^{1\wedge\beta_f}}\frac{\lambda^d}{n} \lesssim \frac{\lambda^d}{n^2}\left(\frac{\lambda^d}{n} + \frac{1}{\lambda^{1\wedge\beta_f}}\right).
\end{aligned}$$

Now apply the same argument to the other term in the expectation, to see that

$$\left|\mathbb{E}\left[\frac{1}{N_{-ibr\cap b'r'}(x)+N_{-ibr\setminus b'r'}(x)+1} \;\Big|\; \mathbf{T}, N_{-ibr\cap b'r'}(x), N_{-ib'r'\setminus br}(x)\right] \right.$$
$$\left. -\frac{1}{N_{-ibr\cap b'r'}(x)+nf(x)|T_{br}(x)\setminus T_{b'r'}(x)|+1}\right| \lesssim \frac{1+n|T_{br}(x)_j|^{1\wedge\beta_f}|T_{br}(x)\setminus T_{b'r'}(x)|}{(N_{-ibr\cap b'r'}(x)+n|T_{br}(x)\setminus T_{b'r'}(x)|+1)^2}.$$

with probability at least $1-n^{-3}-e^{-\lambda/C}$, and so likewise again with $t=4C\log n$,

$$\frac{n^2}{\lambda^d}\left|\mathbb{E}\left[\frac{|T_{br}(x)\cap T_{b'r'}(x)|}{N_{-ibr}(x)+1}\frac{1}{N_{-ib'r'\cap br}(x)+nf(x)|T_{b'r'}(x)\setminus T_{br}(x)|+1}\right]\right.$$
$$\left. -\mathbb{E}\left[\frac{|T_{br}(x)\cap T_{b'r'}(x)|}{N_{-ibr\cap b'r'}(x)+nf(x)|T_{br}(x)\setminus T_{b'r'}(x)|+1}\frac{1}{N_{-ib'r'\cap br}(x)+nf(x)|T_{b'r'}(x)\setminus T_{br}(x)|+1}\right]\right|$$
$$\lesssim \frac{n^2}{\lambda^d}\mathbb{E}\left[\frac{1+n|T_{br}(x)_j|^{1\wedge\beta_f}|T_{br}(x)\setminus T_{b'r'}(x)|}{(N_{-ibr\cap b'r'}(x)+n|T_{br}(x)\setminus T_{b'r'}(x)|+1)^2}\frac{|T_{br}(x)\cap T_{b'r'}(x)|}{N_{-ib'r'\cap br}(x)+nf(x)|T_{b'r'}(x)\setminus T_{br}(x)|+1}\right]$$
$$+\frac{n^2}{\lambda^d}n^{-3}\lesssim \frac{\lambda^d}{n}+\frac{1}{\lambda^{1\wedge\beta_f}}.$$

Thus far we have proven that

$$\frac{n^2}{\lambda^d}\mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\sigma^2(X_i)}{N_{br}(x)N_{b'r'}(x)}\right]=\sigma^2(x)f(x)\frac{n^2}{\lambda^d}$$
$$\times\mathbb{E}\left[\frac{|T_{br}(x)\cap T_{b'r'}(x)|}{N_{-ibr\cap b'r'}(x)+nf(x)|T_{br}(x)\setminus T_{b'r'}(x)|+1}\frac{1}{N_{-ib'r'\cap br}(x)+nf(x)|T_{b'r'}(x)\setminus T_{br}(x)|+1}\right]$$
$$+O\left(\frac{1}{\lambda^{1\wedge\beta_f\wedge\beta_\sigma}}+\frac{\lambda^d}{n}\right).$$

Next we remove the $N_{-ibr\cap b'r'}(x)$ terms. As before, with probability at least $1-e^{-t/C}-e^{-\lambda/C}$, conditional on $\mathbf{T}$,

$$N_{-ibr\cap b'r'}(x)\le \mathrm{Bin}\left(n,\ |T_{br}(x)\cap T_{b'r'}(x)|\left(f(x)+2L\sum_{j=1}^d|T_{br}(x)_j|^{1\wedge\beta_f}\right)\right),$$
$$N_{-ibr\cap b'r'}(x)\ge \mathrm{Bin}\left(n\left(1-\frac{t^{d+1}}{\lambda^d}-\frac{1}{n}\right),\ |T_{br}(x)\cap T_{b'r'}(x)|\left(f(x)-2L\sum_{j=1}^d|T_{br}(x)_j|^{1\wedge\beta_f}\right)\right).$$

So by Lemma 15 conditionally on $\mathbf{T}$, with $t=4C\log n$ and with probability at least $1-n^{-3}$,

$$\left|\mathbb{E}\left[\frac{1}{N_{-ibr\cap b'r'}(x)+nf(x)|T_{br}(x)\setminus T_{b'r'}(x)|+1}\frac{1}{N_{-ib'r'\cap br}(x)+nf(x)|T_{b'r'}(x)\setminus T_{br}(x)|+1}\;\Big|\;\mathbf{T}\right]\right.$$
$$\left.-\frac{1}{nf(x)|T_{br}(x)|+1}\frac{1}{nf(x)|T_{b'r'}(x)|+1}\right|$$
$$\lesssim\frac{1+n|T_{br}(x)_j|^{1\wedge\beta_f}|T_{br}(x)\cap T_{b'r'}(x)|}{(n|T_{br}(x)|+1)(n|T_{b'r'}(x)|+1)}\left(\frac{1}{n|T_{br}(x)|+1}+\frac{1}{n|T_{b'r'}(x)|+1}\right).$$

Now by Lemma 13,

$$
\begin{aligned}
&\frac{n^2}{\lambda^d}\left|\mathbb{E}\left[\frac{|T_{br}(x)\cap T_{b'r'}(x)|}{N_{-ibr\cap b'r'}(x)+nf(x)|T_{br}(x)\setminus T_{b'r'}(x)|+1}\frac{1}{N_{-ib'r'\cap br}(x)+nf(x)|T_{b'r'}(x)\setminus T_{br}(x)|+1}\right]\right. \\
&\qquad\left.-\mathbb{E}\left[\frac{|T_{br}(x)\cap T_{b'r'}(x)|}{nf(x)|T_{br}(x)|+1}\frac{1}{nf(x)|T_{b'r'}(x)|+1}\right]\right| \\
&\qquad\lesssim \frac{n^2}{\lambda^d}\mathbb{E}\left[\frac{1+n|T_{br}(x)_j|^{1\wedge\beta_f}|T_{br}(x)\cap T_{b'r'}(x)|}{(n|T_{br}(x)|+1)(n|T_{b'r'}(x)|+1)}\left(\frac{|T_{br}(x)\cap T_{b'r'}(x)|}{n|T_{br}(x)|+1}+\frac{|T_{br}(x)\cap T_{b'r'}(x)|}{n|T_{b'r'}(x)|+1}\right)\right]+\frac{1}{n\lambda^d} \\
&\qquad\lesssim \frac{n^2}{\lambda^d}\frac{1}{n^3}\mathbb{E}\left[\frac{1+n|T_{br}(x)_j|^{1\wedge\beta_f}|T_{br}(x)\cap T_{b'r'}(x)|}{|T_{br}(x)||T_{b'r'}(x)|}\right]+\frac{1}{n\lambda^d} \\
&\qquad\lesssim \frac{1}{n\lambda^d}\mathbb{E}\left[\frac{1}{|T_{br}(x)||T_{b'r'}(x)|}\right]+\frac{1}{\lambda^d}\mathbb{E}\left[\frac{|T_{br}(x)_j|^{1\wedge\beta_f}}{|T_{b'r'}(x)|}\right]+\frac{1}{n\lambda^d}\lesssim\frac{\lambda^d}{n}+\frac{1}{\lambda^{1\wedge\beta_f}}.
\end{aligned}
$$

This allows us to deduce that

$$
\begin{aligned}
\frac{n^2}{\lambda^d}\mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\sigma^2(X_i)}{N_{br}(x)N_{b'r'}(x)}\right] &= \sigma^2(x)f(x)\frac{n^2}{\lambda^d}\mathbb{E}\left[\frac{|T_{br}(x)\cap T_{b'r'}(x)|}{(nf(x)|T_{br}(x)|+1)(nf(x)|T_{b'r'}(x)|+1)}\right] \\
&\quad+O\left(\frac{1}{\lambda^{1\wedge\beta_f\wedge\beta_\sigma}}+\frac{\lambda^d}{n}\right),
\end{aligned}
$$

and so

$$
\begin{aligned}
\mathbb{E}\left[\tilde{\Sigma}_{\mathrm{d}}(x)\right] &= \sigma^2(x)f(x)\frac{n^2}{\lambda^d}\sum_{r=0}^{J}\sum_{r'=0}^{J}\omega_r\omega_{r'}\mathbb{E}\left[\frac{|T_{br}(x)\cap T_{b'r'}(x)|}{(nf(x)|T_{br}(x)|+1)(nf(x)|T_{b'r'}(x)|+1)}\right] \\
&\quad+O\left(\frac{1}{\lambda^{1\wedge\beta_f\wedge\beta_\sigma}}+\frac{\lambda^d}{n}+\frac{1}{B}\right).
\end{aligned}
$$

Part 3: calculating the limiting variance $\Sigma_{\mathrm{d}}(x)$
Now that we have reduced the variance to an expression only involving the sizes of Mondrian cells, we can exploit their exact distribution to compute this expectation. Recall from Mourtada et al. (2020, Proposition 1) that we can write

$$
\begin{aligned}
|T_{br}(x)| &= \prod_{j=1}^{d}\left(\frac{E_{1j}}{a_r\lambda}\wedge x_j+\frac{E_{2j}}{a_r\lambda}\wedge(1-x_j)\right), \qquad |T_{b'r'}(x)|=\prod_{j=1}^{d}\left(\frac{E_{3j}}{a_{r'}\lambda}\wedge x_j+\frac{E_{4j}}{a_{r'}\lambda}\wedge(1-x_j)\right), \\
|T_{br}(x)\cap T_{b'r'}(x)| &= \prod_{j=1}^{d}\left(\frac{E_{1j}}{a_r\lambda}\wedge\frac{E_{3j}}{a_{r'}\lambda}\wedge x_j+\frac{E_{2j}}{a_r\lambda}\wedge\frac{E_{4j}}{a_{r'}\lambda}\wedge(1-x_j)\right)
\end{aligned}
$$

where E_{1j}, E_{2j}, E_{3j}, and E_{4j} are independent and $\mathrm{Exp}(1)$. Define their non-truncated versions as

$$
\begin{aligned}
|\tilde{T}_{br}(x)| &= a_r^{-d}\lambda^{-d}\prod_{j=1}^{d}\left(E_{1j}+E_{2j}\right), \qquad |\tilde{T}_{b'r'}(x)|=a_{r'}^{-d}\lambda^{-d}\prod_{j=1}^{d}\left(E_{3j}+E_{4j}\right), \\
|\tilde{T}_{br}(x)\cap\tilde{T}_{b'r'}(x)| &= \lambda^{-d}\prod_{j=1}^{d}\left(\frac{E_{1j}}{a_r}\wedge\frac{E_{3j}}{a_{r'}}+\frac{E_{2j}}{a_r}\wedge\frac{E_{4j}}{a_{r'}}\right),
\end{aligned}
$$

and note that

$$
\begin{aligned}
&\mathbb{P}\Big(\big(\tilde{T}_{br}(x), \tilde{T}_{b'r'}(x), \tilde{T}_{br}(x) \cap T_{b'r'}(x)\big) \neq \big(T_{br}(x), T_{b'r'}(x), T_{br}(x) \cap T_{b'r'}(x)\big)\Big) \\
&\quad \leq \sum_{j=1}^{d} \big(\mathbb{P}(E_{1j} \geq a_r \lambda x_j) + \mathbb{P}(E_{3j} \geq a_{r'} \lambda x_j) + \mathbb{P}(E_{2j} \geq a_r \lambda(1-x_j)) + \mathbb{P}(E_{4j} \geq a_{r'} \lambda(1-x_j))\big) \\
&\quad \leq e^{-C\lambda}
\end{aligned}
$$

for some $C > 0$ and sufficiently large λ. Hence by the Cauchy–Schwarz inequality and Lemma 13,

$$
\begin{aligned}
&\frac{n^2}{\lambda^d} \left| \mathbb{E}\left[\frac{|T_{br}(x) \cap T_{b'r'}(x)|}{nf(x)|T_{br}(x)|+1} \frac{1}{nf(x)|T_{b'r'}(x)|+1}\right] - \mathbb{E}\left[\frac{|\tilde{T}_{br}(x) \cap T_{b'r'}(x)|}{nf(x)|\tilde{T}_{br}(x)|+1} \frac{1}{nf(x)|\tilde{T}_{b'r'}(x)|+1}\right]\right| \\
&\quad \lesssim \frac{n^2}{\lambda^d} e^{-C\lambda} \lesssim \frac{1}{n\lambda^d}
\end{aligned}
$$

as $\lambda \gtrsim (\log n)^3$. Therefore

$$
\begin{aligned}
\frac{n^2}{\lambda^d} \mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x) \mathbb{I}_{ib'r'}(x) \sigma^2(X_i)}{N_{br}(x) N_{b'r'}(x)}\right] &= \sigma^2(x) f(x) \frac{n^2}{\lambda^d} \mathbb{E}\left[\frac{|\tilde{T}_{br}(x) \cap \tilde{T}_{b'r'}(x)|}{(nf(x)|\tilde{T}_{br}(x)|+1)(nf(x)|\tilde{T}_{b'r'}(x)|+1)}\right] \\
&\quad + O\left(\frac{1}{\lambda^{1 \wedge \beta_f \wedge \beta_\sigma}} + \frac{\lambda^d}{n}\right).
\end{aligned}
$$

Now we remove the superfluous units in the denominators. Firstly, by independence of the trees,

$$
\begin{aligned}
&\frac{n^2}{\lambda^d} \left| \mathbb{E}\left[\frac{|\tilde{T}_{br}(x) \cap \tilde{T}_{b'r'}(x)|}{(nf(x)|\tilde{T}_{br}(x)|+1)(nf(x)|\tilde{T}_{b'r'}(x)|+1)}\right] - \mathbb{E}\left[\frac{|\tilde{T}_{br}(x) \cap \tilde{T}_{b'r'}(x)|}{(nf(x)|\tilde{T}_{br}(x)|+1)(nf(x)|\tilde{T}_{b'r'}(x)|)}\right]\right| \\
&\quad \lesssim \frac{n^2}{\lambda^d} \mathbb{E}\left[\frac{|\tilde{T}_{br}(x) \cap \tilde{T}_{b'r'}(x)|}{n|\tilde{T}_{br}(x)|} \frac{1}{n^2|\tilde{T}_{b'r'}(x)|^2}\right] \lesssim \frac{1}{n\lambda^d} \mathbb{E}\left[\frac{1}{|T_{br}(x)|}\right] \mathbb{E}\left[\frac{1}{|T_{b'r'}(x)|}\right] \lesssim \frac{\lambda^d}{n}.
\end{aligned}
$$

Secondly, we have in exactly the same manner that

$$
\frac{n^2}{\lambda^d} \left| \mathbb{E}\left[\frac{|\tilde{T}_{br}(x) \cap T_{b'r'}(x)|}{(nf(x)|\tilde{T}_{br}(x)|+1)(nf(x)|\tilde{T}_{b'r'}(x)|)}\right] - \mathbb{E}\left[\frac{|\tilde{T}_{br}(x) \cap T_{b'r'}(x)|}{n^2 f(x)^2 |\tilde{T}_{br}(x)||\tilde{T}_{b'r'}(x)|}\right]\right| \lesssim \frac{\lambda^d}{n}.
$$

Therefore

$$
\frac{n^2}{\lambda^d} \mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x) \mathbb{I}_{ib'r'}(x) \sigma^2(X_i)}{N_{br}(x) N_{b'r'}(x)}\right] = \frac{\sigma^2(x)}{f(x)} \frac{1}{\lambda^d} \mathbb{E}\left[\frac{|\tilde{T}_{br}(x) \cap \tilde{T}_{b'r'}(x)|}{|\tilde{T}_{br}(x)||\tilde{T}_{b'r'}(x)|}\right] + O\left(\frac{1}{\lambda^{1 \wedge \beta_f \wedge \beta_\sigma}} + \frac{\lambda^d}{n}\right).
$$

It remains to compute this integral. By independence over $1 \le j \le d$,

$$
\begin{aligned}
&\mathbb{E}\left[\frac{|\tilde{T}_{br}(x) \cap \tilde{T}_{b'r'}(x)|}{|\tilde{T}_{br}(x)||\tilde{T}_{b'r'}(x)|}\right] \\
&\quad = a_r^d a_{r'}^d \lambda^d \prod_{j=1}^d \mathbb{E}\left[\frac{(E_{1j}/a_r) \wedge (E_{3j}/a_{r'}) + (E_{2j}a_r) \wedge (E_{4j}/a_{r'})}{(E_{1j}+E_{2j})(E_{3j}+E_{4j})}\right] \\
&\quad = 2^d a_r^d a_{r'}^d \lambda^d \prod_{j=1}^d \mathbb{E}\left[\frac{(E_{1j}/a_r) \wedge (E_{3j}/a_{r'})}{(E_{1j}+E_{2j})(E_{3j}+E_{4j})}\right] \\
&\quad = 2^d a_r^d a_{r'}^d \lambda^d \prod_{j=1}^d \int_0^\infty \int_0^\infty \int_0^\infty \int_0^\infty \frac{(t_1/a_r) \wedge (t_3/a_{r'})}{(t_1+t_2)(t_3+t_4)} e^{-t_1-t_2-t_3-t_4} \,\mathrm{d}t_1 \,\mathrm{d}t_2 \,\mathrm{d}t_3 \,\mathrm{d}t_4 \\
&\quad = 2^d a_r^d a_{r'}^d \lambda^d \prod_{j=1}^d \int_0^\infty \int_0^\infty ((t_1/a_r) \wedge (t_3/a_{r'})) e^{-t_1-t_3} \left(\int_0^\infty \frac{e^{-t_2}}{t_1+t_2} \,\mathrm{d}t_2\right) \left(\int_0^\infty \frac{e^{-t_4}}{t_3+t_4} \,\mathrm{d}t_4\right) \mathrm{d}t_1 \,\mathrm{d}t_3 \\
&\quad = 2^d a_r^d a_{r'}^d \lambda^d \prod_{j=1}^d \int_0^\infty \int_0^\infty ((t/a_r) \wedge (s/a_{r'})) \Gamma(0,t)\Gamma(0,s) \,\mathrm{d}t \,\mathrm{d}s,
\end{aligned}
$$

where we used $\int_0^\infty \frac{e^{-t}}{a+t} \,\mathrm{d}t = e^a \Gamma(0,a)$ with $\Gamma(0,a) = \int_a^\infty \frac{e^{-t}}{t} \,\mathrm{d}t$ the upper incomplete gamma function. Now

$$
\begin{aligned}
&2 \int_0^\infty \int_0^\infty ((t/a_r) \wedge (s/a_{r'})) \Gamma(0,t)\Gamma(0,s) \,\mathrm{d}t \,\mathrm{d}s \\
&\quad = \int_0^\infty \Gamma(0,t) \left(\frac{1}{a_{r'}} \int_0^{a_{r'}t/a_r} 2s\Gamma(0,s) \,\mathrm{d}s + \frac{t}{a_r} \int_{a_{r'}t/a_r}^\infty 2\Gamma(0,s) \,\mathrm{d}s\right) \mathrm{d}t \\
&\quad = \int_0^\infty \Gamma(0,t) \left(\frac{t}{a_r} e^{-\frac{a_{r'}}{a_r}t} - \frac{1}{a_{r'}} e^{-\frac{a_{r'}}{a_r}t} + \frac{1}{a_{r'}} - \frac{a_{r'}}{a_r^2} t^2 \Gamma\left(0, \frac{a_{r'}}{a_r}t\right)\right) \mathrm{d}t \\
&\quad = \frac{1}{a_r} \int_0^\infty t e^{-\frac{a_{r'}}{a_r}t} \Gamma(0,t) \,\mathrm{d}t - \frac{1}{a_{r'}} \int_0^\infty e^{-\frac{a_{r'}}{a_r}t} \Gamma(0,t) \,\mathrm{d}t \\
&\qquad + \frac{1}{a_{r'}} \int_0^\infty \Gamma(0,t) \,\mathrm{d}t - \frac{a_{r'}}{a_r^2} \int_0^\infty t^2 \Gamma\left(0, \frac{a_{r'}}{a_r}t\right) \Gamma(0,t) \,\mathrm{d}t,
\end{aligned}
$$

since $\int_0^a 2t\Gamma(0,t) \,\mathrm{d}t = a^2\Gamma(0,a) - ae^{-a} - e^{-a} + 1$ and $\int_a^\infty \Gamma(0,t) \,\mathrm{d}t = e^{-a} - a\Gamma(0,a)$. Next, we use $\int_0^\infty \Gamma(0,t) \,\mathrm{d}t = 1$, $\int_0^\infty e^{-at}\Gamma(0,t) \,\mathrm{d}t = \frac{\log(1+a)}{a}$, $\int_0^\infty te^{-at}\Gamma(0,t) \,\mathrm{d}t = \frac{\log(1+a)}{a^2} - \frac{1}{a(a+1)}$ and $\int_0^\infty t^2\Gamma(0,t)\Gamma(0,at) \,\mathrm{d}t = -\frac{2a^2+a+2}{3a^2(a+1)} + \frac{2(a^3+1)\log(a+1)}{3a^3} - \frac{2\log a}{3}$ to see

$$
\begin{aligned}
&2 \int_0^\infty \int_0^\infty ((t/a_r) \wedge (s/a_{r'})) \Gamma(0,t)\Gamma(0,s) \,\mathrm{d}t \,\mathrm{d}s \\
&\quad = \frac{a_r \log(1 + a_{r'}/a_r)}{a_{r'}^2} - \frac{a_r/a_{r'}}{a_r + a_{r'}} - \frac{a_r \log(1 + a_{r'}/a_r)}{a_{r'}^2} + \frac{1}{a_{r'}} \\
&\qquad + \frac{2a_{r'}^2 + a_r a_{r'} + 2a_r^2}{3a_r a_{r'}(a_r + a_{r'})} - \frac{2(a_{r'}^3 + a_r^3)\log(a_{r'}/a_r + 1)}{3a_r^2 a_{r'}^2} + \frac{2a_{r'}\log(a_{r'}/a_r)}{3a_r^2} \\
&\quad = \frac{2}{3a_r}\left(1 - \frac{a_{r'}}{a_r}\log\left(\frac{a_r}{a_{r'}} + 1\right)\right) + \frac{2}{3a_{r'}}\left(1 - \frac{a_r}{a_{r'}}\log\left(\frac{a_{r'}}{a_r} + 1\right)\right).
\end{aligned}
$$

Finally we conclude this part by giving the limiting variance.

$$\begin{aligned}
&\mathbb{E}\left[\tilde{\Sigma}_{\mathrm{d}}(x)\right] \\
&\quad= \frac{\sigma^2(x)}{f(x)} \sum_{r=0}^{J} \sum_{r'=0}^{J} \omega_r \omega_{r'} \left(\frac{2 a_{r'}}{3} \left(1 - \frac{a_{r'}}{a_r} \log\left(\frac{a_r}{a_{r'}} + 1 \right) \right) + \frac{2 a_r}{3} \left(1 - \frac{a_r}{a_{r'}} \log\left(\frac{a_{r'}}{a_r} + 1 \right) \right) \right)^d \\
&\qquad + O\left(\frac{1}{\lambda^{1 \wedge \beta_f \wedge \beta_\sigma}} + \frac{\lambda^d}{n} + \frac{1}{B} \right) \\
&\quad= \Sigma_{\mathrm{d}}(x) + O\left(\frac{1}{\lambda^{1 \wedge \beta_f \wedge \beta_\sigma}} + \frac{\lambda^d}{n} + \frac{1}{B} \right).
\end{aligned}$$

It follows from this and the previous part that

$$\mathbb{E}\left[\left(\tilde{\Sigma}_{\mathrm{d}}(x) - \Sigma_{\mathrm{d}}(x) \right)^2 \right] \lesssim \frac{\lambda^d}{n} + \frac{1}{B} + \frac{1}{\lambda^{2(1 \wedge \beta_f \wedge \beta_\sigma)}}.$$

Part 4: a lower bound for the second moment with a single tree

We finally show here that if $B = 1$ (a forest with a single tree), then $\tilde{\Sigma}_{\mathrm{d}}(x)$ has a divergent second moment. We take $J = 0$ for brevity (no debiasing), and recall that $\sigma^2(x)$ and $f(x)$ are bounded below. Further, since $\lambda T(x)_j \leq \Gamma(2, 1)$, we have by Jensen's inequality that

$$\begin{aligned}
\mathbb{E}\left[\tilde{\Sigma}_{\mathrm{d}}(x)^2\right] &= \mathbb{E}\left[\left(\frac{n}{\lambda^d} \sum_{i=1}^n \frac{\mathbb{I}\{X_i \in T(x)\} \sigma^2(X_i)}{N(x)^2} \right)^2 \right] \gtrsim \frac{n^2}{\lambda^{2d}} \mathbb{E}\left[\frac{1}{N(x)^4} \left(\sum_{i=1}^n \mathbb{I}\{X_i \in T(x)\} \right)^2 \right] \\
&= \frac{n^2}{\lambda^{2d}} \sum_{i=1}^n \mathbb{E}\left[\frac{\mathbb{I}\{X_i \in T(x)\}}{N(x)^3} \right] = \frac{n^3}{\lambda^{2d}} \mathbb{E}\left[\frac{\mathbb{I}\{X_i \in T(x)\}}{(N_{-i}(x) + 1)^3} \right] \gtrsim \frac{n^3}{\lambda^{2d}} \mathbb{E}\left[\frac{|T(x)|}{\mathbb{E}\left[(N_{-i}(x) + 1)^3 \mid T \right]} \right] \\
&\gtrsim \frac{n^3}{\lambda^{2d}} \mathbb{E}\left[\frac{|T(x)|}{n^3 |T(x)|^3 + 1} \right] \gtrsim \frac{1}{\lambda^{2d}} \mathbb{E}\left[\frac{\mathbb{I}\{|T(x)| \geq 1/n\}}{|T(x)|^2} \right] \gtrsim \frac{1}{\lambda^{2d}} \mathbb{E}\left[\frac{\mathbb{I}\{|T(x)_j| \geq 1/n\}}{|T(x)_j|^2} \right]^d \\
&\gtrsim \left(\int_{1/n}^1 \frac{s e^{-s}}{s^2} \,\mathrm{d}s \right)^d \gtrsim \left(\int_{1/n}^1 \frac{1}{s} \,\mathrm{d}s \right)^d \gtrsim (\log n)^d .
\end{aligned}$$

□

Proof (Theorem 4)**.** The bias–variance decomposition with Lemma 4 and the proof of Theorem 5 with $J = \lfloor \beta / 2 \rfloor$ give

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{\mu}_{\mathrm{d}}(x) - \mu(x) \right)^2 \right] &= \mathbb{E}\left[\left(\hat{\mu}_{\mathrm{d}}(x) - \mathbb{E}\left[\hat{\mu}_{\mathrm{d}}(x) \mid \mathbf{X}, \mathbf{T} \right] \right)^2 \right] + \mathbb{E}\left[\left(\mathbb{E}\left[\hat{\mu}_{\mathrm{d}}(x) \mid \mathbf{X}, \mathbf{T} \right] - \mu(x) \right)^2 \right] \\
&\lesssim \frac{\lambda^d}{n} + \frac{1}{\lambda^{2\beta}} + \frac{1}{\lambda^{2(1 \wedge \beta)} B}.
\end{aligned}$$

As $\lambda \asymp n^{\frac{1}{d + 2\beta}}$ and $B \gtrsim n^{\frac{2\beta - 2(1 \wedge \beta)}{d + 2\beta}}$, we have

$$\mathbb{E}\left[\left(\hat{\mu}_{\mathrm{d}}(x) - \mu(x) \right)^2 \right] \lesssim n^{-\frac{2\beta}{d + 2\beta}}.$$

□

Proof (Theorem 5)**.** Define $S_i(x) = \sqrt{n/\lambda^d} \sum_{r=0}^{J} \omega_r \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{I}_{ibr}(x)\varepsilon_i}{N_{br}(x)}$, which are independent and zero mean conditional on $(\mathbf{X}, \mathbf{T})$, and satisfy

$$\sqrt{\frac{n}{\lambda^d}} \big(\hat{\mu}_{\mathrm{d}}(x) - \mathbb{E}\left[\hat{\mu}_{\mathrm{d}}(x) \mid \mathbf{X}, \mathbf{T}\right]\big) = \sum_{i=1}^{n} S_i(x).$$

Therefore by Petrov (1995, Theorem 5.7) conditional on $(\mathbf{X}, \mathbf{T})$, with $\zeta = \delta \wedge 1$,

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\tilde{\Sigma}_{\mathrm{d}}(x)^{-1/2} \sum_{i=1}^{n} S_i \leq t \;\middle|\; \mathbf{X}, \mathbf{T} \right) - \Phi(t) \right| \lesssim 1 \wedge \left(\tilde{\Sigma}_{\mathrm{d}}(x)^{-1-\zeta/2} \sum_{i=1}^{n} \mathbb{E}\left[|S_i|^{2+\zeta} \mid \mathbf{X}, \mathbf{T} \right] \right).$$

It immediately follows that

$$\begin{aligned}
\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\tilde{\Sigma}_{\mathrm{d}}(x)^{-1/2} \sum_{i=1}^{n} S_i \leq t \right) - \Phi(t) \right| &= \sup_{t \in \mathbb{R}} \left| \mathbb{E}\left[\mathbb{P}\left(\tilde{\Sigma}_{\mathrm{d}}(x)^{-1/2} \sum_{i=1}^{n} S_i \leq t \;\middle|\; \mathbf{X}, \mathbf{T} \right) \right] - \Phi(t) \right| \\
&\leq \mathbb{E}\left[\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\tilde{\Sigma}_{\mathrm{d}}(x)^{-1/2} \sum_{i=1}^{n} S_i \leq t \;\middle|\; \mathbf{X}, \mathbf{T} \right) - \Phi(t) \right| \right] \\
&\lesssim \mathbb{E}\left[1 \wedge \left(\tilde{\Sigma}_{\mathrm{d}}(x)^{-1-\zeta/2} \sum_{i=1}^{n} \mathbb{E}\left[|S_i|^{2+\zeta} \mid \mathbf{X}, \mathbf{T} \right] \right) \right].
\end{aligned}$$

To bound this quantity, we first partition by the event that $\tilde{\Sigma}_{\mathrm{d}}(x)$ is bounded away from zero:

$$\begin{aligned}
\mathbb{E}\left[1 \wedge \left(\tilde{\Sigma}_{\mathrm{d}}(x)^{-1-\zeta/2} \sum_{i=1}^{n} \mathbb{E}\left[|S_i|^{2+\zeta} \mid \mathbf{X}, \mathbf{T} \right] \right) \right] &\lesssim \mathbb{E}\big[\tilde{\Sigma}_{\mathrm{d}}(x)\big]^{-1-\zeta/2} \sum_{i=1}^{n} \mathbb{E}\left[|S_i|^{2+\zeta} \right] \\
&\quad + \mathbb{P}\left(\left| \tilde{\Sigma}_{\mathrm{d}}(x) - \mathbb{E}\big[\tilde{\Sigma}_{\mathrm{d}}(x)\big] \right| > \frac{\mathbb{E}\big[\tilde{\Sigma}_{\mathrm{d}}(x)\big]}{2} \right). \qquad (26)
\end{aligned}$$

The first term in (26) is bounded as follows. We already have from the proof of Lemma 5 that eventually $\mathbb{E}\big[\tilde{\Sigma}_{\mathrm{d}}(x)\big] \geq \Sigma_{\mathrm{d}}(x)/2 \gtrsim 1$. Since $\mathbb{E}[\varepsilon_i^{2+\delta} \mid \mathbf{X}]$ is bounded, by Jensen's inequality,

$$\begin{aligned}
\sum_{i=1}^{n} \mathbb{E}\left[|S_i|^{2+\zeta} \right] &= \sum_{i=1}^{n} \mathbb{E}\left[\left| \sqrt{\frac{n}{\lambda^d}} \sum_{r=0}^{J} \omega_r \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{I}_{ibr}(x)\varepsilon_i}{N_{br}(x)} \right|^{2+\zeta} \right] \\
&\leq \left(\frac{n}{\lambda^d} \right)^{1+\zeta/2} (J+1)^{1+\zeta} \sum_{r=0}^{J} |\omega_r|^3 \, \mathbb{E}\left[\sum_{i=1}^{n} \left(\frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{I}_{ibr}(x)}{N_{br}(x)} \right)^{2+\zeta} \right] \\
&\lesssim \left(\frac{n}{\lambda^d} \right)^{1+\zeta/2} \mathbb{E}\left[\sum_{i=1}^{n} \left(\frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{I}_{ibr}(x)}{N_{br}(x)} \right)^{2+\zeta} \right].
\end{aligned}$$

We now proceed by cases. If $\zeta = 1$, note that by Lemma 14 and with $B \gtrsim (\log n)^d$,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} \left(\frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{I}_{ibr}(x)}{N_{br}(x)} \right)^3 \right] &= \mathbb{E}\left[\sum_{i=1}^{n} \frac{1}{B^3} \sum_{b=1}^{B} \sum_{b'=1}^{B} \sum_{b''=1}^{B} \frac{\mathbb{I}_{ibr}(x)}{N_{br}(x)} \frac{\mathbb{I}_{ib'r'}(x)}{N_{b'r'}(x)} \frac{\mathbb{I}_{ib''r''}(x)}{N_{b''r''}(x)} \right] \\
&\leq \frac{1}{B^2} \sum_{b=1}^{B} \sum_{b'=1}^{B} \mathbb{E}\left[\frac{\mathbb{I}_{br}(x)}{N_{br}(x)} \frac{\mathbb{I}_{b'r'}(x)}{N_{b'r'}(x)} \right] \lesssim \frac{\lambda^{2d}}{n^2} + \frac{1}{B} \frac{\lambda^{2d} (\log n)^d}{n^2} \lesssim \frac{\lambda^{2d}}{n^2}.
\end{aligned}$$

Alternatively, if $\zeta \in (0,1)$, by Jensen's inequality and Lemma 14,

$$\mathbb{E}\left[\sum_{i=1}^{n}\left(\frac{1}{B}\sum_{b=1}^{B}\frac{\mathbb{I}_{ibr}(x)}{N_{br}(x)}\right)^{2+\zeta}\right] \leq \mathbb{E}\left[\sum_{i=1}^{n}\left(\frac{\mathbb{I}_{ibr}(x)}{N_{br}(x)}\right)^{2+\zeta}\right] \leq \mathbb{E}\left[\sum_{i=1}^{n}\frac{\mathbb{I}_{ibr}(x)}{N_{br}(x)}\left(\frac{\mathbb{I}_{br}(x)}{N_{br}(x)}\right)^{1+\zeta}\right]$$
$$\leq \mathbb{E}\left[\left(\frac{\mathbb{I}_{br}(x)}{N_{br}(x)}\right)^{1+\zeta}\right] \lesssim \left(\frac{\lambda^d}{n}\right)^{1+\zeta}.$$

Both cases lead to the conclusion that

$$\mathbb{E}\big[\tilde{\Sigma}_{\mathrm{d}}(x)\big]^{-1-\zeta/2}\sum_{i=1}^{n}\mathbb{E}\left[|S_i|^{2+\zeta}\right] \lesssim \left(\frac{n}{\lambda^d}\right)^{1+\zeta/2}\left(\frac{\lambda^d}{n}\right)^{1+\zeta} = \left(\frac{\lambda^d}{n}\right)^{\zeta/2}.$$

For the second term in (26), Chebyshev's inequality along with the proof of Lemma 5 give

$$\mathbb{P}\left(\left|\tilde{\Sigma}_{\mathrm{d}}(x) - \mathbb{E}\big[\tilde{\Sigma}_{\mathrm{d}}(x)\big]\right| > \frac{\mathbb{E}\big[\tilde{\Sigma}_{\mathrm{d}}(x)\big]}{2}\right) \lesssim \mathrm{Var}\left[\tilde{\Sigma}_{\mathrm{d}}(x)\right] \lesssim \frac{1}{B} + \frac{\lambda^d}{n}.$$

Therefore

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\tilde{\Sigma}_{\mathrm{d}}(x)^{-1/2}\sum_{i=1}^{n}S_i \leq t\right) - \Phi(t)\right| \lesssim \left(\frac{\lambda^d}{n}\right)^{\frac{1\wedge\delta}{2}} + \frac{1}{B}.$$

□

Proof (Lemma 6)**.** We begin by showing that $\hat{\sigma}^2(x)$ is consistent for $\sigma^2(x)$.

Part 1: consistency of $\hat{\sigma}^2(x)$
Recall that

$$\hat{\sigma}^2(x) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}\frac{Y_i^2\,\mathbb{I}_{ib}(x)}{N_b(x)} - \hat{\mu}(x)^2. \tag{27}$$

The first term in (27) is simply a Mondrian forest estimator of $\mathbb{E}[Y_i^2 \mid X_i = x] = \sigma^2(x) + \mu(x)^2$, which is bounded and in $\mathcal{H}^{\beta_\mu \wedge \beta_\sigma}$. Therefore, by Lemma 1, its conditional bias is

$$\mathbb{E}\left[\left(\frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}\frac{\left(\sigma^2(X_i) + \mu(X_i)^2 - \sigma^2(x) - \mu(x)^2\right)\mathbb{I}_{ib}(x)}{N_b(x)}\right)^2\right]$$
$$\lesssim \frac{1}{\lambda^{2(2\wedge\beta\wedge\beta_\sigma)}} + \frac{1}{\lambda^{2(1\wedge\beta\wedge\beta_\sigma)}B} + \frac{1}{\lambda^{2(1\wedge\beta\wedge\beta_\sigma)}}\frac{\lambda^d}{n}.$$

We handle the stochastic part with a truncation argument. Let $S_i = Y_i^2 - \sigma^2(X_i) - \mu(X_i)^2$ and $\tilde{S}_i = S_i\,\mathbb{I}\{|S_i| \leq M\} - \mathbb{E}\left[S_i\,\mathbb{I}\{|S_i| \leq M\} \mid X_i\right]$ where $M > 0$ is to be determined. We bound

$$\frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}\frac{\tilde{S}_i\,\mathbb{I}_{ib}(x)}{N_b(x)} + \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}\frac{(S_i - \tilde{S}_i)\,\mathbb{I}_{ib}(x)}{N_b(x)}. \tag{28}$$

The first term in (28) is controlled with a variance bound, noting $\tilde{S}_i \leq 2M$ almost surely.

$$\mathrm{Var}\left[\frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}\frac{\tilde{S}_i\,\mathbb{I}_{ib}(x)}{N_b(x)}\right] \leq \sum_{i=1}^{n}\mathbb{E}\left[\frac{\tilde{S}_i^2\,\mathbb{I}_{ib}(x)}{N_b(x)^2}\right] \leq 4M^2\sum_{i=1}^{n}\mathbb{E}\left[\frac{\mathbb{I}_b(x)}{N_b(x)}\right] \lesssim M^2\frac{\lambda^d}{n}.$$

For the second term in (28), note that $S_i - \tilde{S}_i = S_i \mathbb{I}\{|S_i| > M\} - \mathbb{E}\left[S_i \mathbb{I}\{|S_i| > M\} \mid X_i\right]$ because $\mathbb{E}[S_i \mid X_i] = 0$. Since $\mathbb{E}\left[|Y_i|^{2+\delta} \mid X_i\right]$ is bounded, so is $\mathbb{E}\left[|S_i|^{1+\delta/2} \mid X_i\right]$. Thus

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}\frac{(S_i - \tilde{S}_i)\,\mathbb{I}_{ib}(x)}{N_b(x)}\right|\right] &\leq \sum_{i=1}^{n}\mathbb{E}\left[\frac{\mathbb{I}_{ib}(x)}{N_b(x)}\mathbb{E}\left[|S_i - \tilde{S}_i| \mid X_i\right]\right] \\
&\leq 2\sum_{i=1}^{n}\mathbb{E}\left[\frac{\mathbb{I}_{ib}(x)}{N_b(x)}\mathbb{E}\left[|S_i|\,\mathbb{I}\{|S_i| > M\} \mid X_i\right]\right] \\
&\leq \frac{2}{M^{\delta/2}}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\mathbb{I}_{ib}(x)}{N_b(x)}\mathbb{E}\left[|S_i|^{1+\delta/2} \mid X_i\right]\right] \lesssim \frac{1}{M^{\delta/2}}.
\end{aligned}$$

Consistency of the second term in (27) follows directly from Lemma 1 and Theorem 5 with the same bias and variance bounds. Therefore

$$\begin{aligned}
\mathbb{E}\left[\left|\hat{\sigma}^2(x) - \sigma^2(x)\right|\right] &\lesssim \frac{1}{\lambda^{2\wedge\beta\wedge\beta_\sigma}} + \frac{1}{\lambda^{1\wedge\beta\wedge\beta_\sigma}\sqrt{B}} + \frac{1}{\lambda^{1\wedge\beta\wedge\beta_\sigma}}\sqrt{\frac{\lambda^d}{n}} + M\sqrt{\frac{\lambda^d}{n}} + \frac{1}{M^{\delta/2}} \\
&\lesssim \frac{1}{\lambda^{2\wedge\beta\wedge\beta_\sigma}} + \frac{1}{\lambda^{1\wedge\beta\wedge\beta_\sigma}\sqrt{B}} + \left(\frac{\lambda^d}{n}\right)^{\frac{\delta}{4+2\delta}},
\end{aligned}$$

where we set $M = \left(\frac{\lambda^d}{n}\right)^{-\frac{1}{2+\delta}}$. Note that if $\delta \geq 2$ then the variance argument applies directly, without the need for truncation, yielding

$$\mathbb{E}\left[\left|\hat{\sigma}^2(x) - \sigma^2(x)\right|^2\right] \lesssim \frac{1}{\lambda^{2(2\wedge\beta\wedge\beta_\sigma)}} + \frac{1}{\lambda^{2(1\wedge\beta\wedge\beta_\sigma)}B} + \frac{\lambda^d}{n}.$$

Part 2: consistency of the sum

Note that

$$\frac{n}{\lambda^d}\sum_{i=1}^{n}\left(\sum_{r=0}^{J}\omega_r\frac{1}{B}\sum_{b=1}^{B}\frac{\mathbb{I}\{X_i \in T_{br}(x)\}}{\sum_{i=1}^{n}\mathbb{I}\{X_i \in T_{br}(x)\}}\right)^2 = \frac{n}{\lambda^d}\frac{1}{B^2}\sum_{i=1}^{n}\sum_{r=0}^{J}\sum_{r'=0}^{J}\omega_r\omega_{r'}\sum_{b=1}^{B}\sum_{b'=1}^{B}\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)}{N_{br}(x)N_{b'r'}(x)}.$$

This is exactly the same as $\tilde{\Sigma}_{\mathrm{d}}(x)$, if we were to take $\sigma^2(x) = 1$. Thus by Lemma 5, we obtain

$$\mathbb{E}\left[\left(\frac{n}{\lambda^d}\frac{1}{B^2}\sum_{i=1}^{n}\sum_{r=0}^{J}\sum_{r'=0}^{J}\omega_r\omega_{r'}\sum_{b=1}^{B}\sum_{b'=1}^{B}\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)}{N_{br}(x)N_{b'r'}(x)} - \frac{\Sigma_{\mathrm{d}}(x)}{\sigma^2(x)}\right)^2\right] \lesssim \frac{\lambda^d}{n} + \frac{1}{B} + \frac{1}{\lambda^{2(1\wedge\beta_f\wedge\beta_\sigma)}}.$$

Part 3: conclusion

By the previous parts and the Cauchy–Schwarz inequality,

$$\mathbb{E}\left[\left|\hat{\Sigma}_{\mathrm{d}}(x) - \Sigma_{\mathrm{d}}(x)\right|^{1/2}\right]^2 \lesssim \left(\frac{\lambda^d}{n}\right)^{\frac{\delta}{4+2\delta}} + \frac{1}{\sqrt{B}} + \frac{1}{\lambda^{1\wedge\beta_\mu\wedge\beta_f\wedge\beta_\sigma}}.$$

If $\delta \geq 2$ then we obtain

$$\mathbb{E}\left[\left|\hat{\Sigma}_{\mathrm{d}}(x) - \Sigma_{\mathrm{d}}(x)\right|\right] \lesssim \sqrt{\frac{\lambda^d}{n}} + \frac{1}{\sqrt{B}} + \frac{1}{\lambda^{1\wedge\beta_\mu\wedge\beta_f\wedge\beta_\sigma}}.$$

Combining these yields

$$\mathbb{E}\left[\left|\hat{\Sigma}_{\mathrm{d}}(x) - \Sigma_{\mathrm{d}}(x)\right|^{\frac{2-\mathbb{I}\{\delta<2\}}{2}}\right]^{\frac{2}{2-\mathbb{I}\{\delta<2\}}} \lesssim \left(\frac{\lambda^d}{n}\right)^{\frac{1}{2}-\frac{\mathbb{I}\{\delta<2\}}{2+\delta}} + \frac{1}{\sqrt{B}} + \frac{1}{\lambda^{1\wedge\beta_\mu\wedge\beta_f\wedge\beta_\sigma}}.$$

□

Proof (Theorem 6)**.** Let τ and $\hat{\tau}$ be real-valued random variables. Then for any $\varepsilon > 0$,

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\big(\hat{\tau} \le t\big) - \Phi(t)\right| \le \sup_{t\in\mathbb{R}} \left|\mathbb{P}\big(\tau \le t\big) - \Phi(t)\right| + \varepsilon\sqrt{2/\pi} + \mathbb{P}\big(|\hat{\tau} - \tau| > \varepsilon\big).$$

Defining $a/0 = 0$ for all $a \in \mathbb{R}$ to accommodate the event $\hat{\Sigma}_{\mathrm{d}}(x) = 0$, we apply this result to

$$\hat{\tau} = \sqrt{\frac{n}{\lambda^d}}\left(\frac{\hat{\mu}_{\mathrm{d}}(x) - \mu(x)}{\sqrt{\hat{\Sigma}_{\mathrm{d}}(x)}} - \frac{\mathbb{E}\left[\hat{\mu}_{\mathrm{d}}(x)\right] - \mu(x)}{\sqrt{\Sigma_{\mathrm{d}}(x)}}\right) \qquad \text{and} \qquad \tau = \sqrt{\frac{n}{\lambda^d}}\frac{\hat{\mu}_{\mathrm{d}}(x) - \mathbb{E}\left[\hat{\mu}_{\mathrm{d}}(x) \mid \mathbf{X}, \mathbf{T}\right]}{\sqrt{\tilde{\Sigma}_{\mathrm{d}}(x)}}$$

respectively, noting that $\sup_{t\in\mathbb{R}} \left|\mathbb{P}\big(\tau \le t\big) - \Phi(t)\right| \lesssim \left(\frac{\lambda^d}{n}\right)^{\frac{1\wedge\delta}{2}} + \frac{1}{B}$ by Theorem 5. With

$$v = \sqrt{\frac{n}{\lambda^d}}\frac{\mathbb{E}\left[\hat{\mu}_{\mathrm{d}}(x)\right] - \mu(x)}{\sqrt{\Sigma_{\mathrm{d}}(x)}} \lesssim \sqrt{\frac{\lambda^d}{n}}\frac{1}{\lambda^{1\wedge\beta_\mu}} + \sqrt{\frac{n}{\lambda^d}}\frac{1}{\lambda^\beta}$$

by the proof of Lemma 1, and by Taylor's theorem, for some $s, s' \in \mathbb{R}$, we have

$$\begin{aligned}
&\left|\mathbb{P}\big(\mu(x) \in \mathrm{CI}_{\mathrm{d}}(x)\big) - (1-\alpha)\right| \\
&\quad= \left|\mathbb{P}\left(q_{\alpha/2} \le \sqrt{\frac{n}{\lambda^d}}\frac{\hat{\mu}_{\mathrm{d}}(x) - \mu(x)}{\sqrt{\hat{\Sigma}_{\mathrm{d}}(x)}} \le q_{1-\alpha/2}\right) - (1-\alpha)\right| \\
&\quad= \left|\mathbb{P}\left(q_{\alpha/2} - v \le \hat{\tau} \le q_{1-\alpha/2} - v\right) - (1-\alpha)\right| \\
&\quad\le \left|\mathbb{P}\left(\hat{\tau} \le q_{1-\alpha/2} - v\right) - \Phi(q_{1-\alpha/2} - v)\right| + \left|\mathbb{P}\left(\hat{\tau} < q_{\alpha/2} - v\right) - \Phi(q_{\alpha/2} - v)\right| \\
&\qquad+ \left|\Phi(q_{1-\alpha/2} - v) - (1-\alpha/2) - \Phi(q_{\alpha/2} - v) + \alpha/2\right| \\
&\quad\lesssim \left(\frac{\lambda^d}{n}\right)^{\frac{1\wedge\delta}{2}} + \frac{1}{B} + \varepsilon + \mathbb{P}\left(|\hat{\tau} - \tau| > \varepsilon\right) + \left|-v\phi(1-\alpha/2) + v^2\phi'(s)/2 + v\phi(\alpha/2) - v^2\phi'(s')/2\right| \\
&\quad\lesssim \left(\frac{\lambda^d}{n}\right)^{\frac{1\wedge\delta}{2}} + \frac{1}{B} + \varepsilon + \frac{n}{\lambda^d}\frac{1}{\lambda^{2\beta}} + \mathbb{P}\left(|\hat{\tau} - \tau| > \varepsilon\right).
\end{aligned}$$

It remains to bound $\mathbb{P}\left(|\hat{\tau} - \tau| > \varepsilon\right)$. Observe that

$$\begin{aligned}
|\hat{\tau} - \tau| &\le R_1 + R_2 + R_3, \\
R_1 &= \sqrt{\frac{n}{\lambda^d}}\left|\hat{\mu}_{\mathrm{d}}(x) - \mu(x)\right|\left|\frac{1}{\sqrt{\hat{\Sigma}_{\mathrm{d}}(x)}} - \frac{1}{\sqrt{\tilde{\Sigma}_{\mathrm{d}}(x)}}\right|, \\
R_2 &= \sqrt{\frac{n}{\lambda^d}}\left|\mathbb{E}\left[\hat{\mu}_{\mathrm{d}}(x) \mid \mathbf{X}, \mathbf{T}\right] - \mu(x)\right|\left|\frac{1}{\sqrt{\tilde{\Sigma}_{\mathrm{d}}(x)}} - \frac{1}{\sqrt{\Sigma_{\mathrm{d}}(x)}}\right|, \\
R_3 &= \sqrt{\frac{n}{\lambda^d}}\frac{\left|\mathbb{E}\left[\hat{\mu}_{\mathrm{d}}(x) \mid \mathbf{X}, \mathbf{T}\right] - \mathbb{E}\left[\hat{\mu}_{\mathrm{d}}(x)\right]\right|}{\sqrt{\Sigma_{\mathrm{d}}(x)}}.
\end{aligned}$$

We begin with R_1. Take $a > 1$ and $b^{2/3} = \Sigma_{\mathrm{d}}(x)/2$, so by the proof of Lemma 5,

$$\begin{aligned}
\mathbb{P}(R_1 > \varepsilon) &\leq \mathbb{P}\left(\sqrt{\frac{n}{\lambda^d}}\big|\hat{\mu}_{\mathrm{d}}(x) - \mu(x)\big| > a\varepsilon\right) + \mathbb{P}\left(\left|\frac{1}{\sqrt{\hat{\Sigma}_{\mathrm{d}}(x)}} - \frac{1}{\sqrt{\tilde{\Sigma}_{\mathrm{d}}(x)}}\right| > \frac{1}{a}\right) \\
&\leq \frac{n}{a^2\varepsilon^2\lambda^d}\mathbb{E}\big[(\hat{\mu}_{\mathrm{d}}(x) - \mu(x))^2\big] + \mathbb{P}\left(\tilde{\Sigma}_{\mathrm{d}}(x) < b^{2/3}\right) + \mathbb{P}\left(\hat{\Sigma}_{\mathrm{d}}(x) < b^{2/3}\right) \\
&\quad + \mathbb{P}\left(\left|\hat{\Sigma}_{\mathrm{d}}(x) - \tilde{\Sigma}_{\mathrm{d}}(x)\right| > \frac{b}{a}\right) \\
&\lesssim \frac{n}{a^2\varepsilon^2\lambda^d}\left(\frac{\lambda^d}{n} + \frac{1}{\lambda^{2\beta}} + \frac{1}{\lambda^{2(1\wedge\beta)}B}\right) + \mathrm{Var}\left[\tilde{\Sigma}_{\mathrm{d}}(x)\right] + \mathbb{P}\left(|\hat{\Sigma}_{\mathrm{d}}(x) - \Sigma_{\mathrm{d}}(x)| > \Sigma_{\mathrm{d}}(x)/2\right) \\
&\quad + \mathbb{P}\left(\left|\tilde{\Sigma}_{\mathrm{d}}(x) - \Sigma_{\mathrm{d}}(x)\right| > \frac{b}{2a}\right) + \mathbb{P}\left(\left|\hat{\Sigma}_{\mathrm{d}}(x) - \Sigma_{\mathrm{d}}(x)\right| > \frac{b}{2a}\right) \\
&\lesssim \frac{1}{a^2\varepsilon^2} + \frac{1}{B} + \frac{\lambda^d}{n} + a^2\left(\frac{\lambda^d}{n} + \frac{1}{B} + \frac{1}{\lambda^{2(1\wedge\beta_f\wedge\beta_\sigma)}}\right) + \mathbb{P}\left(\left|\hat{\Sigma}_{\mathrm{d}}(x) - \Sigma_{\mathrm{d}}(x)\right| > \frac{b}{2a}\right).
\end{aligned}$$

If $\delta < 2$ then the proof of Lemma 6 along with Markov's inequality gives

$$\begin{aligned}
\mathbb{P}\left(\left|\hat{\Sigma}_{\mathrm{d}}(x) - \Sigma_{\mathrm{d}}(x)\right| > \frac{b}{2a}\right) &\lesssim \sqrt{a}\,\mathbb{E}\left[\left|\hat{\Sigma}_{\mathrm{d}}(x) - \Sigma_{\mathrm{d}}(x)\right|^{1/2}\right] \\
&\lesssim \sqrt{a}\left(\left(\frac{\lambda^d}{n}\right)^{\frac{\delta}{8+4\delta}} + \frac{1}{B^{1/4}} + \frac{1}{\lambda^{(1\wedge\beta_\mu\wedge\beta_f\wedge\beta_\sigma)/2}}\right),
\end{aligned}$$

and whenever this converges to zero, minimizing over a yields

$$\begin{aligned}
\mathbb{P}(R_1 > \varepsilon) &\lesssim \frac{1}{a^2\varepsilon^2} + \frac{1}{B} + \frac{\lambda^d}{n} + \sqrt{a}\left(\left(\frac{\lambda^d}{n}\right)^{\frac{\delta}{8+4\delta}} + \frac{1}{B^{1/4}} + \frac{1}{\lambda^{(1\wedge\beta_\mu\wedge\beta_f\wedge\beta_\sigma)/2}}\right) \\
&\lesssim \frac{1}{\varepsilon^{2/5}}\left(\left(\frac{\lambda^d}{n}\right)^{\frac{\delta}{10+5\delta}} + \frac{1}{B^{1/5}} + \frac{1}{\lambda^{2(1\wedge\beta_\mu\wedge\beta_f\wedge\beta_\sigma)/5}}\right).
\end{aligned}$$

If $\delta \geq 2$ however then we instead obtain

$$\mathbb{P}\left(\left|\hat{\Sigma}_{\mathrm{d}}(x) - \Sigma_{\mathrm{d}}(x)\right| > \frac{b}{2a}\right) \lesssim a\,\mathbb{E}\left[\left|\hat{\Sigma}_{\mathrm{d}}(x) - \Sigma_{\mathrm{d}}(x)\right|\right] \lesssim a\left(\sqrt{\frac{\lambda^d}{n}} + \frac{1}{\sqrt{B}} + \frac{1}{\lambda^{1\wedge\beta_\mu\wedge\beta_f\wedge\beta_\sigma}}\right),$$

and again if this converges to zero then minimizing over a gives

$$\begin{aligned}
\mathbb{P}(R_1 > \varepsilon) &\lesssim \frac{1}{a^2\varepsilon^2} + \frac{1}{B} + \frac{\lambda^d}{n} + a\left(\sqrt{\frac{\lambda^d}{n}} + \frac{1}{\sqrt{B}} + \frac{1}{\lambda^{1\wedge\beta_\mu\wedge\beta_f\wedge\beta_\sigma}}\right) \\
&\lesssim \frac{1}{\varepsilon^{2/3}}\left(\left(\frac{\lambda^d}{n}\right)^{1/3} + \frac{1}{B^{1/3}} + \frac{1}{\lambda^{2(1\wedge\beta_\mu\wedge\beta_f\wedge\beta_\sigma)/3}}\right).
\end{aligned}$$

For R_2, note that the same arguments used for R_1 apply again, yielding a bound no worse than that for R_1. Finally, for R_3, we have by Lemma 4 that

$$\begin{aligned}
\mathbb{P}(R_3 > \varepsilon) &\leq \frac{1}{\varepsilon^2}\mathbb{E}\left[R_3^2\right] \lesssim \frac{1}{\varepsilon^2}\frac{n}{\lambda^d}\mathbb{E}\left[\left(\mathbb{E}\left[\hat{\mu}_{\mathrm{d}}(x) \mid \mathbf{X}, \mathbf{T}\right] - \mathbb{E}\left[\hat{\mu}_{\mathrm{d}}(x)\right]\right)^2\right] \\
&\lesssim \frac{1}{\varepsilon^2}\frac{n}{\lambda^d}\left(\frac{1}{\lambda^{2(1\wedge\beta_\mu)}B} + \frac{1}{\lambda^{2(1\wedge\beta_\mu)}}\frac{\lambda^d}{n}\right) \lesssim \frac{1}{\varepsilon^2}\left(\frac{n}{\lambda^d}\frac{1}{\lambda^{2(1\wedge\beta_\mu)}B} + \frac{1}{\lambda^{2(1\wedge\beta_\mu)}}\right).
\end{aligned}$$

So far we have shown that if $\varepsilon \to 0$, then for $\delta < 2$,

$$
\begin{aligned}
\big|\mathbb{P}\big(\mu(x) \in \mathrm{CI_d}(x)\big) - (1-\alpha)\big| &\lesssim \left(\frac{\lambda^d}{n}\right)^{\frac{1\wedge\delta}{2}} + \frac{1}{B} + \varepsilon + \frac{n}{\lambda^d}\frac{1}{\lambda^{2\beta}} + \mathbb{P}\left(|\hat\tau - \tau| > \varepsilon\right) \\
&\lesssim \varepsilon + \frac{n}{\lambda^d}\frac{1}{\lambda^{2\beta}} + \frac{1}{\varepsilon^{2/5}}\left(\left(\frac{\lambda^d}{n}\right)^{\frac{\delta}{10+5\delta}} + \frac{1}{B^{1/5}} + \frac{1}{\lambda^{2(1\wedge\beta_\mu\wedge\beta_f\wedge\beta_\sigma)/5}} + \left(\frac{n}{\lambda^d}\right)^{1/5}\frac{1}{\lambda^{2(1\wedge\beta_\mu)/5}B^{1/5}}\right) \\
&\lesssim \frac{n}{\lambda^d}\frac{1}{\lambda^{2\beta}} + \left(\left(\frac{\lambda^d}{n}\right)^{\frac{\delta}{2+\delta}} + \frac{1}{B} + \frac{1}{\lambda^{2(1\wedge\beta_\mu\wedge\beta_f\wedge\beta_\sigma)}} + \frac{n}{\lambda^d}\frac{1}{\lambda^{2(1\wedge\beta)}B}\right)^{1/7},
\end{aligned}
$$

while for $\delta \geq 2$,

$$
\begin{aligned}
\big|\mathbb{P}\big(\mu(x) \in \mathrm{CI_d}(x)\big) - (1-\alpha)\big| &\lesssim \left(\frac{\lambda^d}{n}\right)^{\frac{1\wedge\delta}{2}} + \frac{1}{B} + \varepsilon + \frac{n}{\lambda^d}\frac{1}{\lambda^{2\beta}} + \mathbb{P}\left(|\hat\tau - \tau| > \varepsilon\right) \\
&\lesssim \varepsilon + \frac{n}{\lambda^d}\frac{1}{\lambda^{2\beta}} + \frac{1}{\varepsilon^{2/3}}\left(\left(\frac{\lambda^d}{n}\right)^{1/3} + \frac{1}{B^{1/3}} + \frac{1}{\lambda^{2(1\wedge\beta_\mu\wedge\beta_f\wedge\beta_\sigma)/3}} + \left(\frac{n}{\lambda^d}\right)^{1/3}\frac{1}{\lambda^{2(1\wedge\beta_\mu)/3}B^{1/3}}\right) \\
&\lesssim \frac{n}{\lambda^d}\frac{1}{\lambda^{2\beta}} + \left(\frac{\lambda^d}{n} + \frac{1}{B} + \frac{1}{\lambda^{2(1\wedge\beta_\mu\wedge\beta_f\wedge\beta_\sigma)}} + \frac{n}{\lambda^d}\frac{1}{\lambda^{2(1\wedge\beta_\mu)}B}\right)^{1/5},
\end{aligned}
$$

where in both displays we minimized over $\varepsilon > 0$. □

A.5 Proofs for Section 6

Proof (Lemma 7)**.** All computational complexities in this proof are understood to be upper bounds up to constants. The first step is to select λ using polynomial fitting as in Section 5.1. Constructing the design matrix $\mathbf{P}$ requires raising a number to a power of at most $J+1$ a total of $nd(J+1)$ times, giving a complexity of $nd(J+1)^2$. Multiplying the design matrix to obtain $\mathbf{P}^\mathsf{T}\mathbf{P}$ is $nd^2(J+1)^2$, and inverting this is $d^3(J+1)^3 \lesssim nd^2(J+1)^2$, giving an overall complexity of $nd^2(J+1)^2$ for selecting the lifetime.

Calculating the debiasing coefficients ω_r as in Section 4 involves inverting a $(J+1)\times(J+1)$ matrix, so is $(J+1)^3 \lesssim n(J+1)^2$. Next, constructing $U(x)$ as in (10) requires $Bd(J+1)$ comparisons, and forming $I(x)$ then needs nd comparisons.

Once $I(x)$ is available, Calculating $N_{br}(x)$, $S_{br}(x)$ and $V_{br}(x)$ as in (11) each take $Bd(J+1)|I(x)|$ operations, and from these we compute $\hat\mu(x)$ and $\hat\sigma^2(x)$ in $(J+1)B$ using (12). Constructing $\hat\Sigma_\mathrm{d}(x)$ as in (13) is $Bd(J+1)|I(x)|$, and calculating $\mathrm{CI_d}(x)$ with (9) has complexity 1.

Thus the overall complexity of Algorithm 1 is $nd^2(J+1)^2 + Bd(J+1) + Bd(J+1)|I(x)|$. To obtain the average case behavior we present a bound for $\mathbb{E}\left[|I(x)|\right]$. Firstly, since $f(x)$ is bounded and by the distribution of Mondrian cells,

$$
\mathbb{E}\left[|I(x)|\right] = \mathbb{E}\left[\sum_{i=1}^n \mathbb{I}\{X_i \in U(x)\}\right] = \sum_{i=1}^n \mathbb{E}\left[\mathbb{P}\big(X_i \in U(x) \mid U(x)\big)\right] \lesssim n\,\mathbb{E}\left[|U(x)|\right] \lesssim n\,\mathbb{E}\left[|U(x)_j|\right]^d.
$$

Next, by Lemma 9, we have that

$$
\mathbb{P}\left(|U(x)_j| \geq \frac{4t + 4\log(2B(J+1))}{\lambda}\right) \leq \mathbb{P}\left(\max_{0\leq r\leq J}\max_{1\leq b\leq B}|T_b(x)_j| \geq \frac{2t + 2\log(2B(J+1))}{\lambda}\right) \leq e^{-t},
$$

and integrating the tail probability yields

$$\mathbb{E}\left[|U(x)_j|\right] \lesssim \frac{\log(2B(J+1))}{\lambda}, \qquad \text{so that} \qquad \mathbb{E}\left[|I(x)|\right] \lesssim \frac{n\log(2B(J+1))^d}{\lambda^d}.$$

□

Proof (Lemma 8)**.** As in the proof of Lemma 7, the complexity of selecting the lifetime is $(n+k)d^2(J+1)^2$. Since this occurs with probability at most k/K, and $k \leq n$, the average case time complexity is $\frac{knd^2(J+1)^2}{K}$.

To update the trees, we sample and perform comparisons with at most $B^* d(J+1) \lesssim Bd(J+1)$ exponential random variables. We verify here that the resulting trees have the correct distribution, since by Mourtada et al. (2020, Proposition 1) and the memoryless property of the exponential distribution, with E'_{brj1} and E'_{brj1} i.i.d. copies of E_{brj1},

$$\begin{aligned} T^*_{br}(x)^-_j &= T_{br}(x)^-_j \vee \left(x_j - \frac{E_{brj1}}{\lambda^* - \lambda}\right) = 0 \vee \left(x_j - \frac{E'_{brj1}}{\lambda}\right) \vee \left(x_j - \frac{E_{brj1}}{\lambda^* - \lambda}\right) \\ &= x_j - \left(x_j \wedge \frac{E'_{brj1}}{\lambda} \wedge \frac{E_{brj1}}{\lambda^* - \lambda}\right) = x_j - \left(x_j \wedge \frac{E''_{brj1}}{\lambda^*}\right), \end{aligned}$$

as required. The same argument applies to $T^*_{br}(x)^+_j$. We also bound the expected number of trees which have changed. By a union bound and with E_1 and E_2 i.i.d. $\mathrm{Exp}(1)$,

$$\begin{aligned} \mathbb{E}\left[\sum_{r=0}^{J}\sum_{b=1}^{B} \mathbb{I}\big\{T^*_{br}(x) \neq T_{br}(x)\big\}\right] &= B(J+1)\,\mathbb{P}\big(T^*_{br}(x) \neq T_{br}(x)\big) \\ &\leq 2dB(J+1)\,\mathbb{P}\left(\frac{E_1}{\lambda^* - \lambda} < \frac{E_2}{\lambda}\right) \leq 2dB(J+1)\frac{\lambda^* - \lambda}{\lambda^*} \lesssim \frac{B(J+1)k}{n}, \end{aligned}$$

since $\frac{\lambda^* - \lambda}{\lambda} = \left(\frac{n+k}{n}\right)^\zeta - 1 \leq \frac{k\zeta}{n} \leq \frac{k}{nd}$. Constructing $U(x)$ requires $Bd(J+1)$ comparisons, and since for $b \leq B$ we have $T^*_{br}(x) \subseteq T_{br}(x)$,

$$\begin{aligned} \mathbb{P}\left(U^*(x) \nsubseteq U(x)\right) &\leq d\,\mathbb{P}\left(U^*(x)^-_j < U(x)^-_j\right) + d\,\mathbb{P}\left(U^*(x)^+_j > U(x)^+_j\right) \\ &\leq 2d\,\mathbb{P}\left(\max_{B+1 \leq b \leq B^*}\max_{0 \leq r \leq J} T^*_{br}(x)^+_j > \max_{1 \leq b \leq B}\max_{0 \leq r \leq J} T^*_{br}(x)^+_j\right) \\ &\leq 2(B^* - B)d(J+1)\,\mathbb{P}\left(T^*_{br}(x)^+_j > \max_{1 \leq b \leq B}\max_{0 \leq r \leq J} T^*_{br}(x)^+_j\right) \\ &\lesssim \frac{(B^* - B)d(J+1)}{B} \lesssim \frac{kd(J+1)}{n}, \end{aligned}$$

since $\frac{B^* - B}{B} \leq \left(\frac{n+k}{n}\right)^\xi - 1 \leq \frac{k}{n}$. The average case complexity of calculating $I^*(x)$ is therefore

$$\begin{aligned} &\mathbb{E}\left[d|I(x)| + dk + dn\mathbb{I}\{U^*(x) \nsubseteq U(x)\}\right] \\ &\quad\leq d\,\mathbb{E}\left[|I(x)|\right] + dk + dn\,\mathbb{P}\left(U^*(x) \nsubseteq U(x)\right) \lesssim \frac{nd\log(2B(J+1))^d}{\lambda^d} + kd^2(J+1). \end{aligned}$$

A similar calculation shows the cost of calculating all of $N^*_{br}(x)$, $S^*_{br}(x)$, and $V^*_{br}(x)$ is at most

$$\begin{aligned}
&\sum_{b=1}^{B}\sum_{r=0}^{J}\mathbb{E}\left[\frac{dk}{n}|I^*(x)| + d|I^*(x)|\,\mathbb{I}\{T_{br}(x) \neq T^*_{br}(x)\}\right] + \sum_{b=B+1}^{B^*}\sum_{r=0}^{J}\mathbb{E}\left[d|I^*(x)|\right] \\
&\quad \lesssim \mathbb{E}\left[\frac{B(J+1)dk}{n}|I^*(x)| + d|I^*(x)|\sum_{b=1}^{B}\sum_{r=0}^{J}\mathbb{I}\{T_{br}(x) \neq T^*_{br}(x)\} + d(B^*-B)(J+1)|I^*(x)|\right] \\
&\quad \lesssim \frac{Bkd(J+1)\log(2B(J+1))^d}{\lambda^d}.
\end{aligned}$$

where we used that the bounds for $|I(x)|$ and $\sum_{r=0}^{J}\sum_{b=1}^{B}\mathbb{I}\{T^*_{br}(x) \neq T_{br}(x)\}$ hold also in L^2 and applied the Cauchy–Schwarz inequality.

Finally, updating $\hat{\Sigma}_{\mathrm{d}}(x)$ takes $Bd(J+1)|I(x)|$ computations, which is done with probability at most k/K, yielding a time complexity of $\frac{nBd(J+1)\log(2B(J+1))^d}{K\lambda^d}$ on average. The overall average case time complexity is therefore bounded by

$$d(J+1)\left(\frac{knd(J+1)}{K} + kd + B\right) + \frac{d(J+1)\log(2B(J+1))^d}{\lambda^d}\left(n + Bk + \frac{nB}{K}\right).$$

□

B Additional empirical results

Tables 3, 4, 5 and 6 present some additional empirical results not given in the main paper. The data generating process is identical to that in Section 5, and we demonstrate here the effect of a smaller forest size B, taking $B = 1$ in Tables 3 and 4, and $B = 10$ in Tables 5 and 6. Note that the bias in Table 3 is not significantly larger than that in Table 1, even though Lemma 1 suggests that the bias should be much greater. This is because Lemma 1 is stated for the *conditional* bias. In fact, the repeated experiments (we use 3000 independent trials) have the same effect as using a large forest in reducing the apparent bias of the estimator. As such, the error incurred appears in the standard deviation column instead; indeed the standard deviations in Table 3 are substantially higher than those in Table 1.

	J	LS	LM	λ	RMSE	Bias	SD	Bias/SD	$\widehat{\text{SD}}$	$\hat{\sigma}^2$	ARMSE	ABias	ASD	CR	CIW
No debiasing	0	$\hat{\lambda}_0$	1.0	14.72	0.0606	-0.0235	0.0558	0.4211	0.0322	0.0907	0.0358	-0.0259	0.0232	77.1%	0.126
		λ_0	1.2	23.10	0.0526	-0.0093	0.0518	0.1792	0.0397	0.0883	0.0306	-0.0092	0.0292	88.9%	0.156
			1.1	21.18	0.0565	-0.0102	0.0555	0.1828	0.0381	0.0886	0.0300	-0.0110	0.0279	87.2%	0.149
			1.0	19.25	0.0551	-0.0150	0.0530	0.2834	0.0361	0.0890	0.0298	-0.0133	0.0266	84.9%	0.142
			0.9	17.33	0.0504	-0.0157	0.0479	0.3268	0.0345	0.0895	0.0301	-0.0164	0.0253	83.3%	0.135
			0.8	15.40	0.0568	-0.0202	0.0530	0.3814	0.0323	0.0902	0.0316	-0.0208	0.0238	79.2%	0.127
Debiasing	1	$\hat{\lambda}_1$	1.0	11.20	0.1150	-0.0058	0.1149	0.0508	0.0633	0.1184	0.0302	-0.0031	0.0287	83.0%	0.248
		λ_1	1.2	7.86	0.1274	-0.0046	0.1274	0.0362	0.0568	0.1389	0.0245	-0.0038	0.0242	71.7%	0.223
			1.1	7.21	0.1360	-0.0097	0.1357	0.0719	0.0546	0.1471	0.0238	-0.0053	0.0232	66.1%	0.214
			1.0	6.55	0.1465	-0.0147	0.1457	0.1008	0.0531	0.1551	0.0235	-0.0078	0.0221	63.5%	0.208
			0.9	5.90	0.1595	-0.0150	0.1587	0.0946	0.0543	0.1692	0.0241	-0.0119	0.0210	60.1%	0.213
			0.8	5.24	0.1799	-0.0256	0.1781	0.1439	0.0531	0.1862	0.0275	-0.0191	0.0198	54.0%	0.208
Robust BC	1	$\hat{\lambda}_0$	1.0	14.72	0.1156	-0.0004	0.1156	0.0036	0.0702	0.1111	0.0332	-0.0006	0.0330	90.1%	0.275
		λ_0	1.2	23.10	0.1187	-0.0042	0.1186	0.0353	0.0855	0.1044	0.0415	-0.0001	0.0415	96.2%	0.335
			1.1	21.18	0.1093	-0.0011	0.1093	0.0101	0.0842	0.1025	0.0398	-0.0001	0.0398	95.5%	0.330
			1.0	19.25	0.1006	-0.0026	0.1005	0.0259	0.0808	0.1020	0.0379	-0.0001	0.0379	94.4%	0.317
			0.9	17.33	0.0873	0.0007	0.0873	0.0077	0.0751	0.1012	0.0360	-0.0002	0.0360	93.4%	0.294
			0.8	15.40	0.1032	-0.0018	0.1032	0.0174	0.0710	0.1050	0.0339	-0.0003	0.0339	92.3%	0.278

Table 3: Simulation results with $d = 1$, $n = 1000$, and $B = 1$, over 3000 repeats

	J	LS	LM	λ	RMSE	Bias	SD	Bias/SD	$\widehat{\text{SD}}$	$\hat{\sigma}^2$	ARMSE	ABias	ASD	CR	CIW
No debiasing	0	$\hat{\lambda}_0$	1.0	12.32	0.3192	-0.1082	0.3003	0.3603	0.0709	0.0846	0.0831	-0.0670	0.0478	64.2%	0.278
		λ_0	1.2	18.39	0.5522	-0.1706	0.5252	0.3248	0.0788	0.0658	0.0771	-0.0292	0.0714	67.1%	0.309
			1.1	16.85	0.4737	-0.1344	0.4542	0.2960	0.0780	0.0707	0.0741	-0.0347	0.0654	70.3%	0.306
			1.0	15.32	0.4220	-0.1178	0.4052	0.2908	0.0764	0.0743	0.0728	-0.0420	0.0595	69.6%	0.299
			0.9	13.79	0.3473	-0.1027	0.3318	0.3096	0.0746	0.0796	0.0746	-0.0519	0.0535	70.4%	0.293
			0.8	12.26	0.3137	-0.1021	0.2966	0.3442	0.0702	0.0835	0.0811	-0.0657	0.0476	65.5%	0.275
Debiasing	1	$\hat{\lambda}_1$	1.0	9.22	0.7451	-0.1351	0.7328	0.1843	0.2742	0.6519	0.0731	-0.0093	0.0692	85.5%	1.075
		λ_1	1.2	7.18	0.5078	-0.0781	0.5018	0.1556	0.2272	0.4124	0.0550	-0.0108	0.0540	81.7%	0.891
			1.1	6.58	0.4920	-0.0686	0.4872	0.1408	0.2227	0.4197	0.0518	-0.0154	0.0495	79.7%	0.873
			1.0	5.99	0.4391	-0.0700	0.4335	0.1616	0.2083	0.3817	0.0503	-0.0225	0.0450	74.9%	0.816
			0.9	5.39	0.4185	-0.0725	0.4121	0.1759	0.1954	0.3862	0.0530	-0.0343	0.0405	72.5%	0.766
			0.8	4.79	0.3645	-0.0789	0.3559	0.2217	0.1766	0.3631	0.0656	-0.0549	0.0360	68.5%	0.692
Robust BC	1	$\hat{\lambda}_0$	1.0	12.32	0.9612	-0.1957	0.9411	0.2079	0.3495	0.9536	0.0926	-0.0014	0.0926	88.3%	1.370
		λ_0	1.2	18.39	1.3984	-0.4142	1.3357	0.3101	0.5041	1.8378	0.1381	-0.0003	0.1381	80.1%	1.976
			1.1	16.85	1.2900	-0.3571	1.2396	0.2881	0.4627	1.5834	0.1266	-0.0004	0.1266	83.3%	1.814
			1.0	15.32	1.1782	-0.3052	1.1379	0.2682	0.4299	1.3422	0.1151	-0.0005	0.1151	85.4%	1.685
			0.9	13.79	1.0995	-0.2733	1.0650	0.2567	0.3901	1.2125	0.1036	-0.0008	0.1036	87.7%	1.529
			0.8	12.26	0.9288	-0.1874	0.9097	0.2060	0.3361	0.8958	0.0921	-0.0013	0.0921	89.3%	1.318

Table 4: Simulation results with $d = 2$, $n = 1000$, and $B = 1$, over 3000 repeats

	J	LS	LM	λ	RMSE	Bias	SD	Bias/SD	$\widehat{\text{SD}}$	$\hat{\sigma}^2$	ARMSE	ABias	ASD	CR	CIW
No debiasing	0	$\hat{\lambda}_0$	1.0	14.72	0.0376	-0.0241	0.0289	0.8324	0.0250	0.0931	0.0363	-0.0263	0.0232	79.9%	0.098
		λ_0	1.2	23.10	0.0325	-0.0089	0.0313	0.2837	0.0310	0.0895	0.0306	-0.0092	0.0292	93.8%	0.122
			1.1	21.18	0.0317	-0.0104	0.0300	0.3462	0.0296	0.0898	0.0300	-0.0110	0.0279	92.6%	0.116
			1.0	19.25	0.0320	-0.0126	0.0293	0.4306	0.0285	0.0903	0.0298	-0.0133	0.0266	91.5%	0.112
			0.9	17.33	0.0324	-0.0153	0.0286	0.5368	0.0270	0.0907	0.0301	-0.0164	0.0253	89.0%	0.106
			0.8	15.40	0.0334	-0.0200	0.0268	0.7464	0.0255	0.0917	0.0316	-0.0208	0.0238	85.6%	0.100
Debiasing	1	$\hat{\lambda}_1$	1.0	11.05	0.0434	-0.0029	0.0433	0.0678	0.0350	0.1023	0.0297	-0.0027	0.0285	89.0%	0.137
		λ_1	1.2	7.86	0.0458	-0.0062	0.0454	0.1376	0.0316	0.1138	0.0245	-0.0038	0.0242	82.0%	0.124
			1.1	7.21	0.0484	-0.0088	0.0476	0.1843	0.0307	0.1181	0.0238	-0.0053	0.0232	78.2%	0.120
			1.0	6.55	0.0511	-0.0119	0.0497	0.2404	0.0301	0.1239	0.0235	-0.0078	0.0221	75.3%	0.118
			0.9	5.90	0.0572	-0.0187	0.0541	0.3466	0.0291	0.1299	0.0241	-0.0119	0.0210	68.5%	0.114
			0.8	5.24	0.0641	-0.0249	0.0591	0.4209	0.0284	0.1389	0.0275	-0.0191	0.0198	63.0%	0.111
Robust BC	1	$\hat{\lambda}_0$	1.0	14.72	0.0419	-0.0010	0.0419	0.0229	0.0393	0.0951	0.0334	-0.0009	0.0330	93.7%	0.154
		λ_0	1.2	23.10	0.0500	-0.0001	0.0500	0.0014	0.0483	0.0910	0.0415	-0.0001	0.0415	95.4%	0.189
			1.1	21.18	0.0483	0.0009	0.0483	0.0193	0.0468	0.0912	0.0398	-0.0001	0.0398	95.4%	0.183
			1.0	19.25	0.0460	0.0008	0.0460	0.0164	0.0445	0.0917	0.0379	-0.0001	0.0379	95.2%	0.174
			0.9	17.33	0.0439	0.0006	0.0439	0.0134	0.0424	0.0923	0.0360	-0.0002	0.0360	94.7%	0.166
			0.8	15.40	0.0424	0.0004	0.0424	0.0084	0.0400	0.0932	0.0339	-0.0003	0.0339	93.9%	0.157

Table 5: Simulation results with $d = 1$, $n = 1000$, and $B = 10$, over 3000 repeats

	J	LS	LM	λ	RMSE	Bias	SD	Bias/SD	$\widehat{\text{SD}}$	$\hat{\sigma}^2$	ARMSE	ABias	ASD	CR	CIW
No debiasing	0	$\hat{\lambda}_0$	1.0	12.27	0.0893	-0.0648	0.0614	1.0566	0.0551	0.0979	0.0836	-0.0678	0.0476	72.7%	0.216
		λ_0	1.2	18.39	0.0857	-0.0310	0.0799	0.3884	0.0707	0.0866	0.0771	-0.0292	0.0714	87.7%	0.277
			1.1	16.85	0.0826	-0.0346	0.0750	0.4619	0.0675	0.0881	0.0741	-0.0347	0.0654	87.9%	0.265
			1.0	15.32	0.0819	-0.0435	0.0694	0.6267	0.0635	0.0906	0.0728	-0.0420	0.0595	85.2%	0.249
			0.9	13.79	0.0815	-0.0505	0.0640	0.7891	0.0595	0.0934	0.0746	-0.0519	0.0535	81.4%	0.233
			0.8	12.26	0.0863	-0.0626	0.0593	1.0557	0.0553	0.0972	0.0811	-0.0657	0.0476	75.0%	0.217
Debiasing	1	$\hat{\lambda}_1$	1.0	9.24	0.1034	-0.0138	0.1024	0.1348	0.1017	0.1356	0.0723	-0.0082	0.0694	92.4%	0.399
		λ_1	1.2	7.18	0.0959	-0.0221	0.0933	0.2368	0.0925	0.1577	0.0550	-0.0108	0.0540	88.8%	0.363
			1.1	6.58	0.0964	-0.0284	0.0921	0.3082	0.0901	0.1698	0.0518	-0.0154	0.0495	87.1%	0.353
			1.0	5.99	0.1007	-0.0378	0.0934	0.4046	0.0855	0.1831	0.0503	-0.0225	0.0450	83.3%	0.335
			0.9	5.39	0.1079	-0.0503	0.0955	0.5269	0.0814	0.2002	0.0530	-0.0343	0.0405	77.4%	0.319
			0.8	4.79	0.1208	-0.0748	0.0949	0.7879	0.0740	0.2138	0.0656	-0.0549	0.0360	68.3%	0.290
Robust BC	1	$\hat{\lambda}_0$	1.0	12.27	0.1231	-0.0049	0.1230	0.0399	0.1182	0.1146	0.0923	-0.0015	0.0922	95.1%	0.463
		λ_0	1.2	18.39	0.1577	-0.0036	0.1576	0.0226	0.1361	0.1058	0.1381	-0.0003	0.1381	92.7%	0.534
			1.1	16.85	0.1491	-0.0021	0.1491	0.0139	0.1338	0.1065	0.1266	-0.0004	0.1266	93.4%	0.525
			1.0	15.32	0.1404	-0.0012	0.1404	0.0088	0.1287	0.1071	0.1151	-0.0005	0.1151	94.3%	0.504
			0.9	13.79	0.1294	-0.0018	0.1294	0.0138	0.1231	0.1093	0.1036	-0.0008	0.1036	95.3%	0.482
			0.8	12.26	0.1179	-0.0028	0.1178	0.0240	0.1172	0.1130	0.0921	-0.0013	0.0921	95.6%	0.459

Table 6: Simulation results with $d = 2$, $n = 1000$, and $B = 10$, over 3000 repeats

References

Abdulsalam, H., Skillicorn, D. B., and Martin, P. (2010). Classification using streaming random forests. *IEEE Transactions on Knowledge and Data Engineering*, 23(1):22–36.

Arnould, L., Boyer, C., and Scornet, E. (2023). Is interpolation benign for random forest regression? In *International Conference on Artificial Intelligence and Statistics*, pages 5493–5548. PMLR.

Baptista, R., O'Reilly, E., and Xie, Y. (2024). TrIM: Transformed iterative Mondrian forests for gradient-based dimension reduction and high-dimensional regression. *arXiv:2407.09964*.

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13:1063–1095.

Birgé, L. (2001). An alternative point of view on Lepski's method. *Lecture Notes – Monograph Series*, 36:113–133.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2018). On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association*, 113(522):767–779.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2022). Coverage error optimal confidence intervals for local polynomial regression. *Bernoulli*, 28(4):2998–3022.

Caruana, R., Niculescu-Mizil, A., Crew, G., and Ksikes, A. (2004). Ensemble selection from libraries of models. In *Proceedings of the Twenty-First International Conference on Machine Learning*.

Cattaneo, M. D., Chandak, R., and Klusowski, J. M. (2024a). Convergence rates of oblique regression trees for flexible function libraries. *The Annals of Statistics*, 52(2):466–490.

Cattaneo, M. D., Farrell, M. H., and Feng, Y. (2020). Large sample properties of partitioning-based series estimators. *The Annals of Statistics*, 48(3):1718–1741.

Cattaneo, M. D., Klusowski, J. M., and Tian, P. M. (2024b). On the pointwise behavior of recursive partitioning and its implications for heterogeneous causal effect estimation. *arXiv:2211.10805*.

Chi, C.-M., Vossler, P., Fan, Y., and Lv, J. (2022). Asymptotic properties of high-dimensional random forests. *The Annals of Statistics*, 50(6):3415–3438.

Duroux, R. and Scornet, E. (2018). Impact of subsampling and tree depth on random forests. *ESAIM: Probability and Statistics*, 22:96–128.

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596.

Gao, W., Xu, F., and Zhou, Z.-H. (2022). Towards convergence rate analysis of random forests for classification. *Artificial Intelligence*, 313:103788.

Goldenshluger, A. and Lepski, O. (2008). Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(4):1150–1190.

Gomes, H. M., Bifet, A., Read, J., Barddal, J. P., Enembreck, F., Pfharinger, B., Holmes, G., and Abdessalem, T. (2017). Adaptive random forests for evolving data stream classification. *Machine Learning*, 106:1469–1495.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics*, 31(5):1600–1635.

Ibragimov, I. A. and Linnik, Y. V. (1975). *Independent and Stationary Sequences of Random Variables*. Wolters–Noordhoff, Groningen.

Klusowski, J. (2021). Sharp analysis of a simple model for random forests. In *International Conference on Artificial Intelligence and Statistics*, pages 757–765. PMLR.

Klusowski, J. M. and Tian, P. (2024). Large scale prediction with decision trees. *Journal of the American Statistical Association*, 119(545):525–537.

Krauss, C., Do, X. A., and Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, 259(2):689–702.

Lafferty, J. and Wasserman, L. (2008). Rodeo: Sparse, greedy nonparametric regression. *The Annals of Statistics*, 36(1):28–63.

Lakshminarayanan, B., Roy, D. M., and Teh, Y. W. (2014). Mondrian forests: Efficient online random forests. *Advances in Neural Information Processing Systems*, 27.

Lepski, O. V. (1992). Asymptotically minimax adaptive estimation I: Upper bounds. optimally adaptive estimates. *Theory of Probability & its Applications*, 36(4):682–697.

Li, X., Wang, Z., Wang, L., Hu, R., and Zhu, Q. (2018). A multi-dimensional context-aware recommendation approach based on improved random forest algorithm. *IEEE Access*, 6:45071–45085.

Ma, H., Ghojogh, B., Samad, M. N., Zheng, D., and Crowley, M. (2020). Isolation Mondrian forest for batch and online anomaly detection. In *2020 IEEE International Conference on Systems, Man, and Cybernetics*, pages 3051–3058. IEEE.

Martindale, N., Ismail, M., and Talbert, D. A. (2020). Ensemble-based online machine learning algorithms for network intrusion detection systems using streaming data. *Information*, 11(6):315.

Miftachov, R. and Reiß, M. (2025). Early stopping for regression trees. *arXiv:2502.04709*.

Mourtada, J., Gaïffas, S., and Scornet, E. (2020). Minimax optimal rates for Mondrian trees and forests. *The Annals of Statistics*, 48(4):2253–2276.

Mourtada, J., Gaïffas, S., and Scornet, E. (2021). AMF: Aggregated Mondrian forests for online learning. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(3):505–533.

Näf, J., Emmenegger, C., Bühlmann, P., and Meinshausen, N. (2023). Confidence and uncertainty assessment for distributional random forests. *Journal of Machine Learning Research*, 24(366):1–77.

O'Reilly, E. (2024). Statistical advantages of oblique randomized decision trees and forests. *arXiv: 2407.02458*.

O'Reilly, E. and Tran, N. M. (2022). Stochastic geometry to generalize the Mondrian process. *SIAM Journal on Mathematics of Data Science*, 4(2):531–552.

O'Reilly, E. and Tran, N. M. (2024). Minimax rates for high-dimensional random tessellation forests. *Journal of Machine Learning Research*, 25(89):1–32.

Osborne, C. and O'Reilly, E. (2025). The uniformly rotated Mondrian kernel. *arXiv:2502.04323*.

Petrov, V. V. (1995). *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*. Oxford University Press, New York.

Roy, D. M., Teh, Y. W., et al. (2008). The Mondrian process. In *Neural Information Processing Systems*, volume 21.

Schucany, W. R. and Sommers, J. P. (1977). Improvement of kernel type density estimators. *Journal of the American Statistical Association*, 72(358):420–423.

Scillitoe, A., Seshadri, P., and Girolami, M. (2021). Uncertainty quantification for data-driven turbulence modelling with Mondrian forests. *Journal of Computational Physics*, 430:110116.

Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053.

Vicuna, M., Khannouz, M., Kiar, G., Chatelain, Y., and Glatard, T. (2021). Reducing numerical precision preserves classification accuracy in Mondrian forests. In *2021 IEEE International Conference on Big Data*, pages 2785–2790. IEEE.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Zhan, H., Wang, J., and Xia, Y. (2024). Non-asymptotic properties of generalized Mondrian forests in statistical learning. *arXiv:2406.00660*.

Zhou, Z.-H. and Feng, J. (2019). Deep forest. *National Science Review*, 6(1):74–86.