Inference with Mondrian Random Forests

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

October 15, 2023

Abstract

Random forests are popular methods for classification and regression, and many different variants have been proposed in recent years. One interesting example is the Mondrian random forest, in which the underlying trees are constructed according to a Mondrian process. In this paper we give a central limit theorem for the estimates made by a Mondrian random forest in the regression setting. When combined with a bias characterization and a consistent variance estimator, this allows one to perform asymptotically valid statistical inference, such as constructing confidence intervals, on the unknown regression function. We also provide a debiasing procedure for Mondrian random forests which allows them to achieve minimax-optimal estimation rates with β-Hölder regression functions, for all β and in arbitrary dimension, assuming appropriate parameter tuning.

Keywords: Random forests, regression trees, central limit theorem, bias correction, statistical inference, minimax rates, nonparametric estimation.

[1]Department of Operations Research and Financial Engineering, Princeton University

[*]Corresponding author: wgu2@princeton.edu

Contents

1 Introduction

Random forests, first introduced by Breiman (2001), are a workhorse in modern machine learning for classification and regression tasks. Their desirable traits include computational efficiency (via parallelization and greedy heuristics) in big data settings, 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. Random forests have achieved great empirical successes in many fields of study, including healthcare, finance, online commerce, 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: see Scornet et al. (2015), Chi et al. (2022), Klusowski and Tian (2023), and references therein, for a sample of recent developments. 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.

1.1 Prior work

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. These include centered random forests (Biau, 2012; Arnould et al., 2023) and median random forests (Duroux and Scornet, 2016; 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—albeit 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 in the statistics and machine learning communities in recent years (Ma et al., 2020; Mourtada et al., 2020; Scillitoe et al., 2021; Mourtada et al., 2021; Vicuna et al., 2021; Gao et al., 2022; O'Reilly and Tran, 2022). 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 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 as well as being subject to a rigorous statistical analysis, while also retaining some of the appealing features of other random forest methods.

This paper studies the statistical properties of Mondrian random forests. 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 is, to date, the only known recursive tree mechanism involving randomization, pure or data-dependent, for which the resulting random forest is minimax-optimal for point estimation over a class of smooth regression functions in arbitrary dimension (Mourtada et al., 2020). 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 precision and confidence.

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 for 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.

1.2 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, and propose valid large-sample inference procedures employing a consistent standard error estimator. We establish this result by deploying a martingale central limit theorem (Hall and Heyde, 2014, Theorem 3.2) 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 a single Mondrian tree) remain random even in the limit due to the lack of ensembling. The presence of "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 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.

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 inference procedures based on robust bias correction (Calonico et al., 2018, 2022).

1.3 Organization

Our paper is structured as follows. In Section 2 we introduce the Mondrian process and give our assumptions on the data generating process, using a Hölder smoothness condition on the regression function to control the bias of various estimators. We define the Mondrian random forest estimator and give our assumptions on its lifetime parameter and the number of trees. We give our notation for the following sections.

Section 3 presents our first set of main results, beginning with a central limit theorem for the centered Mondrian random forest estimator (Theorem 1) in which we characterize the limiting variance. Theorem 2 complements this result by precisely calculating the limiting bias of the estimator, with the aim of subsequently applying a debiasing procedure. To enable valid feasible statistical inference, we provide a consistent variance estimator in Theorem 3 and briefly discuss implications for lifetime parameter selection.

In Section 5 we define debiased Mondrian random forests, a collection 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 provably superior bias properties (Theorem 6). In Theorem 5 we verify that a central limit theorem continues to hold for the debiased Mondrian random forest. We again state the limiting variance, discuss the implications for the lifetime parameter, and provide a consistent variance estimator (Theorem 7) for constructing confidence intervals (Theorem 8). As a final corollary of the improved bias properties, we demonstrate in Theorem 9 that the debiased Mondrian random forest estimator is minimax-optimal in pointwise mean squared error for all $\beta > 0$, provided that β is known a priori.

Section 6 discusses tuning parameter selection, beginning with a data-driven approach to selecting the crucial lifetime parameter using local polynomial estimation, alongside other practical suggestions. We also give advice on choosing the number of trees, and other parameters associated with the debiasing procedure.

Concluding remarks are given in Section 7, while Appendix A contains all the mathematical proofs of our theoretical results.

1.4 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 a set A, we use A^{c} for the complement whenever the background space is clear from context. We use C to denote a positive constant whose value may change from line to line. For non-negative sequences a_n and b_n, write $a_n \lesssim b_n$ or $a_n = O(b_n)$ to indicate that a_n/b_n is bounded for $n \geq 1$. Write $a_n \ll b_n$ or $a_n = o(b_n)$ if $a_n/b_n \to 0$. 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 $A_n \lesssim_{\mathbb{P}} B_n$ or $A_n = O_{\mathbb{P}}(B_n)$ if A_n/B_n is bounded in probability, and $A_n = o_{\mathbb{P}}(B_n)$ if $A_n/B_n \to 0$ in probability. Convergence of random variables X_n in distribution to a law $\mathbb{P}$ is denoted by $X_n \rightsquigarrow \mathbb{P}$.

2 Setup

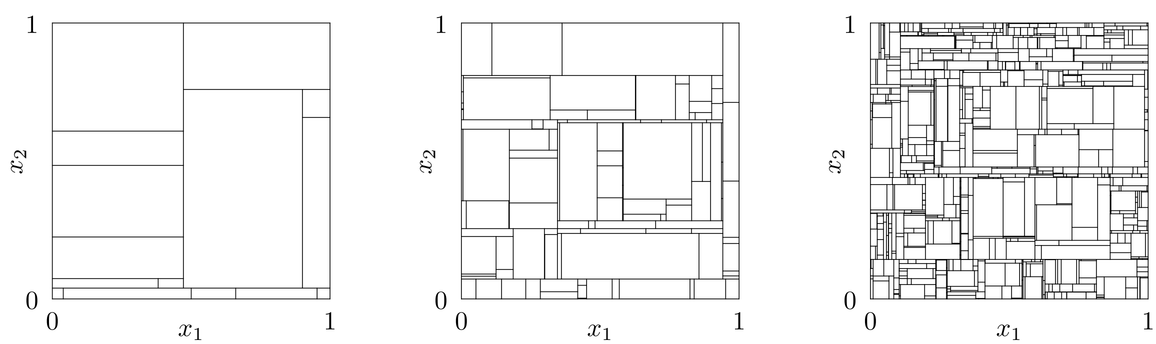
When using a Mondrian random forest, there are two sources of randomness. The first is of course the data, and here we consider the nonparametric regression setting with d-dimensional covariates. The second source is a collection of independent trees drawn from a Mondrian process, which we define in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

2.1 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 definition.

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 $\mathcal{M}\big([0,1]^d, \lambda\big)$ we start at time $t = 0$ with the trivial partition of $[0,1]^d$ which has no splits. We 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 $E_D \sim \operatorname{Exp}(|D|_1)$. If $t_D + E_D \leq \lambda$, then we 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 $S_J \sim \operatorname{Unif}\big[D_J^-, D_J^+\big]$. 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 final outcome 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.

2.2 Data generation

Throughout this paper, we assume that the data satisfies Assumption 1. We begin with a definition of Hölder continuity which will be used for controlling the bias of various estimators.

Definition 1 (Hölder continuity)
Take $\beta > 0$ and define $\underline{\beta}$ to be the largest integer which is 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 $\partial^\nu g(x) = \partial^{|\nu|} g(x) / \prod_{j=1}^d \partial x_j^{\nu_j}$. We say g is Lipschitz if $g \in \mathcal{H}^1$.

Assumption 1 (Data generation)
Fix $d \geq 1$ and let (X_i, Y_i) be 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 a Lebesgue density function $f(x)$ on $[0,1]^d$ which is bounded away from zero and satisfies $f \in \mathcal{H}^\beta$ for some $\beta \geq 1$. Suppose $\mathbb{E}[Y_i^2 \mid X_i]$ is bounded, let $\mu(X_i) = \mathbb{E}[Y_i \mid X_i]$ and assume $\mu \in \mathcal{H}^\beta$. Write $\varepsilon_i = Y_i - \mu(X_i)$ and assume $\sigma^2(X_i) = \mathbb{E}[\varepsilon_i^2 \mid X_i]$ is Lipschitz and bounded away from zero.

Some comments are in order surrounding Assumption 1. The requirement that the covariate density $f(x)$ be strictly positive on all of $[0,1]^d$ may seem strong, particularly when d is moderately large. However, since our theory is presented pointwise in x, it is sufficient for this 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, we can map this hypercube onto $[0,1]^d$. The same argument of course holds for the Hölder smoothness assumptions and the upper and lower bounds on the conditional variance function.

2.3 Mondrian random forests

We define the basic 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 5. 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 $N_b(x) = \sum_{i=1}^n \mathbb{I}\left\{X_i \in T_b(x)\right\}$ 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 an interior point of the support of X_i, $\frac{\lambda^d}{n} \to 0$, $\log \lambda \asymp \log n$, and $B \asymp n^\xi$ for some $\xi \in (0,1)$, which may depend on the dimension d and smoothness β.

Assumption 2 implies that the size of the forest B grows with n. For the purpose of mitigating the computational burden, we suggest the sub-linear polynomial growth $B \asymp n^\xi$, satisfying the conditions imposed in our main results. Large forests usually do not present computational challenges in practice as the ensemble estimator is easily parallelizable over the trees. We emphasize places where this "large forest" condition is important to our theory as they arise throughout the paper.

3 Inference with Mondrian random forests

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

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

Our approach to inference is summarized as follows. Firstly, we provide a central limit theorem (weak convergence to a Gaussian) for the first "variance" term in (2). Secondly, we precisely compute the probability limit of the second "bias" term. By ensuring that the standard deviation dominate the bias, we can conclude that a corresponding central limit theorem holds for the Mondrian random forest. With an appropriate estimator for the limiting variance, we establish procedures for valid and feasible statistical inference on the unknown regression function $\mu(x)$.

We begin with the aforementioned central limit theorem, which forms the core of our methodology for performing statistical inference. Before stating our main result, we highlight some of the challenges involved. At first glance, the summands in the first term in (2) seem to be independent over $1 \leq i \leq n$, 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 $\mathbb{E}\left[\frac{1}{|T_b(x)|^2}\right] = \infty$; 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 Lyapunov central limit theorem, are ineffective.

We circumvent these problems by directly analyzing $\frac{\mathbb{I}\{N_b(x) \geq 1\}}{N_b(x)}$. We establish concentration properties for this non-linear function of X_i via the Efron–Stein inequality (Boucheron et al., 2016, Section 3.1) along with a sequence of somewhat delicate preliminary lemmas regarding inverse moments of truncated (conditional) binomial random variables. In particular, we show that $\mathbb{E}\left[\frac{\mathbb{I}\{N_b(x) \geq 1\}}{N_b(x)}\right] \lesssim \frac{\lambda^d}{n}$ and $\mathbb{E}\left[\frac{\mathbb{I}\{N_b(x) \geq 1\}}{N_b(x)^2}\right] \lesssim \frac{\lambda^{2d} \log n}{n^2}$. Asymptotic normality is then established using a central limit theorem for martingale difference sequences (Hall and Heyde, 2014, Theorem 3.2) with respect to an appropriate filtration. Section 4 gives an overview our proof strategy in which we further discuss the underlying challenges, while Appendix A gives all the technical details.

3.1 Central limit theorem

Theorem 1 gives our first main result.

Theorem 1 (Central limit theorem for the centered Mondrian random forest estimator)
Suppose Assumptions 1 and 2 hold, $\mathbb{E}[Y_i^4 \mid X_i]$ is bounded almost surely, and $\frac{\lambda^d \log n}{n} \to 0$. Then

$$\sqrt{\frac{n}{\lambda^d}} \Big(\hat{\mu}(x) - \mathbb{E}\big[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\big] \Big) \rightsquigarrow \mathcal{N}\big(0, \Sigma(x)\big) \qquad \text{where} \qquad \Sigma(x) = \frac{\sigma^2(x)}{f(x)} \left(\frac{4 - 4\log 2}{3} \right)^d.$$

The condition of $B \to \infty$ is crucial, 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 $\mathbb{E}\left[\frac{1}{|T_b(x)||T_{b'}(x)|}\right] = \mathbb{E}\left[\frac{1}{|T_b(x)|}\right] \mathbb{E}\left[\frac{1}{|T_{b'}(x)|}\right] \approx \lambda^{2d} < \infty$ where $b \neq b'$, by independence of the trees, rather than the "no ensembling" single tree analog $\mathbb{E}\left[\frac{1}{|T_b(x)|^2}\right] = \infty$.

We take this opportunity to contrast Mondrian random forests with classical kernel-based smoothing methods. The lifetime λ plays a similar role to the inverse bandwidth in determining the effective sample size n/λ^d, and thus the associated rate of convergence. 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. Similarly, classical methods based on non-random partitioning such as spline estimators (Huang, 2003; Cattaneo et al., 2020) typically impose a quasi-uniformity assumption to ensure all the cells are of comparable size, a property which does not hold for the Mondrian process (not even with probability approaching one).

Bias characterization

We turn to the second term in (2), which captures the bias of the Mondrian random forest estimator conditional on the covariates $\mathbf{X}$ and the forest $\mathbf{T}$. As such it is a random quantity which, as we will demonstrate, converges in probability. We precisely characterize the limiting non-random bias, including high-degree polynomials in λ which for now may seem ignorable. Indeed the magnitude of the bias is determined by its leading term, typically of order $1/\lambda^2$ whenever $\beta \geq 2$, and this suffices for ensuring a negligible contribution from the bias with an appropriate choice of lifetime parameter. However, the advantage of specifying higher-order bias terms is made apparent in Section 5 when we construct a debiased Mondrian random forest estimator. There, we target and annihilate the higher-order terms in order to furnish superior estimation and inference properties.

Theorem 2 gives our main result on the bias of the Mondrian random forest estimator.

Theorem 2 (Bias of the Mondrian random forest estimator)
Suppose Assumptions 1 and 2 hold. Then for each $1 \leq r \leq \lfloor \beta/2 \rfloor$ there exists $B_r(x) \in \mathbb{R}$, which is a function only of the derivatives of f and μ at x up to order $2r$, such that

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

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

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

In Theorem 2 we give some explicit examples of calculating the limiting bias if $\beta > 2$ or when 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, crucial facts for our debiasing procedure given in Section 5. If the forest size B does not diverge to infinity then we suffer the first-order bias term $\frac{1}{\lambda\sqrt{B}}$. This phenomenon was explained by Mourtada et al. (2020), who noted that it allows single Mondrian trees to achieve minimax optimality only when $\beta \in (0,1]$. In contrast, large forests remove this first-order bias and as such are optimal for all $\beta \in (0,2]$.

Using Theorem 1 and Theorem 2 together, along with an appropriate choice of lifetime parameter λ, gives a central limit theorem for the Mondrian random forest estimator which can be used, for example, to build confidence intervals for the unknown regression function $\mu(x)$ whenever the bias shrinks faster than the standard deviation. In general this will require $\frac{1}{\lambda^2} + \frac{1}{\lambda^\beta} + \frac{1}{\lambda\sqrt{B}} \ll \sqrt{\frac{\lambda^d}{n}}$, which can be satisfied by imposing the restrictions $\lambda \gg n^{\frac{1}{d+2(2\wedge\beta)}}$ and $B \gg n^{\frac{2(2\wedge\beta)-2}{d+2(2\wedge\beta)}}$ on the lifetime λ and forest size B. If instead we aim for optimal point estimation, then balancing the bias and standard deviation requires $\frac{1}{\lambda^2} + \frac{1}{\lambda^\beta} + \frac{1}{\lambda\sqrt{B}} \asymp \sqrt{\frac{\lambda^d}{n}}$, which can be satisfied by $\lambda \asymp n^{\frac{1}{d+2(2\wedge\beta)}}$ and $B \gtrsim n^{\frac{2(2\wedge\beta)-2}{d+2(2\wedge\beta)}}$. Such a choice of λ gives the convergence rate $n^{\frac{-(2\wedge\beta)}{d+2(2\wedge\beta)}}$ which is the minimax-optimal rate of convergence (Stone, 1982) for β-Hölder functions with $\beta \in (0,2]$ as shown by Mourtada et al. (2020, Theorem 2). In Section 5 we will show how the Mondrian random forest estimator can be debiased, giving both weaker lifetime conditions for inference and also improved rates of convergence, under additional smoothness assumptions.

Variance estimation

The limiting variance $\Sigma(x)$ from the resulting central limit theorem depends on the unknown quantities $\sigma^2(x)$ and $f(x)$. To conduct feasible inference, we must therefore first estimate $\Sigma(x)$. To this end, define

$$
\begin{aligned}
\hat{\sigma}^2(x) &= \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} \frac{\big(Y_i - \hat{\mu}(x)\big)^2 \, \mathbb{I}\{X_i \in T_b(x)\}}{N_b(x)}, \\
\hat{\Sigma}(x) &= \hat{\sigma}^2(x) \frac{n}{\lambda^d} \sum_{i=1}^{n} \left(\frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{I}\{X_i \in T_b(x)\}}{N_b(x)} \right)^2 .
\end{aligned} \tag{3}
$$

In Theorem 3 we show that this variance estimator is consistent and establish its rate of convergence.

Theorem 3 (Variance estimation)
Suppose Assumptions 1 and 2 hold, and that $\mathbb{E}[Y_i^4 \mid X_i]$ is bounded almost surely. Then

$$
\hat{\Sigma}(x) = \Sigma(x) + O_{\mathbb{P}}\left(\frac{(\log n)^{d+1}}{\lambda} + \frac{1}{\sqrt{B}} + \sqrt{\frac{\lambda^d \log n}{n}} \right).
$$

3.2 Confidence intervals

Theorem 4 shows how to construct valid confidence intervals for the regression function $\mu(x)$ under the lifetime and forest size assumptions previously discussed. For details on feasible and practical selection of the lifetime parameter λ, see Section 6.

Theorem 4 (Feasible confidence intervals using a Mondrian random forest)
Suppose Assumptions 1 and 2 hold, $\mathbb{E}[Y_i^4 \mid X_i]$ is bounded almost surely, and $\frac{\lambda^d \log n}{n} \to 0$. Assume that $\lambda \gg n^{\frac{1}{d+2(2\wedge\beta)}}$ and $B \gg n^{\frac{2(2\wedge\beta)-2}{d+2(2\wedge\beta)}}$. For a confidence level $\alpha \in (0,1)$, let $q_{1-\alpha/2}$ be the normal quantile satisfying $\mathbb{P}\left(\mathcal{N}(0,1) \le q_{1-\alpha/2}\right) = 1 - \alpha/2$. Then

$$
\mathbb{P}\left(\mu(x) \in \left[\hat{\mu}(x) - \sqrt{\frac{\lambda^d}{n}} \hat{\Sigma}(x)^{1/2} q_{1-\alpha/2}, \ \hat{\mu}(x) + \sqrt{\frac{\lambda^d}{n}} \hat{\Sigma}(x)^{1/2} q_{1-\alpha/2} \right] \right) \to 1 - \alpha.
$$

When coupled with an appropriate lifetime selection method, Theorem 4 gives a fully 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 weak convergence result given in Theorem 1, and follows immediately from Slutsky's theorem with a consistent variance estimator. Other applications include hypothesis testing on the value of $\mu(x)$ at a design point x by inversion of the confidence interval, as well as parametric 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 theorem and formulating a consistent multivariate variance estimator.

4 Overview of proof strategy

This section provides some insight into the general approach we use to establish the main results in the preceding sections. We focus on the technical innovations forming the core of our arguments, and refer the reader to Appendix A for detailed proofs, including those for the debiased estimator discussed in the upcoming Section 5.

Preliminary results

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

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

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 $\mathrm{Exp}(1)$ 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

$$|T(x)_j| = \frac{E_{j1} + E_{j2}}{\lambda}$$

with high probability. Controlling the size of the largest cell in the forest containing x is now straightforward with a union bound, exploiting the sharp tail decay of the exponential distribution, and thus

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

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 the upcoming bias characterization.

Having provided upper bounds for the sizes of Mondrian cells, we also must establish some lower bounds in order 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

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

These bounds are computed directly using the exact distribution of $|T(x)|$. Note that $\mathbb{E}\left[\frac{1}{|T(x)|^2}\right] = \infty$ because $\frac{1}{E_{j1}+E_{j2}}$ has only $2-\delta$ finite moments, so the truncation is crucial here. Since we "nearly" have 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 $2+\delta$ 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

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

as well as similar bounds for "mixed" terms such as $\mathbb{E}\left[\frac{\mathbb{I}\{N_b(x)\geq 1\}}{N_b(x)} \frac{\mathbb{I}\{N_{b'}(x)\geq 1\}}{N_{b'}(x)}\right] \lesssim \frac{\lambda^{2d}}{n^2}$ 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.

Central limit theorem

To prove our main central limit theorem result (Theorem 1), we use the martingale central limit theorem given by Hall and Heyde (2014, Theorem 3.2). For each $1 \leq i \leq n$ define $\mathcal{H}_{ni}$ to be the filtration generated by $\mathbf{T}$, $\mathbf{X}$ and $(\varepsilon_j : 1 \leq j \leq i)$, noting that $\mathcal{H}_{ni} \subseteq \mathcal{H}_{(n+1)i}$ because B increases as n increases. Define the $\mathcal{H}_{ni}$-measurable and square integrable 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 satisfy the martingale difference property $\mathbb{E}[S_i(x) \mid \mathcal{H}_{ni}] = 0$. Further,

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

To establish weak convergence to $\mathcal{N}\big(0, \Sigma(x)\big)$, it suffices to check that $\max_i |S_i(x)| \to 0$ in probability, $\mathbb{E}\left[\max_i S_i(x)^2\right] \lesssim 1$, and $\sum_i S_i(x)^2 \to \Sigma(x)$ in probability. Checking the first two of these is straightforward given the denominator moment bounds derived above. For the third condition, we demonstrate that $\sum_i S_i(x)^2$ concentrates by checking its variance is vanishing. To do this, first observe that $S_i(x)^2$ is the square of a sum over the B trees. Expanding this square, we see that the diagonal terms (where $b = b'$) provide a negligible contribution due to the large forest assumption. For the other terms, we apply the law of total variance and the moment bounds detailed earlier. Here, it is crucial that $b \neq b'$ in order to exploit the independence of the trees and avoid having to control any higher moments. The law of total variance requires that we bound

$$\operatorname{Var}\left[\mathbb{E}\left[\sum_{i=1}^{n}\sum_{b=1}^{B}\sum_{b'\neq b}\frac{\mathbb{I}\{X_i \in T_b(x)\cap T_{b'}(x)\}\varepsilon_i^2}{N_b(x)N_{b'}(x)} \;\Big|\; \mathbf{X}, \mathbf{Y}\right]\right],$$

which is the variance of a non-linear function of the i.i.d. variables (X_i, ε_i), and so we apply the Efron–Stein inequality. The important insight here is that replacing a sample (X_i, ε_i) with an independent copy $(\tilde{X}_i, \tilde{\varepsilon}_i)$ 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).

The final part of the central limit theorem proof is to calculate the limiting variance $\Sigma(x)$. The penultimate step showed that we must have

$$\Sigma(x) = \lim_{n\to\infty}\sum_{i=1}^{n}\mathbb{E}\left[S_i(x)^2\right] = \lim_{n\to\infty}\frac{n^2}{\lambda^d}\,\mathbb{E}\left[\frac{\mathbb{I}\{X_i \in T_b(x)\cap T_{b'}(x)\}\varepsilon_i^2}{N_b(x)N_{b'}(x)}\right],$$

assuming the limit exists, so it remains to check this and calculate the limit. It is a straightforward but tedious exercise to verify that each term can be replaced with its conditional expectation given T_b and $T_{b'}$, using some further properties of the binomial and exponential distributions. This yields

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

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

Bias characterization

Our second substantial technical result is the bias characterization given as Theorem 2, 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

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

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, the error $\frac{1}{\lambda\sqrt{B}}$ exactly accounts for the first-order bias of individual Mondrian trees noted by Mourtada et al. (2020).

Next we show that $\mathbb{E}[\hat{\mu}(x) \mid \mathbf{X}]$ converges in probability to its expectation, again using the Efron–Stein theorem for this non-linear function of the i.i.d. variables X_i. The Lipschitz property of μ and the upper bound on the maximum cell size give $|\mu(X_i) - \mu(x)| \lesssim \max_{1 \leq j \leq d} |T_b(x)_j| \lesssim_{\mathbb{P}} \frac{\log B}{\lambda}$ whenever $X_i \in T_b(x)$, so we combine this with moment bounds for the denominator $N_b(x)$ to see

$$|\mathbb{E}[\hat{\mu}(x) \mid \mathbf{X}] - \mathbb{E}[\hat{\mu}(x)]| \lesssim_{\mathbb{P}} \frac{\log n}{\lambda} \sqrt{\frac{\lambda^d}{n}}$$

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

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

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

$$\mathbb{E}[\hat{\mu}(x)] - \mu(x) \approx \mathbb{E}\left[\frac{\int_{T_b(x)} (\mu(s) - \mu(x)) f(s) \, \mathrm{d}s}{\int_{T_b(x)} f(s) \, \mathrm{d}s}\right]. \tag{4}$$

Next we apply the multivariate version of Taylor's theorem to the integrands in both the numerator and the denominator in (4), and then apply the Maclaurin series of $\frac{1}{1+x}$ 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.

5 Debiased Mondrian random forests

In this section we give our next main contribution, proposing a variant of the Mondrian random forest estimator which 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 estimator 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 Theorem 2 to construct a debiased version of the Mondrian forest estimator. Let $J \geq 0$ be the bias correction order. As such, with $J = 0$ we retain the original Mondrian forest estimator, 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 $\mathcal{H}^\beta$. 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 ω_r in order to annihilate the leading J bias terms of the debiased Mondrian random forest estimator

$$\hat{\mu}_{\mathrm{d}}(x) = \sum_{r=0}^{J} \omega_r \hat{\mu}_r(x) = \sum_{r=0}^{J} \omega_r \frac{1}{B} \sum_{b=1}^{B} \frac{\sum_{i=1}^n Y_i \, \mathbb{I}\big\{X_i \in T_{rb}(x)\big\}}{N_{rb}(x)}. \tag{5}$$

This ensemble estimator retains the "forest" structure of the original estimators, but with varying lifetime parameters λ_r and coefficients ω_r. Thus we want to solve

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

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 $e_0 = (1, 0, \ldots, 0)^\mathsf{T} \in \mathbb{R}^{J+1}$. 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, we 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.

5.1 Central limit theorem

In Theorem 5, we verify that a central limit theorem holds for the debiased random forest estimator $\hat{\mu}_\mathrm{d}(x)$ and give its limiting variance. The strategy and challenges associated with proving Theorem 5 are identical to those discussed earlier surrounding Theorem 1. In fact in Appendix A we provide a direct proof only for Theorem 5 and deduce Theorem 1 as a special case.

Theorem 5 (Central limit theorem for the debiased Mondrian random forest estimator)
Suppose Assumptions 1 and 2 hold, $\mathbb{E}[Y_i^4 \mid X_i]$ is bounded almost surely, and $\frac{\lambda^d \log n}{n} \to 0$. Then

$$\sqrt{\frac{n}{\lambda^d}} \Big(\hat{\mu}_\mathrm{d}(x) - \mathbb{E}\big[\hat{\mu}_\mathrm{d}(x) \mid \mathbf{X}, \mathbf{T}\big] \Big) \rightsquigarrow \mathcal{N}\big(0, \Sigma_\mathrm{d}(x)\big)$$

where, with $\ell_{rr'} = \frac{2a_r}{3} \left(1 - \frac{a_r}{a_{r'}} \log \left(\frac{a_{r'}}{a_r} + 1 \right) \right)$, the limiting variance of the debiased estimator is

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

It is easy to verify that in the case of no debiasing we have $J = 0$ and $a_0 = \omega_0 = 1$, yielding $\Sigma_\mathrm{d}(x) = \Sigma(x)$, and recovering Theorem 1.

Bias characterization

In Theorem 6 we verify that this debiasing procedure does indeed annihilate the desired bias terms, and its proof is a consequence of Theorem 2 and the construction of the debiased Mondrian random forest estimator $\hat{\mu}_\mathrm{d}(x)$.

Theorem 6 (Bias of the debiased Mondrian random forest estimator)
Suppose Assumptions 1 and 2 hold. Then in the notation of Theorem 2 and with $\bar{\omega} = \sum_{r=0}^{J} \omega_r a_r^{-2J-2}$,

$$\begin{aligned} \mathbb{E}\big[\hat{\mu}_\mathrm{d}(x) \mid \mathbf{X}, \mathbf{T}\big] &= \mu(x) + \mathbb{I}\{2J+2 < \beta\} \frac{\bar{\omega} B_{J+1}(x)}{\lambda^{2J+2}} \\ &\quad + O_\mathbb{P} \left(\frac{1}{\lambda^{2J+4}} + \frac{1}{\lambda^\beta} + \frac{1}{\lambda \sqrt{B}} + \frac{\log n}{\lambda} \sqrt{\frac{\lambda^d}{n}} \right). \end{aligned}$$

Theorem 6 has the following consequence: the leading bias term is characterized in terms of $B_{J+1}(x)$ whenever $J < \beta/2 - 1$, or equivalently $J < \lfloor \underline{\beta}/2 \rfloor$, 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 and the resulting leading bias term is bounded above by $1/\lambda^\beta$ up to constants but its form is left unspecified.

Variance estimation

As before, we propose a variance estimator in order to conduct feasible inference and show that it is consistent. With $\hat{\sigma}^2(x)$ as in (3) in Section 3, define the estimator

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

Theorem 7 (Variance estimation)
Suppose Assumptions 1 and 2 hold and that $\mathbb{E}[Y_i^4 \mid X_i]$ is bounded almost surely. Then

$$\hat{\Sigma}_{\mathrm{d}}(x) = \Sigma_{\mathrm{d}}(x) + O_{\mathbb{P}} \left(\frac{(\log n)^{d+1}}{\lambda} + \frac{1}{\sqrt{B}} + \sqrt{\frac{\lambda^d \log n}{n}} \right).$$

5.2 Confidence intervals

In analogy to Section 3, we now demonstrate the construction of feasible valid confidence intervals using the debiased Mondrian random forest estimator in Theorem 8. Once again 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 assume for simplicity that the estimator has been fully debiased by setting $J \geq \lfloor \underline{\beta}/2 \rfloor$ to yield a leading bias of order $1/\lambda^\beta$, but intermediate "partially debiased" versions can easily be provided, with leading bias terms of order $1/\lambda^{\beta \wedge (2J+2)}$ in general. We thus require $\frac{1}{\lambda^\beta} + \frac{1}{\lambda\sqrt{B}} \ll \sqrt{\frac{\lambda^d}{n}}$, which can be satisfied by imposing the restrictions $\lambda \gg n^{\frac{1}{d+2\beta}}$ and $B \gg n^{\frac{2\beta-2}{d+2\beta}}$ on the lifetime parameter λ and forest size B.

Theorem 8 (Feasible confidence intervals using a debiased Mondrian random forest)
Suppose Assumptions 1 and 2 hold, $\mathbb{E}[Y_i^4 \mid X_i]$ is bounded almost surely, and $\frac{\lambda^d \log n}{n} \to 0$. Fix $J \geq \lfloor \underline{\beta}/2 \rfloor$ and assume that $\lambda \gg n^{\frac{1}{d+2\beta}}$ and $B \gg n^{\frac{2\beta-2}{d+2\beta}}$. For a confidence level $\alpha \in (0,1)$, let $q_{1-\alpha/2}$ be as in Theorem 4. Then

$$\mathbb{P} \left(\mu(x) \in \left[\hat{\mu}_{\mathrm{d}}(x) - \sqrt{\frac{\lambda^d}{n}} \hat{\Sigma}_{\mathrm{d}}(x)^{1/2} q_{1-\alpha/2}, \ \hat{\mu}_{\mathrm{d}}(x) + \sqrt{\frac{\lambda^d}{n}} \hat{\Sigma}_{\mathrm{d}}(x)^{1/2} q_{1-\alpha/2} \right] \right) \to 1 - \alpha.$$

One important benefit of our debiasing technique is made clear in Theorem 8: 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 6 for more details. Nonetheless, such improvements do not come without concession. The limiting variance $\Sigma_{\mathrm{d}}(x)$ of the debiased estimator is larger than that of the unbiased version (the extent of this increase depends on the choice of the debiasing parameters a_r), leading to wider confidence intervals and larger estimation error in small samples despite the theoretical asymptotic improvements.

5.3 Minimax optimality

Our final result, Theorem 9 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 estimating a regression function $\mu \in \mathcal{H}^\beta$ in d dimensions (Stone, 1982). This result holds for all $d \geq 1$ and all $\beta > 0$, complementing a previous result established only for $\beta \in (0, 2]$ by Mourtada et al. (2020).

Theorem 9 (Minimax optimality of the debiased Mondrian random forest estimator)
Grant Assumptions 1 and 2. Fix $J \geq \lfloor \beta/2 \rfloor$, take $\lambda \asymp n^{\frac{1}{d+2\beta}}$, and suppose $B \gtrsim n^{\frac{2\beta-2}{d+2\beta}}$. Then

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

The sequence of lifetime parameters λ required in Theorem 9 are chosen to balance the bias and standard deviation bounds implied by Theorem 6 and Theorem 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 challenging and beyond the scope of this paper; we provide some practical advice for tuning parameter selection in Section 6.

Theorem 9 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 $\delta \in (0, 1/2)$ is fixed then under the same conditions as Theorem 9,

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

with appropriate tuning of λ and B.

5.4 Interpretation

The debiased Mondrian random forest estimator defined in (5) is a linear combination of 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 thus far been interpreting this estimator as a debiased version of the standard Mondrian random forest given in (1), but it is of course 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; Friedberg et al., 2020).

6 Tuning parameter selection

We discuss various 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 $0 \leq r \leq J$.

6.1 Selecting the base lifetime parameter λ

The most important parameter is the base Mondrian lifetime parameter 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 $\lambda \gg n^{\frac{1}{d+2\beta}}$. Secondly, the minimax optimality result of Theorem 9 is valid only in the regime $\lambda \asymp n^{\frac{1}{d+2\beta}}$, and thus requires careful determination in the more realistic finite-sample setting. For clarity, in this section we use the notation $\hat{\mu}_{\mathrm{d}}(x; \lambda, J)$ for the debiased Mondrian random forest with lifetime λ and debiasing order J. Similarly write $\hat{\Sigma}_{\mathrm{d}}(x; \lambda, J)$ for the associated variance estimator.

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 9. For an explicit example of how to choose the lifetime, one can instead use $\hat{\mu}_{\mathrm{d}}\big(x; \hat{\lambda}_{\mathrm{AIMSE}}(J-1), J-1\big)$ so that the leading bias is explicitly characterized by Theorem 6, and with $\hat{\lambda}_{\mathrm{AIMSE}}(J-1)$ as defined below. This is no longer minimax-optimal as $J - 1 < J$ does not satisfy the conditions of Theorem 9.

For performing inference, a more careful procedure is required; we suggest the following method assuming $\beta > 2$. Set $J = \lfloor \underline{\beta}/2 \rfloor$ as before, and use $\hat{\mu}_{\mathrm{d}}\big(x; \hat{\lambda}_{\mathrm{AIMSE}}(J-1), J\big)$ and $\hat{\Sigma}_{\mathrm{d}}\big(x; \hat{\lambda}_{\mathrm{AIMSE}}(J-1), J\big)$ to construct a confidence interval. The reasoning for this is that we select a lifetime tailored for a more biased estimator than we actually use. 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 valid 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 sampling variability introduced by the bias correction. This leads to a more refined distributional approximation but it does not necessarily exhaust the underlying smoothness of the regression function. An alternative inference approach based on Lepskii's method (Lepskii, 1992; Birgé, 2001) could be developed with the latter goal in mind.

It remains to propose a concrete method for computing $\hat{\lambda}_{\mathrm{AIMSE}}(J)$ in the finite-sample setting; we suggest two such procedures based on plug-in selection with local polynomials and cross-validation respectively, building on classical ideas from the nonparametric smoothing literature (Fan et al., 2020).

Lifetime selection with local polynomial smoothing

Firstly suppose $X_i \sim \mathrm{Unif}\big([0,1]^d\big)$ 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 $\partial_j^{2J+2}\mu(x)/\partial x_j^{2J+2}$,

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

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

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

the asymptotic integrated mean squared error (AIMSE) is

$$\begin{aligned} \mathrm{AIMSE}(\lambda, J) &= \frac{1}{\lambda^{4J+4}} \frac{\bar{\omega}^2}{(J+2)^2} \int_{[0,1]^d} \left(\sum_{j=1}^{d} \partial_j^{2J+2} \mu(x) \right)^2 \mathrm{d}x \\ &\quad + \frac{\lambda^d \sigma^2}{n} \sum_{r=0}^{J} \sum_{r'=0}^{J} \omega_r \omega_{r'} \left(\ell_{rr'} + \ell_{r'r}\right)^d . \end{aligned}$$

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

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

An estimator of $\lambda_{\mathrm{AIMSE}}(J)$ is therefore given by

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

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 of order $2J+4$ without interaction terms. To do this, define the $n \times ((2J+4)d+1)$ design matrix P with rows

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

and set

$$P_x = \left(1, x_1, x_1^2, \ldots, x_1^{2J+4}, x_2, x_2^2, \ldots, x_2^{2J+4}, \ldots, x_d, x_d^2, \ldots, x_d^{2J+4}\right).$$

Then the derivative estimator is

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

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

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

Lifetime selection with cross-validation

As an alternative to the analytic plug-in methods described above, one can use a cross-validation approach. While leave-one-out cross-validation (LOOCV) can be applied directly, the linear smoother structure of the (debiased) Mondrian random forest estimator allows a computationally simpler formulation. Writing $\hat{\mu}_{\mathrm{d}}^{-i}(x)$ for a debiased Mondrian random forest estimator fitted without using the ith data sample, it is easy to show that

$$\begin{aligned}
\mathrm{LOOCV}(\lambda, J) &= \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{\mu}_{\mathrm{d}}^{-i}(X_i)\right)^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{r=0}^{J} \omega_r \frac{1}{B} \sum_{b=1}^{B} \frac{1}{1 - 1/N_{rb}(X_i)} \left(Y_i - \sum_{j=1}^{n} \frac{Y_j \mathbb{I}\left\{X_j \in T_{rb}(X_i)\right\}}{N_{rb}(X_i)} \right) \right)^2,
\end{aligned}$$

avoiding refitting the model leaving each sample out in turn. Supposing $X_i \sim \mathrm{Unif}\left([0,1]^d\right)$ and replacing $1/N_{rb}(X_i)$ with their average expectation $\frac{1}{J+1} \sum_{r=0}^{J} \mathbb{E}\left[1/N_{rb}(X_i)\right] \approx \bar{a}^d \lambda^d / n$ where $\bar{a}^d = \frac{1}{J+1} \sum_{r=0}^{J} a_r^d$ gives the generalized cross-validation (GCV) formula

$$\mathrm{GCV}(\lambda, J) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i - \hat{\mu}_{\mathrm{d}}(X_i)}{1 - \bar{a}^d \lambda^d / n} \right)^2 .$$

The lifetime can then be selected by computing $\hat{\lambda}_{\mathrm{LOOCV}} \in \mathrm{argmin}_\lambda \, \mathrm{LOOCV}(\lambda, J)$ or $\hat{\lambda}_{\mathrm{GCV}} \in \mathrm{argmin}_\lambda \, \mathrm{GCV}(\lambda, J)$.

6.2 Choosing the number B of trees in each forest

The next parameter to choose is the number of trees in each forest, B. If no debiasing is applied, we suggest taking $B = \sqrt{n}$ to satisfy the constraint in Theorem 4. If debiasing is used then we recommend setting $B = n^{\frac{2J-1}{2J}}$, consistent with Theorem 8 and Theorem 9.

6.3 Setting the debiasing order J

When constructing a debiased Mondrian random forest estimator, we must decide how many orders of bias to remove. Of course this requires having some form of oracle 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 overly inflating the variance of the estimator.

6.4 Selecting the debiasing coefficients

As mentioned in Section 5, we take a_r to be a fixed geometric or arithmetic sequence. For example, we could set $a_r = (1+\gamma)^r$ or $a_r = 1 + \gamma r$ for some $\gamma > 0$. We suggest for concreteness taking $a_r = 1.05^r$.

7 Conclusion

We presented a central limit theorem 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 a debiased version of Mondrian random forests, exploiting higher order smoothness, and demonstrated their advantages for statistical inference and their minimax optimality properties. Finally we discussed tuning parameter selection, enabling fully feasible and practical estimation and inference procedures.

8 Acknowledgments

The authors would like to thank Krishnakumar Balasubramanian, Rajita Chandak, Jianqing Fan, and Sanjeev Kulkarni for comments and insightful discussions.

9 Funding

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. See Section 4 in the main paper for an overview of the main proof strategies and a discussion of the challenges involved. We use the following simplified notation for convenience, whenever it is appropriate. We 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 $\mathbb{I}_b(x) = \mathbb{I}\{N_b(x) \geq 1\}$.

A.1 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 1 (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 1)
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 2 (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 sufficiently large* n, *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 2)
Note that $N_\cup(x) \sim \operatorname{Bin}\left(n, \int_{\bigcup_{b=1}^B T_b(x)} f(s)\,\mathrm{d}s\right) \leq \operatorname{Bin}\left(n, 2^d \max_{1\leq b\leq B}\max_{1\leq j\leq d}|T_b(x)_j|^d\|f\|_\infty\right)$ conditionally on $\mathbf{T}$. If $N \sim \operatorname{Bin}(n,p)$ 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 $t \geq 2$,

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

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

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

Replacing t by $t/2$ gives that for sufficiently large n such that $n/\lambda^d \geq 1$,

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

□

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 $N_b(x) \sim \operatorname{Bin}\left(n, \int_{T_b(x)} f(s)\,\mathrm{d}s\right)$. Our constants could be significantly suboptimal but they are sufficient for our applications.

Lemma 3 (An inverse moment bound for the binomial distribution)
For $n \geq 1$ *and* $p \in [0,1]$, *let* $N \sim \operatorname{Bin}(n,p)$ *and* $a_1, \ldots, a_k \geq 0$. *Then*

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

Proof (Lemma 3)
By Bernstein's inequality, $\mathbb{P}(N \leq np - t) \leq \exp\left(-\frac{t^2/2}{np(1-p)+t/3}\right) \leq \exp\left(-\frac{3t^2}{6np+2t}\right)$. 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

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

□

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 3 repeatedly to each subcell in $\mathcal{R}$ in turn, replacing the number of samples X_i in that subcell with its volume multiplied by 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 4 (Generalized moment bound for Mondrian random forest denominators)
Suppose Assumptions 1 and 2 hold. Let $T_b \sim \mathcal{M}\big([0,1]^d, \lambda\big)$ 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,

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

Proof (Lemma 4)
Define the common refinement of $\{T_b(x) : 1 \leq b \leq B_0\}$ as the class of sets

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

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 $D \in \mathcal{R} \setminus \mathcal{D}$,

$$
\begin{aligned}
&\mathbb{E}\left[\prod_{b=1}^{B_0} \frac{1}{1 \vee \left(\sum_{D' \in \mathcal{A}(b) \setminus \mathcal{D}} N_b(D') + n \sum_{D' \in \mathcal{A}(b) \cap \mathcal{D}} |D'|\right)^{k_b}}\right] \\
&= \mathbb{E}\left[\prod_{b \notin \mathcal{A}(D)} \frac{1}{1 \vee \left(\sum_{D' \in \mathcal{A}(b) \setminus \mathcal{D}} N_b(D') + n \sum_{D' \in \mathcal{A}(b) \cap \mathcal{D}} |D'|\right)^{k_b}}\right. \\
&\quad \times \left.\mathbb{E}\left[\prod_{b \in \mathcal{A}(D)} \frac{1}{1 \vee \left(\sum_{D' \in \mathcal{A}(b) \setminus \mathcal{D}} N_b(D') + n \sum_{D' \in \mathcal{A}(b) \cap \mathcal{D}} |D'|\right)^{k_b}} \,\middle|\, \mathbf{T}, N_b(D') : D' \in \mathcal{R} \backslash (\mathcal{D} \cup \{D\})\right]\right].
\end{aligned}
$$

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, for sufficiently large t by Lemma 2

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

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 3,

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

Therefore plugging this back into the marginal expectation yields

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

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 $|\mathcal{R}| \leq 2^{B_0}$ steps and yields

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

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,

$$\prod_{b=1}^{B_0} \mathbb{E}\left[\frac{1}{1 \vee (n|T_b(x)|)^{k_b}}\right] \geq \prod_{b=1}^{B_0} \frac{1}{\mathbb{E}\left[1 \vee (n|T_b(x)|)^{k_b}\right]} \geq \prod_{b=1}^{B_0} \frac{1}{n^{k_b}} = n^{-k} \geq 4dB_0 2^{B_0} e^{\frac{-\sqrt{\lambda}}{8\|f\|_\infty}}$$

for sufficiently large n because B_0, d, and k are fixed while $\log \lambda \gtrsim \log n$. □

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 $\log n$ 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 $k \geq 3$. This "small cell" phenomenon in which the inverse volumes of Mondrian cells have heavy tails is a recurring challenge in our analysis.

Lemma 5 (Inverse moments of the volume of a Mondrian cell)
Suppose Assumption 2 holds and let $T \sim \mathcal{M}([0,1]^d, \lambda)$. Then for sufficiently large n,

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

Proof (Lemma 5)
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 $\mathrm{Exp}(1)$ random variables. Thus for any $0 < t < 1$, using the fact that $E_{j1} + E_{j2} \sim \mathrm{Gamma}(2, 1)$,

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

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

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

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

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

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

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

□

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 6 (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 $k = \sum_{b=1}^{B_0} k_b$,*

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

for sufficiently large n.

Proof (Lemma 6)
This follows directly from Lemmas 4 and 5. □

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 3 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 3.

Lemma 7 (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 7)
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 3 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 3 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.2 Proofs for Section 3

We give rigorous proofs of the central limit theorem, bias characterization, and variance estimation results for the Mondrian random forest estimator without debiasing.

Proof of central limit theorem

Proof (Theorem 1)
Follows from the debiased version. See the proof of Theorem 5 and set $J = 0$, $a_0 = 1$, and $\omega_0 = 1$. □

Proof of bias characterization

See Section 4 in the main paper for details on our approach to this proof.

Proof (Theorem 2)
Part 1: Removing the dependence on the trees
By measurability and with $\mu(X_i) = \mathbb{E}[Y_i \mid X_i]$ almost surely,

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

Now conditional on $\mathbf{X}$, the terms in the outer sum depend only on T_b so are i.i.d. Since μ is Lipschitz,

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

where we used the law of $T(x)_j$ from Mourtada et al. (2020, Proposition 1). So by Chebyshev's inequality,

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

Part 2: Showing the conditional bias converges in probability

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 $\tilde{\mathbb{I}}_{ijb}(x) = \mathbb{I}\{\tilde{X}_{ij} \in T_b(x)\}$ and $N_{jb}(x) = \sum_{i=1}^n \tilde{\mathbb{I}}_{ijb}(x)$.

$$\begin{aligned}
&\operatorname{Var}\left[\sum_{i=1}^n \left(\mu(X_i) - \mu(x)\right)\mathbb{E}\left[\frac{\mathbb{I}_{ib}(x)}{N_b(x)} \;\Big|\; \mathbf{X}\right]\right] \\
&\leq \frac{1}{2}\sum_{j=1}^n \mathbb{E}\left[\left(\sum_{i=1}^n \left(\mu(X_i) - \mu(x)\right)\mathbb{E}\left[\frac{\mathbb{I}_{ib}(x)}{N_b(x)} \;\Big|\; \mathbf{X}\right] - \sum_{i=1}^n \left(\mu(\tilde{X}_{ij}) - \mu(x)\right)\mathbb{E}\left[\frac{\tilde{\mathbb{I}}_{ijb}(x)}{\tilde{N}_{jb}(x)} \;\Big|\; \tilde{\mathbf{X}}_j\right]\right)^2\right] \\
&\leq \frac{1}{2}\sum_{j=1}^n \mathbb{E}\left[\left(\sum_{i=1}^n \left(\left(\mu(X_i) - \mu(x)\right)\frac{\mathbb{I}_{ib}(x)}{N_b(x)} - \left(\mu(\tilde{X}_{ij}) - \mu(x)\right)\frac{\tilde{\mathbb{I}}_{ijb}(x)}{\tilde{N}_{jb}(x)}\right)\right)^2\right] \\
&\leq \sum_{j=1}^n \mathbb{E}\left[\left(\sum_{i\neq j} \left(\mu(X_i) - \mu(x)\right)\left(\frac{\mathbb{I}_{ib}(x)}{N_b(x)} - \frac{\mathbb{I}_{ib}(x)}{\tilde{N}_{jb}(x)}\right)\right)^2\right] + 2\sum_{j=1}^n \mathbb{E}\left[\left(\mu(X_j) - \mu(x)\right)^2 \frac{\mathbb{I}_{jb}(x)}{N_b(x)^2}\right].
\end{aligned} \tag{6}$$

For the first term in (6) 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 μ is Lipschitz, writing $\mathbb{I}_{-jb}(x) = \mathbb{I}\left\{N_{-jb}(x) \geq 1\right\}$,

$$\begin{aligned}
&\sum_{j=1}^n \mathbb{E}\left[\left(\sum_{i\neq j} \left(\mu(X_i) - \mu(x)\right)\left(\frac{\mathbb{I}_{ib}(x)}{N_b(x)} - \frac{\mathbb{I}_{ib}(x)}{\tilde{N}_{jb}(x)}\right)\right)^2\right] \\
&\quad\lesssim \sum_{j=1}^n \mathbb{E}\left[\max_{1\leq l\leq d}|T_b(x)_l|^2 \left(\frac{\sum_{i\neq j}\mathbb{I}_{ib}(x)\mathbb{I}_{jb}(x)}{N_{-jb}(x)(N_{-jb}(x)+1)}\right)^2\right] \lesssim \mathbb{E}\left[\max_{1\leq l\leq d}|T_b(x)_l|^2 \frac{\mathbb{I}_b(x)}{N_b(x)}\right].
\end{aligned}$$

For $t > 0$, partition by the event $\{\max_{1 \leq l \leq d} |T_b(x)_l| \geq t/\lambda\}$ and apply Lemma 1 and Lemma 6:

$$\mathbb{E}\left[\max_{1\leq l\leq d} |T_b(x)_l|^2 \frac{\mathbb{I}_b(x)}{N_b(x)}\right] \leq \mathbb{P}\left(\max_{1\leq l\leq d} |T_b(x)_l| \geq t/\lambda\right) + (t/\lambda)^2\, \mathbb{E}\left[\frac{\mathbb{I}_b(x)}{N_b(x)}\right]$$
$$\lesssim e^{-t/2} + \left(\frac{t}{\lambda}\right)^2 \frac{\lambda^d}{n} \lesssim \frac{1}{n^2} + \frac{(\log n)^2}{\lambda^2}\frac{\lambda^d}{n} \lesssim \frac{(\log n)^2}{\lambda^2}\frac{\lambda^d}{n},$$

where we set $t = 4\log n$. For the second term in (6) we have

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

in the same manner. Hence

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

and so by Chebyshev's inequality,

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

Part 3: Computing the limiting bias

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

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

By Lemma 7, since $N_{-ib}(x) \sim \mathrm{Bin}\left(n-1, \int_{T_b(x)} f(s)\, \mathrm{d}s\right)$ given $\mathbf{T}$ and f is bounded away from zero,

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

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 1 and Lemma 5, since f is Lipschitz and bounded, using the Cauchy–Schwarz inequality,

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

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

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

Note that since f and μ are Lipschitz and by integrating the tail probability given in Lemma 1, the Maclaurin remainder term is bounded by

$$\begin{aligned}
&\mathbb{E}\left[\frac{\int_{T_b(x)} |\mu(s) - \mu(x)|\, f(s)\, \mathrm{d}s}{f(x) |T_b(x)|} |A|^{\underline{\beta}+1}\right] \\
&\quad= \mathbb{E}\left[\frac{\int_{T_b(x)} |\mu(s) - \mu(x)|\, f(s)\, \mathrm{d}s}{f(x) |T_b(x)|} \left(\frac{1}{f(x) |T_b(x)|} \int_{T_b(x)} (f(s) - f(x))\, \mathrm{d}s\right)^{\underline{\beta}+1}\right] \\
&\quad\lesssim \mathbb{E}\left[\max_{1 \le l \le d} |T_b(x)_l|^{\underline{\beta}+2}\right] = \int_0^\infty \mathbb{P}\left(\max_{1 \le l \le d} |T_b(x)_l| \ge t^{\frac{1}{\underline{\beta}+2}}\right) \mathrm{d}t \le \int_0^\infty 2d e^{-\lambda t^{\frac{1}{\underline{\beta}+2}}/2}\, \mathrm{d}t \\
&\quad= \frac{2^{\underline{\beta}+3} d (\underline{\beta} + 2)!}{\lambda^{\underline{\beta}+2}} \lesssim \frac{1}{\lambda^\beta}
\end{aligned}$$

since $\int_0^\infty e^{-a x^{1/k}}\, \mathrm{d}x = a^{-k} k!$. Hence to summarize the progress so far, we have

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

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

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

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

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

where $\binom{k}{u}$ is a multinomial coefficient. By Taylor's theorem with $f, \mu \in \mathcal{H}^\beta$,

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

Now by integrating the tail probabilities in Lemma 1, $\mathbb{E}\left[\max_{1 \leq l \leq d} |T_b(x)_l|^\beta\right] \lesssim \frac{1}{\lambda^\beta}$. Therefore by Lemma 5, writing $T_b(x)^\nu$ for $\int_{T_b(x)} (s-x)^\nu \, \mathrm{d}s$,

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

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,

$$\begin{aligned}
&\mathbb{E}\left[\frac{T_b(x)^{\nu'+\nu''} (T_b(x)^\nu)^u}{|T_b(x)|^{|u|+1}} \right] \\
&\quad = \lambda^{-|\nu'|-|\nu''|-|\nu| \cdot u} \prod_{j=1}^d \mathbb{E}\left[\frac{E_{2j*}^{\nu'_j+\nu''_j+1} + (-1)^{\nu'_j+\nu''_j} E_{1j*}^{\nu'_j+\nu''_j+1}}{(\nu'_j + \nu''_j + 1)(E_{2j*} + E_{1j*})} \frac{\left(E_{2j*}^{\nu_j+1} + (-1)^{\nu_j} E_{1j*}^{\nu_j+1} \right)^u}{(\nu_j + 1)^u (E_{2j*} + E_{1j*})^{|u|}} \right]. \qquad (7)
\end{aligned}$$

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,

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

Further, the quantity inside the expectation in (7) 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 (7) is at most $2de^{-C\lambda} \lesssim \lambda^{-\beta}$. Thus the limiting bias is

$$\mathbb{E}\left[\hat{\mu}(x)\right] - \mu(x) = \sum_{|\nu'|=1}^{\beta} \sum_{|\nu''|=0}^{\beta} \sum_{|u|=0}^{\beta} \frac{\partial^{\nu'}\mu(x)}{\nu'!} \frac{\partial^{\nu''}f(x)}{\nu''!} \left(\frac{\partial^{\nu}f(x)}{\nu!}\right)^{u} \binom{|u|}{u} \frac{(-1)^{|u|}}{f(x)^{|u|+1}} \lambda^{-|\nu'|-|\nu''|-|\nu|\cdot u} \tag{8}$$

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

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 $\underline{\beta}$ of f and μ 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 $\sum_{|\nu|=1}^{\beta} \nu_j u_\nu$ is odd.

If $\nu'_j + \nu''_j$ is odd, then $\sum_{|\nu|=1}^{\beta} \nu_j u_\nu$ 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

$$\frac{E_{2j}^{\nu'_j+\nu''_j+1} + (-1)^{\nu'_j+\nu''_j} E_{1j}^{\nu'_j+\nu''_j+1}}{E_{2j}+E_{1j}} \tag{9}$$

and

$$\frac{\left(E_{2j}^{\nu_j+1} + (-1)^{\nu_j} E_{1j}^{\nu_j+1}\right)^{u}}{(E_{2j}+E_{1j})^{|u|}} = \prod_{\substack{|\nu|=1 \\ \nu_j u_\nu \text{ even}}}^{\beta} \frac{\left(E_{2j}^{\nu_j+1} + (-1)^{\nu_j} E_{1j}^{\nu_j+1}\right)^{u_\nu}}{(E_{2j}+E_{1j})^{u_\nu}} \prod_{\substack{|\nu|=1 \\ \nu_j u_\nu \text{ odd}}}^{\beta} \frac{\left(E_{2j}^{\nu_j+1} + (-1)^{\nu_j} E_{1j}^{\nu_j+1}\right)^{u_\nu}}{(E_{2j}+E_{1j})^{u_\nu}}. \tag{10}$$

Clearly $\nu'_j + \nu''_j$ being odd inverts the sign of (9). For (10), 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 (10) 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 (9) and (10) is zero by symmetry.

If however $\nu'_j + \nu''_j$ is even, then $\sum_{|\nu|=1}^{\beta} \nu_j u_\nu$ is odd so $|\{\nu : \nu_j u_\nu \text{ is odd}\}|$ is odd. Clearly the sign of (9) is preserved. Again the sign of the first product in (10) is preserved, and the sign of every term in (10) 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 (9) and (10) 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 (8) 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 (8) 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 $X_i \sim \text{Unif}\left([0,1]^d\right)$ then $f(x) = 1$ and the bias expansion from (8) 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 of variance estimator consistency

Proof (Theorem 3)
Follows from the debiased version. See the proof of Theorem 7 and set $J = 0$, $a_0 = 1$, and $\omega_0 = 1$. □

Proof of confidence interval validity

Proof (Theorem 4)
By Theorem 2 and Theorem 3,

$$\sqrt{\frac{n}{\lambda^d}} \frac{\hat{\mu}(x) - \mu(x)}{\hat{\Sigma}(x)^{1/2}} = \sqrt{\frac{n}{\lambda^d}} \frac{\hat{\mu}(x) - \mathbb{E}\left[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\right]}{\hat{\Sigma}(x)^{1/2}} + \sqrt{\frac{n}{\lambda^d}} \frac{\mathbb{E}\left[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\right] - \mu(x)}{\hat{\Sigma}(x)^{1/2}}$$
$$= \sqrt{\frac{n}{\lambda^d}} \frac{\hat{\mu}(x) - \mathbb{E}\left[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\right]}{\hat{\Sigma}(x)^{1/2}} + \sqrt{\frac{n}{\lambda^d}} O_{\mathbb{P}}\left(\frac{1}{\lambda^{\beta \wedge 2}} + \frac{1}{\lambda \sqrt{B}} + \frac{\log n}{\lambda} \sqrt{\frac{\lambda^d}{n}}\right).$$

The first term now converges weakly to $\mathcal{N}(0,1)$ by Slutsky's theorem, Theorem 1, and Theorem 3, while the second term is $o_{\mathbb{P}}(1)$ by assumption. Validity of the confidence interval follows immediately. □

A.3 Proofs for Section 5

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

Proof of central limit theorem with debiasing

Proof (Theorem 5)
We use the martingale central limit theorem given by Hall and Heyde (2014, Theorem 3.2). For each $1 \leq i \leq n$ define $\mathcal{H}_{ni}$ to be the filtration generated by $\mathbf{T}$, $\mathbf{X}$ and $(\varepsilon_j : 1 \leq j \leq i)$, noting that $\mathcal{H}_{ni} \subseteq \mathcal{H}_{(n+1)i}$ because B increases weakly as n increases. Let $\mathbb{I}_{ibr}(x) = \mathbb{I}\{X_i \in T_{br}(x)\}$ where $T_{br}(x)$ is the cell containing x in tree b used to construct $\hat{\mu}_r(x)$, and similarly let $N_{br}(x) = \sum_{i=1}^n \mathbb{I}_{ibr}(x)$ and $\mathbb{I}_{br}(x) = \mathbb{I}\{N_{br}(x) \geq 1\}$. Define the $\mathcal{H}_{ni}$-measurable and square integrable variables

$$S_i(x) = \sqrt{\frac{n}{\lambda^d}} \sum_{r=0}^{J} \omega_r \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{I}_{ibr}(x) \varepsilon_i}{N_{br}(x)},$$

which satisfy the martingale difference property $\mathbb{E}[S_i(x) \mid \mathcal{H}_{ni}] = 0$. Further,

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

By Hall and Heyde (2014, Theorem 3.2) it suffices to check that

(i) $\max_i |S_i(x)| \to 0$ in probability,

(ii) $\mathbb{E}\left[\max_i S_i(x)^2\right] \lesssim 1$,

(iii) $\sum_i S_i(x)^2 \to \Sigma_{\mathrm{d}}(x)$ in probability.

Part 1: checking condition (i)
Since J is fixed and $\mathbb{E}[|\varepsilon_i|^3 \mid X_i]$ is bounded, by Jensen's inequality and Lemma 6,

$$\begin{aligned}
\mathbb{E}\left[\max_{1\le i\le n}|S_i(x)|\right] &= \mathbb{E}\left[\max_{1\le i\le n}\left|\sqrt{\frac{n}{\lambda^d}}\sum_{r=0}^{J}\omega_r\frac{1}{B}\sum_{b=1}^{B}\frac{\mathbb{I}_{ibr}(x)\varepsilon_i}{N_{br}(x)}\right|\right] \\
&\le \sqrt{\frac{n}{\lambda^d}}\sum_{r=0}^{J}|\omega_r|\frac{1}{B}\mathbb{E}\left[\max_{1\le i\le n}\left|\sum_{b=1}^{B}\frac{\mathbb{I}_{ibr}(x)\varepsilon_i}{N_{br}(x)}\right|\right] \\
&\le \sqrt{\frac{n}{\lambda^d}}\sum_{r=0}^{J}|\omega_r|\frac{1}{B}\mathbb{E}\left[\sum_{i=1}^{n}\left(\sum_{b=1}^{B}\frac{\mathbb{I}_{ibr}(x)|\varepsilon_i|}{N_{br}(x)}\right)^3\right]^{1/3} \\
&= \sqrt{\frac{n}{\lambda^d}}\sum_{r=0}^{J}|\omega_r|\frac{1}{B}\mathbb{E}\left[\sum_{i=1}^{n}|\varepsilon_i|^3\sum_{b=1}^{B}\sum_{b'=1}^{B}\sum_{b''=1}^{B}\frac{\mathbb{I}_{ibr}(x)}{N_{br}(x)}\frac{\mathbb{I}_{ib'r}(x)}{N_{b'r}(x)}\frac{\mathbb{I}_{ib''r}(x)}{N_{b''r}(x)}\right]^{1/3} \\
&\lesssim \sqrt{\frac{n}{\lambda^d}}\sum_{r=0}^{J}|\omega_r|\frac{1}{B^{2/3}}\mathbb{E}\left[\sum_{b=1}^{B}\sum_{b'=1}^{B}\frac{\mathbb{I}_{br}(x)}{N_{br}(x)}\frac{\mathbb{I}_{b'r}(x)}{N_{b'r}(x)}\right]^{1/3} \\
&\lesssim \sqrt{\frac{n}{\lambda^d}}\sum_{r=0}^{J}|\omega_r|\frac{1}{B^{2/3}}\left(B^2\frac{a_r^{2d}\lambda^{2d}}{n^2}+B\frac{a_r^{2d}\lambda^{2d}\log n}{n^2}\right)^{1/3} \\
&\lesssim \left(\frac{\lambda^d}{n}\right)^{1/6}+\left(\frac{\lambda^d}{n}\right)^{1/6}\left(\frac{\log n}{B}\right)^{1/3}\to 0.
\end{aligned}$$

Part 2: checking condition (ii)
Since $\mathbb{E}[\varepsilon_i^2 \mid X_i]$ is bounded and by Lemma 6,

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

Part 3: checking condition (iii)
Next, we have

$$\begin{aligned}
\sum_{i=1}^{n}S_i(x)^2 &= \sum_{i=1}^{n}\left(\sqrt{\frac{n}{\lambda^d}}\sum_{r=0}^{J}\omega_r\frac{1}{B}\sum_{b=1}^{B}\frac{\mathbb{I}_{ibr}(x)\varepsilon_i}{N_{br}(x)}\right)^2 \\
&= \frac{n}{\lambda^d}\frac{1}{B^2}\sum_{i=1}^{n}\sum_{r=0}^{J}\sum_{r'=0}^{J}\omega_r\omega_{r'}\sum_{b=1}^{B}\sum_{b'=1}^{B}\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)} \\
&= \frac{n}{\lambda^d}\frac{1}{B^2}\sum_{i=1}^{n}\sum_{r=0}^{J}\sum_{r'=0}^{J}\omega_r\omega_{r'}\sum_{b=1}^{B}\left(\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ibr'}(x)\varepsilon_i^2}{N_{br}(x)N_{br'}(x)}+\sum_{b'\ne b}\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)}\right). \qquad (11)
\end{aligned}$$

By boundedness of $\mathbb{E}[\varepsilon_i^2 \mid X_i]$ and Lemma 6, the first term in (11) converges to zero in probability as

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

For the second term in (11), the law of total variance gives

$$\begin{aligned}
&\operatorname{Var}\left[\frac{n}{\lambda^d}\frac{1}{B^2}\sum_{i=1}^{n}\sum_{r=0}^{J}\sum_{r'=0}^{J}\omega_r\omega_{r'}\sum_{b=1}^{B}\sum_{b'\ne b}\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)}\right] \\
&\le (J+1)^4\max_{0\le r,r'\le J}\omega_r\omega_{r'}\operatorname{Var}\left[\frac{n}{\lambda^d}\frac{1}{B^2}\sum_{i=1}^{n}\sum_{b=1}^{B}\sum_{b'\ne b}\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)}\right] \\
&\lesssim \max_{0\le r,r'\le J}\mathbb{E}\left[\operatorname{Var}\left[\frac{n}{\lambda^d}\frac{1}{B^2}\sum_{i=1}^{n}\sum_{b=1}^{B}\sum_{b'\ne b}\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)}\;\Big|\;\mathbf{X},\mathbf{Y}\right]\right] \\
&\quad+\max_{0\le r,r'\le J}\operatorname{Var}\left[\mathbb{E}\left[\frac{n}{\lambda^d}\frac{1}{B^2}\sum_{i=1}^{n}\sum_{b=1}^{B}\sum_{b'\ne b}\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)}\;\Big|\;\mathbf{X},\mathbf{Y}\right]\right] \qquad (12)
\end{aligned}$$

For the first term in (12),

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{Var}\left[\frac{n}{\lambda^d}\frac{1}{B^2}\sum_{i=1}^{n}\sum_{b=1}^{B}\sum_{b'\ne b}\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)}\;\Big|\;\mathbf{X},\mathbf{Y}\right]\right] = \frac{n^2}{\lambda^{2d}}\frac{1}{B^4}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{b=1}^{B}\sum_{b'\ne b}\sum_{\tilde b=1}^{B}\sum_{\tilde b'\ne \tilde b} \\
&\mathbb{E}\left[\varepsilon_i^2\varepsilon_j^2\left(\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)}{N_{br}(x)N_{b'r'}(x)}-\mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)}{N_{br}(x)N_{b'r'}(x)}\;\Big|\;\mathbf{X}\right]\right)\left(\frac{\mathbb{I}_{j\tilde br}(x)\mathbb{I}_{j\tilde b'r'}(x)}{N_{\tilde br}(x)N_{\tilde b'r'}(x)}-\mathbb{E}\left[\frac{\mathbb{I}_{j\tilde br}(x)\mathbb{I}_{j\tilde b'r'}(x)}{N_{\tilde br}(x)N_{\tilde b'r'}(x)}\;\Big|\;\mathbf{X}\right]\right)\right].
\end{aligned}$$

Since T_{br} is independent of $T_{b'r'}$ given $\mathbf{X},\mathbf{Y}$, the summands are zero whenever $|\{b,b',\tilde b,\tilde b'\}|=4$. Further, since $\mathbb{E}[\varepsilon_i^2 \mid X_i]$ is bounded and by the Cauchy–Schwarz inequality and Lemma 6,

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{Var}\left[\frac{n}{\lambda^d}\frac{1}{B^2}\sum_{i=1}^{n}\sum_{b=1}^{B}\sum_{b'\ne b}\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)}\;\Big|\;\mathbf{X},\mathbf{Y}\right]\right] \\
&\quad\lesssim \frac{n^2}{\lambda^{2d}}\frac{1}{B^3}\sum_{b=1}^{B}\sum_{b'\ne b}\mathbb{E}\left[\left(\sum_{i=1}^{n}\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)}{N_{br}(x)N_{b'r'}(x)}\right)^2\right] \lesssim \frac{n^2}{\lambda^{2d}}\frac{1}{B}\mathbb{E}\left[\frac{\mathbb{I}_{br}(x)}{N_{br}(x)}\frac{\mathbb{I}_{b'r'}(x)}{N_{b'r'}(x)}\right] \lesssim \frac{1}{B}\to 0.
\end{aligned}$$

For the second term in (12), the random variable inside the variance is a nonlinear function of the i.i.d. variables (X_i,ε_i), so we apply the Efron–Stein inequality (Efron and Stein, 1981). Let $(\tilde X_{ij},\tilde Y_{ij})=(X_i,Y_i)$ if $i\ne j$ and be an independent copy of (X_j,Y_j), denoted $(\tilde X_j,\tilde Y_j)$, if $i=j$, and define $\tilde\varepsilon_{ij}=\tilde Y_{ij}-\mu(\tilde X_{ij})$. Write $\tilde{\mathbb{I}}_{ijbr}(x)=\mathbb{I}\{\tilde X_{ij}\in T_{br}(x)\}$ and $\tilde{\mathbb{I}}_{jbr}(x)=\mathbb{I}\{\tilde X_j\in T_{br}(x)\}$, and also $\tilde N_{jbr}(x)=\sum_{i=1}^{n}\tilde{\mathbb{I}}_{ijbr}(x)$. We use the leave-one-out notation $N_{-jbr}(x)=\sum_{i\ne j}\mathbb{I}_{ibr}(x)$ and also write

$N_{-jbr\cap b'r'} = \sum_{i\neq j} \mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)$. Since $\mathbb{E}[\varepsilon_i^4 \mid X_i]$ is bounded,

$$
\begin{aligned}
&\mathrm{Var}\left[\mathbb{E}\left[\frac{n}{\lambda^d}\frac{1}{B^2}\sum_{i=1}^n\sum_{b=1}^B\sum_{b'\neq b}\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)} \;\Big|\; \mathbf{X},\mathbf{Y}\right]\right] \leq \mathrm{Var}\left[\mathbb{E}\left[\frac{n}{\lambda^d}\sum_{i=1}^n\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)} \;\Big|\; \mathbf{X},\mathbf{Y}\right]\right] \\
&\quad\leq \frac{1}{2}\frac{n^2}{\lambda^{2d}}\sum_{j=1}^n \mathbb{E}\left[\left(\sum_{i=1}^n\left(\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)} - \frac{\tilde{\mathbb{I}}_{ijbr}(x)\tilde{\mathbb{I}}_{ijb'r'}(x)\tilde{\varepsilon}_{ij}^2}{\tilde{N}_{jbr}(x)\tilde{N}_{jb'r'}(x)}\right)\right)^2\right] \\
&\quad\leq \frac{n^2}{\lambda^{2d}}\sum_{j=1}^n \mathbb{E}\left[\left(\left|\frac{1}{N_b(x)N_{b'r'}(x)} - \frac{1}{\tilde{N}_{jbr}(x)\tilde{N}_{jb'r'}(x)}\right|\sum_{i\neq j}\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2\right)^2\right] \\
&\qquad+ \frac{n^2}{\lambda^{2d}}\sum_{j=1}^n \mathbb{E}\left[\left(\left(\frac{\mathbb{I}_{jbr}(x)\mathbb{I}_{jb'r'}(x)\varepsilon_j^2}{N_{br}(x)N_{b'r'}(x)} - \frac{\tilde{\mathbb{I}}_{jbr}(x)\tilde{\mathbb{I}}_{jb'r'}(x)\tilde{\varepsilon}_j^2}{\tilde{N}_{jbr}(x)\tilde{N}_{jb'r'}(x)}\right)\right)^2\right] \\
&\quad\lesssim \frac{n^2}{\lambda^{2d}}\sum_{j=1}^n \mathbb{E}\left[N_{-jbr\cap b'r}(x)^2\left|\frac{1}{N_{br}(x)N_{b'r'}(x)} - \frac{1}{\tilde{N}_{jbr}(x)\tilde{N}_{jb'r'}(x)}\right|^2 + \frac{\mathbb{I}_{jbr}(x)\mathbb{I}_{jb'r'}(x)}{N_{br}(x)^2N_{b'r'}(x)^2}\right].
\end{aligned}
\tag{13}
$$

For the first term in (13), note that

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

since $|N_{br}(x) - \tilde{N}_{jbr}(x)| \leq 1$ and $|N_{b'r'}(x) - \tilde{N}_{jb'r'}(x)| \leq 1$. Further, these terms are non-zero only on the events $\{X_j \in T_{br}(x)\} \cup \{\tilde{X}_j \in T_{br}(x)\}$ and $\{X_j \in T_{b'r'}(x)\} \cup \{\tilde{X}_j \in T_{b'r'}(x)\}$ respectively, so

$$
\begin{aligned}
&\mathrm{Var}\left[\mathbb{E}\left[\frac{n}{\lambda^d}\frac{1}{B^2}\sum_{i=1}^n\sum_{b=1}^B\sum_{b'\neq b}\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)} \;\Big|\; \mathbf{X},\mathbf{Y}\right]\right] \\
&\lesssim \frac{n^2}{\lambda^{2d}}\sum_{j=1}^n \mathbb{E}\left[\frac{\mathbb{I}_{jb'r'}(x) + \tilde{\mathbb{I}}_{jb'r'}(x)}{N_{-jbr}(x)^2}\frac{N_{-jbr\cap b'r}(x)^2}{N_{-jb'r'}(x)^4} + \frac{\mathbb{I}_{jbr}(x) + \tilde{\mathbb{I}}_{jbr}(x)}{N_{-jb'r'}(x)^2}\frac{N_{-jbr\cap b'r}(x)^2}{N_{-jbr}(x)^4} + \frac{\mathbb{I}_{jbr}(x)\mathbb{I}_{jb'r'}(x)}{N_{br}(x)^2N_{b'r'}(x)^2}\right] \\
&\lesssim \frac{n^2}{\lambda^{2d}}\sum_{j=1}^n \mathbb{E}\left[\frac{\mathbb{I}_{jbr}(x)\mathbb{I}_{br}(x)\mathbb{I}_{b'r'}(x)}{N_{br}(x)^2N_{b'r'}(x)^2}\right] \lesssim \frac{n^2}{\lambda^{2d}}\mathbb{E}\left[\frac{\mathbb{I}_{br}(x)\mathbb{I}_{b'r'}(x)}{N_{br}(x)N_{b'r'}(x)^2}\right] \lesssim \frac{n^2}{\lambda^{2d}}\frac{\lambda^d}{n}\frac{\lambda^{2d}\log n}{n^2} \lesssim \frac{\lambda^d\log n}{n} \to 0,
\end{aligned}
$$

where we applied Lemma 6. So $\sum_{i=1}^n S_i(x)^2 - n\,\mathbb{E}\left[S_i(x)^2\right] = O_{\mathbb{P}}\left(\frac{1}{\sqrt{B}} + \sqrt{\frac{\lambda^d \log n}{n}}\right) = o_{\mathbb{P}}(1)$.

Part 4: calculating the limiting variance

Thus by (Hall and Heyde, 2014, Theorem 3.2) we conclude that

$$
\sqrt{\frac{n}{\lambda^d}}\big(\hat{\mu}_{\mathrm{d}}(x) - \mathbb{E}\left[\hat{\mu}_{\mathrm{d}}(x) \mid \mathbf{X},\mathbf{T}\right]\big) \rightsquigarrow \mathcal{N}\big(0, \Sigma_{\mathrm{d}}(x)\big)
$$

as $n \to \infty$, assuming that the limit

$$\Sigma_{\mathrm{d}}(x) = \lim_{n\to\infty} \sum_{r=0}^{J} \sum_{r'=0}^{J} \omega_r \omega_{r'} \frac{n^2}{\lambda^d} \mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)}\right]$$

exists. Now we verify this and calculate the limit. Since J is fixed, it suffices to find

$$\lim_{n\to\infty} \frac{n^2}{\lambda^d} \mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)}\right]$$

for each $0 \le r, r' \le J$. Firstly, note that

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

Since σ^2 is Lipschitz and $\mathbb{P}\left(\max_{1\le l\le d} |T_b(x)_l| \ge t/\lambda\right) \le 2de^{-t/2}$ by Lemma 1, we have by Lemma 6 that

$$\begin{aligned}
\frac{n^2}{\lambda^d} \mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\big|\sigma^2(X_i) - \sigma^2(x)\big|}{N_{br}(x)N_{b'r'}(x)}\right] &\le 2de^{-t/2}\frac{n^2}{\lambda^d} + \frac{n^2}{\lambda^d}\frac{t}{\lambda} \mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)}{N_{br}(x)N_{b'r'}(x)}\right] \\
&\lesssim \frac{n^2}{\lambda^d}\frac{\log n}{\lambda}\frac{\lambda^d}{n^2} \lesssim \frac{\log n}{\lambda},
\end{aligned}$$

where we set $t = 4\log n$. Therefore

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

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

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

by a familiar argument based on Lemma 1, the Lipschitz property of $f(x)$, and Lemma 6. Hence

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

Now we apply Lemma 7 to approximate the expectation. With $N_{-ib'r'\setminus br}(x) = \sum_{j\ne i} \mathbb{I}\{X_j \in T_{b'r'}(x) \setminus T_{br}(x)\}$,

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

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

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

Now we bound these parameters above and below. Firstly, by applying Lemma 2 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 all sufficiently large t. Next, note that if f is L-Lipschitz in ℓ^2, by Lemma 1

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{\int_{T_{b'r'}(x)\setminus T_{br}(x)} f(\xi)\,\mathrm{d}\xi}{1 - \int_{T_{br}(x)} f(\xi)\,\mathrm{d}\xi} - f(x)|T_{b'r'}(x) \setminus T_{br}(x)|\right| > t\,\frac{|T_{b'r'}(x) \setminus T_{br}(x)|}{\lambda}\right) \\
&\leq \mathbb{P}\left(\int_{T_{b'r'}(x)\setminus T_{br}(x)} |f(\xi) - f(x)|\,\mathrm{d}\xi > t\,\frac{|T_{b'r'}(x) \setminus T_{br}(x)|}{2\lambda}\right) \\
&\quad + \mathbb{P}\left(\frac{\int_{T_{b'r'}(x)\setminus T_{br}(x)} f(\xi)\,\mathrm{d}\xi \cdot \int_{T_{br}(x)} f(\xi)\,\mathrm{d}\xi}{1 - \int_{T_{br}(x)} f(\xi)\,\mathrm{d}\xi} > t\,\frac{|T_{b'r'}(x) \setminus T_{br}(x)|}{2\lambda}\right) \\
&\leq \mathbb{P}\left(Ld\,|T_{b'r'}(x) \setminus T_{br}(x)| \max_{1\leq j\leq d} |T_{b'r'}(x)_j| > t\,\frac{|T_{b'r'}(x) \setminus T_{br}(x)|}{2\lambda}\right) \\
&\quad + \mathbb{P}\left(\|f\|_\infty\,|T_{b'r'}(x) \setminus T_{br}(x)|\frac{\|f\|_\infty|T_{br}(x)|}{1 - \|f\|_\infty|T_{br}(x)|} > t\,\frac{|T_{b'r'}(x) \setminus T_{br}(x)|}{2\lambda}\right) \\
&\leq \mathbb{P}\left(\max_{1\leq j\leq d} |T_{b'r'}(x)_j| > \frac{t}{2\lambda Ld}\right) + \mathbb{P}\left(|T_{br}(x)| > \frac{t}{4\lambda\|f\|_\infty^2}\right) \\
&\leq 2de^{-ta_r/(4Ld)} + 2de^{-ta_r/(8\|f\|_\infty^2)} \leq e^{-t/C},
\end{aligned}$$

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

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

So by Lemma 7 conditionally on $\mathbf{T}$, $N_{-ib'r'\cap br}(x)$, and $N_{-ibr\setminus b'r'}(x)$, with probability at least $1 - e^{-t/C}$,

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

Therefore by the same approach as the proof of Lemma 4, taking $t = 3C \log n$,

$$
\begin{aligned}
&\left|\mathbb{E}\left[\frac{|T_{br}(x) \cap T_{b'r'}(x)|}{(N_{-ibr}(x)+1)(N_{-ib'r'}(x)+1)} - \frac{|T_{br}(x) \cap T_{b'r'}(x)|}{(N_{-ibr}(x)+1)(N_{-ib'r'\cap br}(x)+nf(x)|T_{b'r'}(x) \setminus T_{br}(x)|+1)}\right]\right| \\
&\quad\lesssim \mathbb{E}\left[\frac{|T_{br}(x) \cap T_{b'r'}(x)|}{N_{-ibr}(x)+1} \frac{1+\frac{nt}{\lambda}|T_{b'r'}(x) \setminus T_{br}(x)|}{(N_{-ib'r'\cap br}(x)+n|T_{b'r'}(x) \setminus T_{br}(x)|+1)^2}\right] + e^{-t/C} \\
&\quad\lesssim \mathbb{E}\left[\frac{|T_{br}(x) \cap T_{b'r'}(x)|}{n|T_{br}(x)|+1} \frac{1+\frac{nt}{\lambda}|T_{b'r'}(x) \setminus T_{br}(x)|}{(n|T_{b'r'}(x)|+1)^2}\right] + e^{-t/C} \\
&\quad\lesssim \mathbb{E}\left[\frac{1}{n}\frac{1}{(n|T_{b'r'}(x)|+1)^2} + \frac{1}{n}\frac{t/\lambda}{n|T_{b'r'}(x)|+1}\right] + e^{-t/C} \\
&\quad\lesssim \frac{\lambda^{2d}\log n}{n^3} + \frac{\log n}{n\lambda}\frac{\lambda^d}{n} \lesssim \frac{\lambda^d}{n^2}\left(\frac{\lambda^d \log n}{n} + \frac{\log n}{\lambda}\right).
\end{aligned}
$$

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

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

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

$$
\begin{aligned}
&\frac{n^2}{\lambda^d}\left|\mathbb{E}\left[\frac{|T_{br}(x) \cap T_{b'r'}(x)|}{N_{-ibr}(x)+1} \frac{1}{N_{-ib'r'\cap br}(x)+nf(x)|T_{b'r'}(x) \setminus T_{br}(x)|+1}\right]\right. \\
&\qquad \left. - \mathbb{E}\left[\frac{|T_{br}(x) \cap T_{b'r'}(x)|}{N_{-ibr\cap b'r'}(x)+nf(x)|T_{br}(x) \setminus T_{b'r'}(x)|+1} \frac{1}{N_{-ib'r'\cap br}(x)+nf(x)|T_{b'r'}(x) \setminus T_{br}(x)|+1}\right]\right| \\
&\quad\lesssim \frac{n^2}{\lambda^d}\mathbb{E}\left[\frac{1+\frac{nt}{\lambda}|T_{br}(x) \setminus T_{b'r'}(x)|}{(N_{-ibr\cap b'r'}(x)+n|T_{br}(x) \setminus T_{b'r'}(x)|+1)^2} \frac{|T_{br}(x) \cap T_{b'r'}(x)|}{N_{-ib'r'\cap br}(x)+nf(x)|T_{b'r'}(x) \setminus T_{br}(x)|+1}\right] \\
&\qquad + \frac{n^2}{\lambda^d} e^{-t/C} \lesssim \frac{\lambda^d \log n}{n} + \frac{\log n}{\lambda}.
\end{aligned}
$$

Thus far we have proven that

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

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

$$
\begin{aligned}
N_{-ibr\cap b'r'}(x) &\leq \mathrm{Bin}\left(n,\, |T_{br}(x) \cap T_{b'r'}(x)|\left(f(x) + \frac{t}{\lambda}\right)\right), \\
N_{-ibr\cap b'r'}(x) &\geq \mathrm{Bin}\left(n\left(1 - \frac{t^{d+1}}{\lambda^d} - \frac{1}{n}\right),\, |T_{br}(x) \cap T_{b'r'}(x)|\left(f(x) - \frac{t}{\lambda}\right)\right).
\end{aligned}
$$

Therefore by Lemma 7 applied conditionally on $\mathbf{T}$, with probability at least $1 - e^{-t/C}$,

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

Now by Lemma 5, with $t = 3C \log n$,

$$\frac{n^2}{\lambda^d} \Bigg| \mathbb{E}\left[\frac{|T_{br}(x) \cap T_{b'r'}(x)|}{N_{-ibr\cap b'r'}(x) + nf(x)|T_{br}(x) \setminus T_{b'r'}(x)| + 1} \frac{1}{N_{-ib'r'\cap br}(x) + nf(x)|T_{b'r'}(x) \setminus T_{br}(x)| + 1} \right] - \mathbb{E}\left[\frac{|T_{br}(x) \cap T_{b'r'}(x)|}{nf(x)|T_{br}(x)| + 1} \frac{1}{nf(x)|T_{b'r'}(x)| + 1} \right] \Bigg|$$
$$\lesssim \frac{n^2}{\lambda^d} \mathbb{E}\left[|T_{br}(x) \cap T_{b'r'}(x)| \frac{1 + \frac{nt}{\lambda}|T_{br}(x) \cap T_{b'r'}(x)|}{(n|T_{br}(x)| + 1)(n|T_{b'r'}(x)| + 1)} \left(\frac{1}{n|T_{br}(x)| + 1} + \frac{1}{n|T_{b'r'}(x)| + 1} \right) \right] + \frac{n^2}{\lambda^d} e^{-t/C}$$
$$\lesssim \frac{n^2}{\lambda^d} \frac{1}{n^3} \mathbb{E}\left[\frac{1 + \frac{nt}{\lambda}|T_{br}(x) \cap T_{b'r'}(x)|}{|T_{br}(x)||T_{b'r'}(x)|} \right] + \frac{n^2}{\lambda^d} e^{-t/C}$$
$$\lesssim \frac{1}{n\lambda^d} \mathbb{E}\left[\frac{1}{|T_{br}(x)||T_{b'r'}(x)|} \right] + \frac{t}{\lambda^{d+1}} \mathbb{E}\left[\frac{1}{|T_{br}(x)|} \right] + \frac{n^2}{\lambda^d} e^{-t/C} \lesssim \frac{\lambda^d}{n} + \frac{\log n}{\lambda}.$$

This allows us to deduce that

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

Now that we have reduced the limiting 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

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

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

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

and note that

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

for some $C > 0$ and sufficiently large λ. Hence by the Cauchy–Schwarz inequality and Lemma 5,

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

as $\log \lambda \gtrsim \log n$. Therefore

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

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

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

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

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

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

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

Finally we conclude by giving the limiting variance.

$$\begin{aligned}
&\sum_{r=0}^{J}\sum_{r'=0}^{J} \omega_r \omega_{r'} \frac{n^2}{\lambda^d} \mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)}\right] \\
&\quad= \frac{\sigma^2(x)}{f(x)} \sum_{r=0}^{J}\sum_{r'=0}^{J} \omega_r \omega_{r'} \left(\frac{2a_{r'}}{3}\left(1 - \frac{a_{r'}}{a_r}\log\left(\frac{a_r}{a_{r'}} + 1\right)\right) + \frac{2a_r}{3}\left(1 - \frac{a_r}{a_{r'}}\log\left(\frac{a_{r'}}{a_r} + 1\right)\right)\right)^d \\
&\qquad + O\left(\frac{(\log n)^{d+1}}{\lambda} + \frac{\lambda^d \log n}{n}\right).
\end{aligned}$$

So the limit exists and

$$\Sigma_{\mathrm{d}}(x) = \frac{\sigma^2(x)}{f(x)} \sum_{r=0}^{J}\sum_{r'=0}^{J} \omega_r \omega_{r'} \left(\frac{2a_r}{3}\left(1 - \frac{a_r}{a_{r'}}\log\left(\frac{a_{r'}}{a_r} + 1\right)\right) + \frac{2a_{r'}}{3}\left(1 - \frac{a_{r'}}{a_r}\log\left(\frac{a_r}{a_{r'}} + 1\right)\right)\right)^d.$$

□

Proof of bias characterization with debiasing

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

Proof (Theorem 6)
By the definition of the debiased estimator and Theorem 2, since J and a_r are fixed,

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

It remains to evaluate the first term. 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 of variance estimator consistency with debiasing

Proof (Theorem 7)
Part 1: consistency of $\hat{\sigma}^2(x)$
Recall that

$$
\hat{\sigma}^2(x) = \frac{1}{B} \sum_{b=1}^{B} \frac{\sum_{i=1}^{n} Y_i^2 \, \mathbb{I}\{X_i \in T_b(x)\}}{\sum_{i=1}^{n} \mathbb{I}\{X_i \in T_b(x)\}} - \hat{\mu}(x)^2. \tag{14}
$$

The first term in (14) 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 Lipschitz, where $\mathbb{E}[Y_i^4 \mid X_i]$ is bounded almost surely. So its conditional bias is controlled by Theorem 2 and is at most $O_{\mathbb{P}}\left(\frac{1}{\lambda} + \frac{\log n}{\lambda} \sqrt{\lambda^d / n} \right)$. Its variance is at most $\frac{\lambda^d}{n}$ by Theorem 5. Consistency of the second term in (14) follows directly from Theorems 2 and 5 with the same bias and variance bounds. Therefore

$$
\hat{\sigma}^2(x) = \sigma^2(x) + O_{\mathbb{P}}\left(\frac{1}{\lambda} + \sqrt{\frac{\lambda^d}{n}} \right).
$$

Part 2: consistency of the sum
Note that

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

This is exactly the same as the quantity in (11), if we were to take ε_i to be ± 1 with equal probability. Thus we immediately have convergence in probability by the proof of Theorem 5:

$$
\begin{aligned}
\frac{n}{\lambda^d} \sum_{i=1}^{n} \left(\sum_{r=0}^{J} \omega_r \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{I}\{X_i \in T_{rb}(x)\}}{\sum_{i=1}^{n} \mathbb{I}\{X_i \in T_{rb}(x)\}} \right)^2 &= \frac{n^2}{\lambda^d} \sum_{r=0}^{J} \sum_{r'=0}^{J} \omega_r \omega_{r'} \mathbb{E}\left[\frac{\mathbb{I}_{ibr}(x) \mathbb{I}_{ib'r'}(x)}{N_{br}(x) N_{b'r'}(x)} \right] \\
&\quad + O_{\mathbb{P}}\left(\frac{1}{\sqrt{B}} + \sqrt{\frac{\lambda^d \log n}{n}} \right).
\end{aligned}
$$

Part 3: conclusion

Again by the proof of Theorem 5 with ε_i being ± 1 with equal probability, and by the previous parts,

$$\hat{\Sigma}_{\mathrm{d}}(x) = \Sigma_{\mathrm{d}}(x) + O_{\mathbb{P}}\left(\frac{(\log n)^{d+1}}{\lambda} + \frac{1}{\sqrt{B}} + \sqrt{\frac{\lambda^d \log n}{n}} \right).$$

□

Proof of confidence interval validity with debiasing

Proof (Theorem 8)
By Theorem 6 and Theorem 7,

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

The first term now converges weakly to $\mathcal{N}(0,1)$ by Slutsky's theorem, Theorem 5, and Theorem 7, while the second term is $o_{\mathbb{P}}(1)$ by assumption. Validity of the confidence interval follows immediately. □

Proof of minimax optimality with debiasing

Proof (Theorem 9)
The bias–variance decomposition along with Theorem 6 and the proof of Theorem 5 with $J = \lfloor \underline{\beta}/2 \rfloor$ gives

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

Note that we used an L^2 version of Theorem 6 which is immediate from the proof of Theorem 2, since we obtain the bound in probability through Chebyshev's inequality. Now since $\lambda \asymp n^{\frac{1}{d+2\beta}}$ and $B \gtrsim n^{\frac{2\beta-2}{d+2\beta}}$,

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

□

References

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.

Biau, G. (2012). Analysis of a random forests model. *The 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. (2016). *Concentration inequalities*. Oxford.

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

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*, page 18.

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.

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. (2016). Impact of subsampling and pruning on random forests. *arXiv preprint arXiv:1603.04261*.

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

Fan, J., Li, R., Zhang, C.-H., and Zou, H. (2020). *Statistical Foundations of Data Science*. CRC press.

Friedberg, R., Tibshirani, J., Athey, S., and Wager, S. (2020). Local linear forests. *Journal of Computational and Graphical Statistics*, 30(2):503–517.

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

Hall, P. and Heyde, C. C. (2014). *Martingale limit theory and its application*. Academic press.

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

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. (2023). Large scale prediction with decision trees. *Journal of the American Statistical Association*.

Lakshminarayanan, B., Roy, D. M., and Teh, Y. W. (2014). Mondrian forests: Efficient online random forests. *Advances in neural information processing systems*, 27.

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

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.

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.

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.

Roy, D. M., Teh, Y. W., et al. (2008). The Mondrian process. In *NIPS*, 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*, pages 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 (Big Data)*, pages 2785–2790. IEEE.

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