Inference with Mondrian Random Forests

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

July 16, 2023

Abstract

Keywords: Mondrian process, regression trees, random forests, central limit theorem, bias correction, inference, confidence intervals.

Contents

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

[*] Corresponding author: wgu2@princeton.edu

1 Introduction

Random forest estimators, first introduced by Breiman (2001), have long established a reputation for providing state-of-the-art methods for classification and regression tasks. Their desirable traits include their ability to adapt to a wide range of large and high-dimensional data sets, their computational efficiency, their simplicity of configuration and their amenability to tuning parameter selection. Random forests have achieved unparalleled successes in many fields of study, including healthcare, finance, online commerce, text analysis, bioinformatics, image classification and ecology.

Despite their widespread empirical success, the theoretical understanding of random forests is somewhat less well developed. In particular, there are several open questions regarding the use of random forests for performing principled statistical inference. Wager and Athey (2018) introduced causal random forests which use trees which satisfy and honesty criterion in order to show asymptotic normality. Oprescu et al. (2019) defined orthogonal random forests, which use a two-step procedure along with Neyman orthogonality to establish a central limit theorem.

In this paper we focus on the nonparametric regression setting. Mondrian random forests (Lakshminarayanan et al., 2014) offer a simplified modification of traditional random forest methods in which the partition is generated independent of the data and according to a canonical stochastic process known as the Mondrian process (Roy et al., 2008). This process takes a single parameter λ known as the "lifetime" and enjoys various mathematical properties which allow for precise analysis of the statistical properties of Mondrian random forests, and further enables them to be fitted in an online manner. Mondrian forests containing just a single tree have recently been shown to be minimax optimal when the regression function is β-Hölder for $\beta \in (0, 1]$, and the use of large forests extends this result to $\beta \in (0, 2]$ (Mourtada et al., 2020). However, neither methods for statistical inference with Mondrian random forests nor extensions to $\beta > 2$ have yet been studied.

As such, our main results include a central limit theorem for the Mondrian random forest estimator and a debiased version of the Mondrian random forest which is minimax-optimal for all $\beta > 0$. The debiasing procedure is also useful when conducting statistical inference, providing a principled method for ensuring that the bias is negligible which is analogous to undersmoothing techniques for kernel estimators.

Our paper is structured as follows. In Section 2 we introduce the Mondrian process (Roy et al., 2008) and give our assumptions on the data generating process, using a Hölder smoothness assumption on the regression function to control the bias of various estimators. We define the Mondrian random forest estimator (Lakshminarayanan et al., 2014) and give our assumptions on its lifetime parameter and the number of trees. We define 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 bias of the estimator, with the aim of applying a debiasing procedure later on. To enable valid feasible statistical inference, we provide a consistent variance estimator in Lemma 1 and briefly discuss implications for lifetime parameter selection.

In Section 4 we define debiased Mondrian random forests, a collection of new estimators based on linear combinations of Mondrian random forests with varying lifetime parameters. The parameters are carefully chosen to annihilate leading terms in our bias characterization, yielding an estimator with provably superior bias properties (Theorem 4). 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 (Lemma 2). As a final corollary of the improved bias properties, we demonstrate in Theorem 7 that the debiased Mondrian random forest estimator is minimax-optimal for all $\beta > 0$, providing that β is known a priori.

Section 5 contains details on tuning parameter selection, beginning with a data-driven approach to selecting the crucial lifetime parameter using local polynomial estimation. We also give advice on choosing the number of trees, the debiasing order and the debiasing coefficients.

Finally we give our concluding remarks in Section 6.

2 Setup

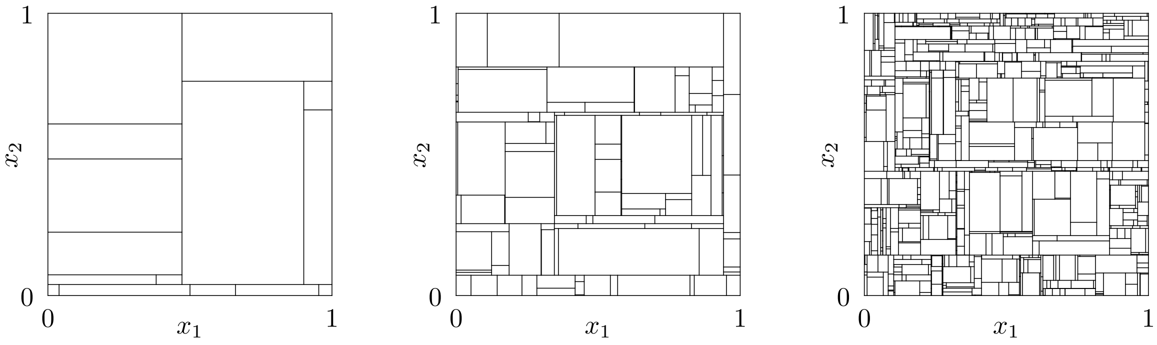
When using a Mondrian random forest, there are two sources of randomness. The first is of course the data, and here we consider the regression setting. The second source is a collection of independent trees drawn from a Mondrian process, which we define in the subsequent section.

2.1 The Mondrian process

The Mondrian process was introduced by Roy et al. (2008) and offers a canonical method for generating random 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 is $|D|_1 = \sum_{j=1}^d |D_j|$.

To sample a partition T from the Mondrian process $\mathcal{M}\left([0,1]^d, \lambda\right)$ we start at time $t = 0$ with the trivial partition 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 $t_D + E_D \le \lambda$, 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 $S_J \sim \operatorname{Unif}\left[D_J^-, D_J^+\right]$. The cell D splits into the two new cells $\{x \in D : x_j \le S_J\}$ 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 $T \sim \mathcal{M}\left([0,1]^d, \lambda\right)$ for $d = 2$ and with different lifetime parameters λ.



(a) $\lambda = 3$

(b) $\lambda = 10$

(c) $\lambda = 30$

Figure 1: The Mondrian process $T \sim \mathcal{M}\left([0,1]^d, \lambda\right)$ with $d = 2$ and varying lifetime parameter 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. Then we say a function $g : [0,1]^d \to \mathbb{R}$ is β-Hölder continuous and write $g \in \mathcal{H}^\beta$ if there is a constant $C > 0$ with

$$\max_{1 \le |\nu| \le \underline{\beta}} |\partial^\nu \mu(x)| \le C \qquad \text{and} \qquad \max_{|\nu| = \underline{\beta}} |\partial^\nu \mu(x) - \partial^\nu \mu(x')| \le C \|x - x'\|_2^{\beta - \underline{\beta}}$$

for 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 $\partial^\nu \mu(x) = \partial^{|\nu|}\mu(x) / \prod_{j=1}^d \partial x_j^{\nu_j}$. 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 that X_i has a Lebesgue density function $f_X(x)$ on $[0,1]^d$ which is bounded away from zero and satisfies $f_X \in \mathcal{H}^\beta$ for some $\beta \geq 1$. Suppose $\mathbb{E}[Y_i^2 \mid X_i]$ is bounded almost surely, 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 that $\sigma^2(X_i) = \mathbb{E}[\varepsilon_i^2 \mid X_i]$ is Lipschitz.

Note that Assumption 1 requires neither X_i nor Y_i to be compactly supported, and $f_X(x)$ must exist and be bounded away from zero only on $[0,1]^d$.

2.3 Mondrian random forests

We define here the basic Mondrian random forest estimator according to Mourtada et al. (2020), and extend it to a debiased version in Section 4. For a lifetime parameter $\lambda > 0$ and forest size $B \geq 1$, let $(T_b : 1 \leq b \leq B)$ be a Mondrian forest where $T_b \sim \mathcal{M}\left([0,1]^d, \lambda\right)$ are mutually independent Mondrian trees which are independent of the data. For $x \in [0,1]^d$, the Mondrian random forest estimator of $\mu(x)$ is

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

To avoid boundary issues we assume $x \in (0,1)^d$ throughout. To ensure the bias and variance of the Mondrian random forest estimator converge to zero (see Section 3), we impose some basic conditions on λ and B in Assumption 2.

Assumption 2 (Mondrian random forest estimator)
Suppose that $x \in (0,1)^d$ is an interior point, $\frac{\lambda^d}{n} \to 0$ and $\log \lambda \asymp \log B \asymp \log n$.

2.4 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. The Lebesgue measure on $\mathbb{R}^d$ is Leb. 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. Write $A_n = o_{\mathbb{P}}(A_n)$ if $A_n / B_n \to 0$ in probability. For $a, b \in \mathbb{R}$, define $a \wedge b = \min\{a, b\}$.

3 Inference with Mondrian random forests

We begin with a novel central limit theorem for the Mondrian random forest estimator (1), which forms the core of our methodology for performing inference with such estimators.

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

$$\sqrt{\frac{n}{\lambda^d}} \Big(\hat{\mu}(x) - \mathbb{E}\big[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\big] \Big) \rightsquigarrow \mathcal{N}\big(0, \Sigma(x)\big) \quad \textit{where} \quad \Sigma(x) = \frac{\sigma^2(x)}{f_X(x)} \left(\frac{4 - 4 \log 2}{3} \right)^d \approx \frac{\sigma^2(x)}{f_X(x)} 0.4091^d.$$

The proof of Theorem 1 is based on a central limit theorem for martingale difference sequences (Hall and Heyde, 2014), and uses the Efron–Stein inequality to handle the non-linear dependence of the estimator on the data. We also rely on several somewhat delicate preliminary lemmas concerning moments of the denominator in (1). All technical lemmas and proofs are available in Appendix A. Our next result characterizes the bias of the Mondrian random forest estimator as a polynomial in $1/\lambda$.

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_X 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 second-order term

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

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

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

Note that if the forest size B does not diverge to infinity then we suffer a first-order bias which decays as $1/\lambda$. This phenomenon where large forests remove first-order bias was noted by Mourtada et al. (2020). The leading second-order bias given in Theorem 2 is reminiscent of the leading bias for the Nadaraya–Watson estimator.

Using Theorem 1 and Theorem 2 together gives a central limit theorem for the Mondrian random forest estimator which can be used, for example, to build (infeasible) confidence intervals for the unknown regression function $\mu(x)$ whenever the bias shrinks faster than the standard deviation. In general this will require $\frac{1}{\lambda^2} \ll \sqrt{\frac{\lambda^d}{n}}$, which can be framed as the restriction $\lambda \gg n^{\frac{1}{d+4}}$ on the lifetime parameter λ.

If instead we aim for optimal point estimation, then balancing the bias and standard deviation requires $\frac{1}{\lambda^2} \asymp \sqrt{\frac{\lambda^d}{n}}$, or equivalently $\lambda \asymp n^{\frac{1}{d+4}}$. Such a choice of λ gives the convergence rate $n^{-\frac{2}{d+4}}$, which is the minimax-optimal rate of convergence for 2-Hölder functions, as shown by Mourtada et al. (2020, Theorem 2). In Section 4 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.

In order to conduct feasible inference, we propose the following variance estimator. First, define

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

Then let

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

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

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

For details on choosing the lifetime parameter λ in practice, see Section 5. As an example application of our feasible inference methodology, Theorem 3 shows how to construct valid confidence intervals for the regression function $\mu(x)$.

Theorem 3 (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 further that* $\lambda \gg n^{\frac{1}{d+4\wedge 2\beta}}$ *and* $B \gg n^{\frac{4\wedge 2\beta - 2}{d+4\wedge 2\beta}}$*. For a confidence level* $\alpha \in (0,1)$*, Let* $q_{1-\alpha/2}$ *be the normal quantile satisfying* $\mathbb{P}\left(\mathcal{N}(0,1) \leq q_{1-\alpha/2}\right) = 1 - \alpha/2$*. Then*

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

4 Debiased Mondrian random forests

In this section we propose a variant of the Mondrian random forest estimator which corrects for higher-order bias. This estimator retains the basic form of a random forest estimator in the sense that it is a linear combination of tree estimators, but in this section we allow for non-identical 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 based on generalized jackknifing. Let $J \geq 0$ be the generalized jackknife correction order. With $J = 0$ we retain the original Mondrian forest estimator and with $J = \lfloor \underline{\beta}/2 \rfloor$ we remove as much bias as is possible in the Hölder class $\mathcal{H}^\beta$. For $0 \leq r \leq J$ let $\hat{\mu}_r(x)$ be a Mondrian forest estimator based on the independent 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

$$\hat{\mu}_{\mathrm{J}}(x) = \sum_{r=0}^{J} \omega_r \hat{\mu}_r(x) = \sum_{r=0}^{J} \omega_r \frac{1}{B} \sum_{b=1}^{B} \frac{\sum_{i=1}^n Y_i \mathbb{I}\big\{X_i \in T_{rb}(x)\big\}}{\sum_{i=1}^n \mathbb{I}\big\{X_i \in T_{rb}(x)\big\}}. \tag{3}$$

Note that this ensemble estimator retains the "forest" structure of the original estimators, but with varying 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 $\sum_{r=0}^J \omega_r = 1$ and $\sum_{r=0}^J \omega_r a_r^{-2s} = 0$ for each $1 \leq s \leq J$. 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 $e_0 = (1, 0, \ldots, 0)^\mathsf{T} \in \mathbb{R}^{J+1}$ so that a solution for the jackknifing coefficients is $\omega = A^{-1} e_0$ whenever A is non-singular. In practice we take a_r to be a fixed geometric or arithmetic sequence to ensure this is the case. For example, we could set $a_r = \gamma^r$ for some $\gamma > 1$ or $a_r = 1 + \gamma r$ for some $\gamma > 0$ as proposed by Cattaneo et al. (2013). The improved bias bound presented in Theorem 4 is an easy consequence of the construction of this debiased Mondrian random forest estimator.

Theorem 4 (Bias of the debiased Mondrian random forest estimator)
Suppose Assumptions 1 and 2 hold. Then in the notation of Theorem 2 and with $\bar{\omega} = \sum_{l=0}^J \omega_l a_l^{-2J-2}$*,*

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

Hence the leading bias can be characterized whenever the debiasing order J *does not fully exhaust the Hölder smoothness* β*.*

Next we verify that a central limit theorem holds for this debiased random forest estimator, and give its limiting variance.

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

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

where

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

In order to conduct feasible inference, we propose the following variance estimator. With $\hat{\sigma}^2(x)$ as in (2) in Section 3, let

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

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

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

Theorem 6 (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 $\frac{\lambda^d \log n}{n} \to 0$. Take $J = \lfloor \beta/2 \rfloor$ and assume further 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 the normal quantile satisfying $\mathbb{P}\left(\mathcal{N}(0,1) \leq q_{1-\alpha/2}\right) = 1 - \alpha/2$. Then

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

4.1 Minimax optimality

Our next result shows that when using an appropriately-chosen sequence of lifetime parameters λ, the debiased Mondrian random forest estimator achieves the minimax-optimal rate for the problem of estimating a regression function $\mu \in \mathcal{H}^\beta$ on $[0,1]^d$ (Stone, 1982). We assume that the smoothness β is known.

Theorem 7 (Minimax optimality of the debiased Mondrian random forest estimator)
Grant Assumptions 1 and 2. Let $J = \lfloor \beta/2 \rfloor$, take $\lambda \asymp n^{\frac{1}{d+2\beta}}$ and suppose $B \gtrsim n^{\frac{2\beta-2}{d+2\beta}}$. Then

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

4.2 Interpretation

The debiased Mondrian random forest estimator defined in (3) is a linear combination of Mondrian random forests, and as such contains 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 the 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 the existing model aggregation literature, with the outermost operation representing a B-fold randomization of base learners, each of which has a small bias component.

5 Tuning parameter selection

In this section 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 bias correction J and the debiasing scale parameters a_r for $0 \leq r \leq J$.

5.1 Selecting the base lifetime parameter λ

The most important parameter is the base Mondrian lifetime parameter λ, which has 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, the minimax optimality result of Theorem 7 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. Secondly, in order to use the central limit theorem established in Theorem 5, we must have that the bias converges to zero strictly faster than the standard deviation, yielding $\frac{1}{\lambda^\beta} \ll \sqrt{\frac{\lambda^d}{n}}$, or equivalently $\lambda \gg n^{\frac{1}{d+2\beta}}$.

Naturally these conditions are incompatible, since the former balances the bias and standard deviation where the latter requires that the standard deviation dominate the bias. One solution to this is to employ an undersmoothing procedure as follows. To obtain point estimates of $\mu(x)$ we simply choose $\lambda \asymp n^{\frac{1}{d+2\beta}}$ as usual. However to establish confidence intervals for $\mu(x)$ we first replace β by $\tilde\beta < \beta$ and take $\tilde\lambda \asymp n^{\frac{1}{d+2\tilde\beta}} \gg n^{\frac{1}{d+2\beta}}$ to guarantee $\frac{1}{\tilde\lambda^\beta} \ll \sqrt{\frac{\tilde\lambda^d}{n}}$. The use of the term undersmoothing is in direct analogy with the classical kernel case, where the bandwidth h plays the same role as $1/\lambda$ in our setting.

It remains to propose a concrete method for selecting λ in the finite-sample setting. We suggest the following method based on Fan and Gijbels (1996, Section 4.2). Firstly suppose that $X_i \sim \mathrm{Unif}\left([0,1]^d\right)$ and that the leading bias of $\hat\mu_{\mathrm{J}}$ is well approximated by an additively separable polynomial:

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

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

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

Thus, writing $\partial_j^{2J+2}\mu(x)$ for $\frac{\partial_j^{2J+2}\mu(x)}{\partial x_j^{2J+2}}$, the asymptotic integrated mean squared error (AIMSE) is

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

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

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

An estimator of λ_{AIMSE} is therefore given by

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

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, as suggested in the one-dimensional case by Fan and Gijbels (1996, Section 4.2). 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:

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

Alternatively one could simply use a general-purpose cross-validation scheme, such as leave-one-out cross-validation or repeated k-fold cross-validation.

5.2 Selecting the number of trees in each forest B

The next parameter to choose is the number of trees in each forest, B. Our results place no upper bound on B, so it is natural to take B to be as large as is computationally feasible. We suggest taking $B = n$, which is certainly large enough to satisfy the constraint in Theorem 7 for instance.

5.3 Selecting 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_X, which is in practice very difficult to estimate statistically. As such we recommend removing only the first few bias terms to avoid overly inflating the variance of the estimator.

5.4 Selecting the debiasing coefficients a

As mentioned in Section 4, we take a_r to be a fixed geometric or arithmetic sequence. For example, we could set $a_r = \gamma^r$ for some $\gamma > 1$ or $a_r = 1 + \gamma r$ for some $\gamma > 0$, as in Cattaneo et al. (2013). We suggest for concreteness taking $a_r = \gamma^r$ with $\gamma = 1.05$, yielding the debiasing coefficient matrix $A_{rs} = a_{r-1}^{2-2s} = \gamma^{(r-1)(2-2s)}$. The determinant of this Vandermonde matrix is $\prod_{0 \leq r < r' \leq J}(a_r^{-2} - a_{r'}^{-2}) = \prod_{0 \leq r < r' \leq J}(\gamma^{-2r} - \gamma^{-2r'})$ so A is invertible for all J.

6 Conclusion

We presented a novel 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, demonstrating their advantages for statistical inference and their minimax optimality properties. Finally we discussed tuning parameter selection, enabling fully feasible estimation and inference procedures.

A Proofs

Throughout this section we will use the following simplified notation for convenience, whenever it is appropriate. We write $\mathbb{I}_{ib}(x) = \mathbb{I}\left\{X_i \in T_b(x)\right\}$ and $N_b(x) = \sum_{i=1}^n \mathbb{I}_{ib}(x)$, as well as $\mathbb{I}N_b(x) = \mathbb{I}\left\{N_b(x) \geq 1\right\}$. We begin by bounding the maximum size of any cell in a Mondrian forest containing x.

Lemma 3 (Upper bound on the largest cell in a Mondrian forest)
Let $T_1, \ldots, T_b \sim \mathcal{M}([0,1]^d, \lambda)$ 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 3)
By Mourtada et al. (2020, Proposition 1), $|T_b(x)_j| = \left(\frac{E_{bj1}}{\lambda} \wedge x_j\right) + \left(\frac{E_{bj2}}{\lambda} \wedge (1 - x_j)\right)$ where E_{bj1} and E_{bj2} are independent $\mathrm{Exp}(1)$ random variables. Thus $|T_b(x)_j| \leq \frac{E_{bj1} + E_{bj2}}{\lambda}$ and so by a union bound

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

□

Next we give a series of results culminating in a generalized moment bound for the denominator appearing in the Mondrian random forest estimator.

Lemma 4 (An inverse moment bound for the binomial distribution)
For $n \geq 1$ and $p \in [0,1]$, let $N \sim \mathrm{Bin}(n,p)$ 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 4)
By Bernstein's inequality, $\mathbb{P}\left(N \leq np - t\right) \leq \exp\left(-\frac{t^2/2}{np(1-p) + t/3}\right) \leq \exp\left(-\frac{3t^2}{6np + 2t}\right)$. Therefore we have $\mathbb{P}\left(N \leq np/4\right) \leq \exp\left(-\frac{27n^2p^2/16}{6np + 3np/2}\right) = e^{-9np/40}$. Partitioning by this event gives

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

□

Lemma 5 (Upper bound on the number of active data points)
Suppose that Assumptions 1 and 2 hold, and let $T_1, \ldots, T_B \sim \mathcal{M}\big([0,1]^d, \lambda\big)$ be independent. Define $N_\cup(x) = \sum_{i=1}^n \mathbb{I}\left\{X_i \in \bigcup_{b=1}^B T_b(x)\right\}$. Then for sufficiently large n and t, with $\|f_X\|_\infty = \sup_{x \in [0,1]^d} f_X(x)$,

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

Proof (Lemma 5)
Note $N_\cup(x) \sim \mathrm{Bin}\left(n, \int_{\bigcup_{b=1}^B T_b(x)} f_X(s) \,\mathrm{d}s\right) \leq \mathrm{Bin}\left(n, 2^d \max_{1 \leq b \leq B} \max_{1 \leq j \leq d} |T_b(x)_j|^d \|f_X\|_\infty\right)$ conditionally on $\mathbf{T}$. Now if $N \sim \mathrm{Bin}(n,p)$ then $\mathbb{P}\left(N \geq (1+t)np\right) \leq \exp\left(-\frac{t^2 n^2 p^2/2}{np(1-p)+tnp/3}\right) \leq \exp\left(-\frac{3t^2 np}{6+2t}\right)$ by Bernstein's inequality. Thus for $t \geq 2$,

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

By Lemma 3, $\mathbb{P}\left(\max_{1 \leq b \leq B} \max_{1 \leq j \leq d} |T_j(x)| > \frac{t}{\lambda}\right) \leq 2dBe^{-t/2}$. Hence

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

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

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

□

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

$$\mathbb{E}\left[\prod_{b=1}^{B_0} \frac{\mathbb{I} N_b(x)}{N_b(x)^{k_b}}\right] \leq \left(\frac{36k}{\inf_x f_X(x)}\right)^{2^{B_0} k} \prod_{b=1}^{B_0} \mathbb{E}\left[1 \wedge \frac{1}{(n|T_b(x)|)^{k_b}}\right]$$

for sufficiently large n.

Proof (Lemma 6)
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 cardinality of $\mathcal{D}$, which represents the subcells we have finished checking, 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 $|D| = \mathrm{Leb}\, D$ and $N_b(D) = \sum_{i=1}^n \mathbb{I}\left\{X_i \in D\right\}$ so that $N_b(x) = \sum_{D \in \mathcal{A}(b)} N_b(D)$. Note that for any $D \in \mathcal{R} \setminus \mathcal{D}$,

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

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

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

conditional on $\mathbf{T}$ and $N_b(D') : D' \in \mathcal{R} \setminus (\mathcal{D} \cup \{D\})$. Further, for sufficiently large t by Lemma 5 we have

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

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

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

Therefore plugging this back into the marginal expectation yields

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

Now we apply induction, starting with $\mathcal{D} = \emptyset$ and adding $D \in \mathcal{R} \setminus \mathcal{D}$ to $\mathcal{D}$ until $\mathcal{D} = \mathcal{R}$. This takes at most $|\mathcal{R}| \leq 2^{B_0}$ steps and yields

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

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

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

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

Lemma 7 (Inverse moments of the volume of a Mondrian cell)
Suppose Assumption 2 holds and let $T \sim \mathcal{M}\left([0,1]^d, \lambda\right)$. 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 7)
Observe that by Mourtada et al. (2020, Proposition 1), we have $|T(x)| = \prod_{j=1}^d \left(\frac{1}{\lambda}E_{j1} \wedge x_j + \frac{1}{\lambda}E_{j2} \wedge (1-x_j)\right)$ 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,

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

□

Lemma 8 (Simplified generalized moment bound for Mondrian random forest denominators)
Grant Assumptions 1 and 2. Let $T_b \sim \mathcal{M}\left([0,1]^d, \lambda\right)$ and $k_b \geq 1$ for $1 \leq b \leq B_0$. Then with $k = \sum_{b=1}^{B_0} k_b$,

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

for sufficiently large n.

Proof (Lemma 8)
This follows directly from Lemmas 6 and 7. □

Lemma 9 (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 9)
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

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

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

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

□

Proof (Theorem 1)
This follows from the proof of Theorem 5 with $J = 0$. □

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,

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

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

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

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

$$\begin{aligned}
&\operatorname{Var}\left[\sum_{i=1}^{n}\big(\mu(X_i) - \mu(x)\big)\mathbb{E}\left[\frac{\mathbb{I}_{ib}(x)}{N_b(x)} \;\middle|\; \mathbf{X}\right]\right] \\
&\quad\le \frac{1}{2}\sum_{j=1}^{n}\mathbb{E}\left[\left(\sum_{i=1}^{n}\big(\mu(X_i) - \mu(x)\big)\mathbb{E}\left[\frac{\mathbb{I}_{ib}(x)}{N_b(x)} \;\middle|\; \mathbf{X}\right] - \sum_{i=1}^{n}\Big(\mu(\tilde{X}_{ij}) - \mu(x)\Big)\mathbb{E}\left[\frac{\tilde{\mathbb{I}}_{ijb}(x)}{\tilde{N}_{jb}(x)} \;\middle|\; \tilde{\mathbf{X}}_j\right]\right)^2\right] \\
&\quad\le \frac{1}{2}\sum_{j=1}^{n}\mathbb{E}\left[\left(\sum_{i=1}^{n}\left(\big(\mu(X_i) - \mu(x)\big)\frac{\mathbb{I}_{ib}(x)}{N_b(x)} - \Big(\mu(\tilde{X}_{ij}) - \mu(x)\Big)\frac{\tilde{\mathbb{I}}_{ijb}(x)}{\tilde{N}_{jb}(x)}\right)\right)^2\right] \\
&\quad\le \sum_{j=1}^{n}\mathbb{E}\left[\left(\sum_{i\neq j}\big(\mu(X_i) - \mu(x)\big)\left(\frac{\mathbb{I}_{ib}(x)}{N_b(x)} - \frac{\mathbb{I}_{ib}(x)}{\tilde{N}_{jb}(x)}\right)\right)^2\right] + 2\sum_{j=1}^{n}\mathbb{E}\left[\big(\mu(X_j) - \mu(x)\big)^2\frac{\mathbb{I}_{jb}(x)}{N_b(x)^2}\right]. \qquad (4)
\end{aligned}$$

For the first term in (4) 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_X is bounded and μ is Lipschitz, writing $\mathbb{I}N_{-jb}(x) = \mathbb{I}\left\{N_{-jb}(x) \ge 1\right\}$,

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

Now for $t > 0$ we partition by the event $\{\max_{1\le l\le d}|T_b(x)_l| \ge t/\lambda\}$ and apply Lemma 3 and Lemma 8 to see

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

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

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

in the same manner. Hence

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

and so by Chebyshev's inequality,

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

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

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

By Lemma 9, since $N_{-ib}(x) \sim \operatorname{Bin}\left(n-1, \int_{T_b(x)} f_X(s) \, \mathrm{d}s\right)$ given $\mathbf{T}$ and f_X is bounded away from zero,

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

and also

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

So by Lemma 3 and Lemma 7, since f_X is Lipschitz and bounded, using the Cauchy–Schwarz inequality,

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

Next set $A = \frac{1}{f_X(x)|T_b(x)|} \int_{T_b(x)} (f_X(s) - f_X(x)) \, \mathrm{d}s \geq \inf_{s \in [0,1]^d} \frac{f_X(s)}{f_X(x)} - 1$. 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

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

Note that since f_X and μ are Lipschitz and by integrating the tail probability given in Lemma 3, the Maclaurin remainder term is bounded by

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

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

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

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

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

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

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

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

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

Now by integrating the tail probabilities in Lemma 3, $\mathbb{E}\left[\max_{1\le l\le d}|T_b(x)_l|^\beta\right] \lesssim \frac{1}{\lambda^\beta}$. Therefore by Lemma 7, writing $T_b(x)^\nu$ for $\int_{T_b(x)}(s-x)^\nu\,\mathrm{d}s$,

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

Now we show that this is a polynomial in λ. For $1 \leq j \leq d$, define the independent random 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,

$$\begin{aligned} &\mathbb{E}\left[\frac{T_b(x)^{\nu'+\nu''} (T_b(x)^\nu)^u}{|T_b(x)|^{|u|+1}}\right] \\ &\quad = \lambda^{-|\nu'|-|\nu''|-|\nu|\cdot u} \prod_{j=1}^d \mathbb{E}\left[\frac{E_{2j*}^{\nu'_j+\nu''_j+1} + (-1)^{\nu'_j+\nu''_j} E_{1j*}^{\nu'_j+\nu''_j+1}}{(\nu'_j+\nu''_j+1)(E_{2j*}+E_{1j*})} \frac{\left(E_{2j*}^{\nu_j+1} + (-1)^{\nu_j} E_{1j*}^{\nu_j+1}\right)^u}{(\nu_j+1)^u (E_{2j*}+E_{1j*})^{|u|}}\right]. \end{aligned} \tag{5}$$

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,

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

Further, the quantity inside the expectation in (5) is bounded almost surely by one and so the error incurred by replacing E_{1j*} and E_{2j*} and E_{1j} by E_{2j} in (5) is at most $2de^{-C\lambda} \lesssim \lambda^{-\beta}$. Thus the limiting bias is

$$\begin{aligned} \mathbb{E}[\hat{\mu}(x)] - \mu(x) &= \sum_{|\nu'|=1}^{\underline{\beta}} \sum_{|\nu''|=0}^{\underline{\beta}} \sum_{|u|=0}^{\underline{\beta}} \frac{\partial^{\nu'} \mu(x)}{\nu'!} \frac{\partial^{\nu''} f_X(x)}{\nu''!} \left(\frac{\partial^\nu f_X(x)}{\nu!}\right)^u \binom{|u|}{u} \frac{(-1)^{|u|}}{f_X(x)^{|u|+1}} \lambda^{-|\nu'|-|\nu''|-|\nu|\cdot u} \\ &\times \prod_{j=1}^d \mathbb{E}\left[\frac{E_{2j}^{\nu'_j+\nu''_j+1} + (-1)^{\nu'_j+\nu''_j} E_{1j}^{\nu'_j+\nu''_j+1}}{(\nu'_j+\nu''_j+1)(E_{2j}+E_{1j})} \frac{\left(E_{2j}^{\nu_j+1} + (-1)^{\nu_j} E_{1j}^{\nu_j+1}\right)^u}{(\nu_j+1)^u (E_{2j}+E_{1j})^{|u|}}\right] + O\left(\frac{\log n}{\lambda} \frac{\lambda^d}{n}\right) + O\left(\frac{1}{\lambda^\beta}\right), \end{aligned} \tag{6}$$

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_X 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 $\sum_{|\nu|=1}^{\underline{\beta}} \nu_j u_\nu$ is odd.

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

$$\frac{E_{2j}^{\nu'_j+\nu''_j+1} + (-1)^{\nu'_j+\nu''_j} E_{1j}^{\nu'_j+\nu''_j+1}}{E_{2j}+E_{1j}} \tag{7}$$

and

$$\frac{\left(E_{2j}^{\nu_j+1} + (-1)^{\nu_j} E_{1j}^{\nu_j+1}\right)^u}{(E_{2j}+E_{1j})^{|u|}} = \prod_{\substack{|\nu|=1 \\ \nu_j u_\nu \text{ even}}}^{\underline{\beta}} \frac{\left(E_{2j}^{\nu_j+1} + (-1)^{\nu_j} E_{1j}^{\nu_j+1}\right)^{u_\nu}}{(E_{2j}+E_{1j})^{u_\nu}} \prod_{\substack{|\nu|=1 \\ \nu_j u_\nu \text{ odd}}}^{\underline{\beta}} \frac{\left(E_{2j}^{\nu_j+1} + (-1)^{\nu_j} E_{1j}^{\nu_j+1}\right)^{u_\nu}}{(E_{2j}+E_{1j})^{u_\nu}}. \tag{8}$$

Clearly $\nu'_j + \nu''_j$ being odd inverts the sign of (7). For (8), 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 (8) 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 (7) and (8) 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 (7) is preserved. Again the sign of the first product in (8) is preserved, and the sign of every term in (8) 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 (7) and (8) 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 (6) 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 (6) is zero. So we need only consider $\nu'_j = \nu''_j = 1$, giving

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

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

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

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

Hence the second-order bias term is

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

Part 5: Calculating the bias if the data is uniformly distributed
If $X_i \sim \text{Unif}\left([0,1]^d\right)$ then $f_X(x) = 1$ and the bias expansion from (6) becomes

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

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

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

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

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

so

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

□

Proof (Lemma 1)
This follows from the proof of Lemma 2 with $J = 0$. □

Proof (Theorem 3)
By Theorem 2 and Lemma 1,

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

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

Proof (Theorem 4)
By the definition of the debiased estimator and Theorem 2, since J and a_r are fixed,

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

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

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

□

Proof (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 $\mathbb{I}N_{br}(x) = \mathbb{I}\{N_{br}(x) \geq 1\}$. 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,

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

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) $\sum_i S_i(x)^2 \to \Sigma_{\mathrm{J}}(x)$ 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 8,

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

Part 2: checking condition (ii)
Since $\mathbb{E}[\varepsilon_i^2 \mid X_i]$ is bounded and by Lemma 8,

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

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

$$
\begin{aligned}
\sum_{i=1}^n S_i(x)^2 &= \sum_{i=1}^n \left(\sqrt{\frac{n}{\lambda^d}} \sum_{r=0}^J \omega_r \frac{1}{B} \sum_{b=1}^B \frac{\mathbb{I}_{ibr}(x)\varepsilon_i}{N_{br}(x)} \right)^2 \\
&= \frac{n}{\lambda^d} \frac{1}{B^2} \sum_{i=1}^n \sum_{r=0}^J \sum_{r'=0}^J \omega_r \omega_{r'} \sum_{b=1}^B \sum_{b'=1}^B \frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)} \\
&= \frac{n}{\lambda^d} \frac{1}{B^2} \sum_{i=1}^n \sum_{r=0}^J \sum_{r'=0}^J \omega_r \omega_{r'} \sum_{b=1}^B \left(\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ibr'}(x)\varepsilon_i^2}{N_{br}(x)N_{br'}(x)} + \sum_{b' \neq b} \frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)} \right). \qquad (9)
\end{aligned}
$$

By boundedness of $\mathbb{E}[\varepsilon_i^2 \mid X_i]$ and Lemma 8, the first term in (9) converges to zero in probability as

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

For the second term in (9), the law of total variance gives

$$
\begin{aligned}
&\operatorname{Var}\left[\frac{n}{\lambda^d} \frac{1}{B^2} \sum_{i=1}^n \sum_{r=0}^J \sum_{r'=0}^J \omega_r \omega_{r'} \sum_{b=1}^B \sum_{b' \neq b} \frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)} \right] \\
&\quad \leq (J+1)^4 \max_{0 \leq r, r' \leq J} \omega_r \omega_{r'} \operatorname{Var}\left[\frac{n}{\lambda^d} \frac{1}{B^2} \sum_{i=1}^n \sum_{b=1}^B \sum_{b' \neq b} \frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)} \right] \\
&\quad \lesssim \max_{0 \leq r, r' \leq J} \mathbb{E}\left[\operatorname{Var}\left[\frac{n}{\lambda^d} \frac{1}{B^2} \sum_{i=1}^n \sum_{b=1}^B \sum_{b' \neq b} \frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)} \;\Big|\; \mathbf{X}, \mathbf{Y} \right] \right] \\
&\qquad + \max_{0 \leq r, r' \leq J} \operatorname{Var}\left[\mathbb{E}\left[\frac{n}{\lambda^d} \frac{1}{B^2} \sum_{i=1}^n \sum_{b=1}^B \sum_{b' \neq b} \frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)} \;\Big|\; \mathbf{X}, \mathbf{Y} \right] \right] \qquad (10)
\end{aligned}
$$

For the first term in (10),

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

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 8,

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

For the second term in (10), 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 $i \neq j$ 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 $\tilde{\mathbb{I}}_{ijbr}(x) = \mathbb{I}\big\{\tilde{X}_{ij} \in T_{br}(x)\big\}$ and $\tilde{\mathbb{I}}_{jbr}(x) = \mathbb{I}\big\{\tilde{X}_j \in T_{br}(x)\big\}$ and also $\tilde{N}_{jbr}(x) = \sum_{i=1}^n \tilde{\mathbb{I}}_{ijbr}(x)$. We use the leave-one-out notation $N_{-jbr}(x) = \sum_{i \neq j} \mathbb{I}_{ibr}(x)$ and also write $N_{-jbr \cap b'r'} = \sum_{i \neq j} \mathbb{I}_{ibr}(x) \mathbb{I}_{ib'r'}(x)$. Since $\mathbb{E}[\varepsilon_i^4 \mid X_i]$ is bounded,

$$
\begin{aligned}
&\operatorname{Var}\left[\mathbb{E}\left[\frac{n}{\lambda^d} \frac{1}{B^2} \sum_{i=1}^n \sum_{b=1}^B \sum_{b' \neq b} \frac{\mathbb{I}_{ibr}(x) \mathbb{I}_{ib'r'}(x) \varepsilon_i^2}{N_{br}(x) N_{b'r'}(x)} \;\Big|\; \mathbf{X}, \mathbf{Y}\right]\right] \leq \operatorname{Var}\left[\mathbb{E}\left[\frac{n}{\lambda^d} \sum_{i=1}^n \frac{\mathbb{I}_{ibr}(x) \mathbb{I}_{ib'r'}(x) \varepsilon_i^2}{N_{br}(x) N_{b'r'}(x)} \;\Big|\; \mathbf{X}, \mathbf{Y}\right]\right] \\
&\quad \leq \frac{1}{2} \frac{n^2}{\lambda^{2d}} \sum_{j=1}^n \mathbb{E}\left[\left(\sum_{i=1}^n \left(\frac{\mathbb{I}_{ibr}(x) \mathbb{I}_{ib'r}(x) \varepsilon_i^2}{N_{br}(x) N_{b'r'}(x)} - \frac{\tilde{\mathbb{I}}_{ijbr}(x) \tilde{\mathbb{I}}_{ijb'r'}(x) \tilde{\varepsilon}_{ij}^2}{\tilde{N}_{jbr}(x) \tilde{N}_{jb'r'}(x)}\right)\right)^2\right] \\
&\quad \leq \frac{n^2}{\lambda^{2d}} \sum_{j=1}^n \mathbb{E}\left[\left(\left|\frac{1}{N_b(x) N_{b'r'}(x)} - \frac{1}{\tilde{N}_{jbr}(x) \tilde{N}_{jb'r'}(x)}\right| \sum_{i \neq j} \mathbb{I}_{ibr}(x) \mathbb{I}_{ib'r'}(x) \varepsilon_i^2\right)^2\right] \\
&\qquad + \frac{n^2}{\lambda^{2d}} \sum_{j=1}^n \mathbb{E}\left[\left(\left(\frac{\mathbb{I}_{jbr}(x) \mathbb{I}_{jb'r'}(x) \varepsilon_j^2}{N_{br}(x) N_{b'r'}(x)} - \frac{\tilde{\mathbb{I}}_{jbr}(x) \tilde{\mathbb{I}}_{jb'r'}(x) \tilde{\varepsilon}_j^2}{\tilde{N}_{jbr}(x) \tilde{N}_{jb'r'}(x)}\right)\right)^2\right] \\
&\quad \lesssim \frac{n^2}{\lambda^{2d}} \sum_{j=1}^n \mathbb{E}\left[N_{-jbr \cap b'r}(x)^2 \left|\frac{1}{N_{br}(x) N_{b'r'}(x)} - \frac{1}{\tilde{N}_{jbr}(x) \tilde{N}_{jb'r'}(x)}\right|^2 + \frac{\mathbb{I}_{jbr}(x) \mathbb{I}_{jb'r'}(x)}{N_{br}(x)^2 N_{b'r'}(x)^2}\right]. \qquad (11)
\end{aligned}
$$

For the first term in (11), note that

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

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

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

where we applied Lemma 8 in the last line. 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

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

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

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

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 3, we have by Lemma 8 that

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

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 $N_{-ib'r'}(x)$,

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

by a familiar argument based on Lemma 3, the Lipschitz property of $f_X(x)$ and Lemma 8. Hence

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

Now we apply Lemma 9 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 $N_{-ibr\setminus b'r'}(x)$,

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

Now we bound these parameters above and below. Firstly, by applying Lemma 5 with $B = 1$, we have

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

for some $C > 0$ and all sufficiently large t. Next, note that if f_X is L-Lipschitz in ℓ^2, by Lemma 3

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

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

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

So by Lemma 9 conditionally on $\mathbf{T}, N_{-ib'r'\cap br}(x)$ and $N_{-ibr\setminus b'r'}(x)$, with probability at least $1-e^{-t/C}$,

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

Therefore by the same approach as the proof of Lemma 6, taking $t = 3C\log n$,

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

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

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

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

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

Thus far we have proven that

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

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

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

Therefore by Lemma 9 applied conditionally on $\mathbf{T}$, with probability at least $1-e^{-t/C}$,

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

Now by Lemma 7, with $t=3C\log n$,

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

This allows us to deduce that

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

Now 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 $\mathrm{Exp}(1)$. Define their non-truncated versions as

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

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

and note that

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

for some $C > 0$ and sufficiently large λ. Hence by the Cauchy–Schwarz inequality and Lemma 7,

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

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

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

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

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

Secondly, we have in exactly the same manner that

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

Therefore

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

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

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

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

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

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

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

Finally we conclude by giving the limiting variance.

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

So the limit exists and

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

□

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

$$
\hat{\sigma}^2(x) = \frac{1}{B}\sum_{b=1}^{B} \frac{\sum_{i=1}^n Y_i^2 \mathbb{I}\{X_i \in T_b(x)\}}{\sum_{i=1}^n \mathbb{I}\{X_i \in T_b(x)\}} - \hat{\mu}(x)^2. \tag{12}
$$

The first term in (12) 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 (12) 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 (9), 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,

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

□

Proof (Theorem 6)
By Theorem 4 and Lemma 2,

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

The first term now converges weakly to $\mathcal{N}(0,1)$ by Slutsky's theorem, Theorem 5 and Lemma 2, while the second term is $o_{\mathbb{P}}(1)$ by assumption. Validity of the confidence interval follows immediately. □

Proof (Theorem 7)
The bias–variance decomposition along with Theorem 4 and the proof of Theorem 5 with $J=\lfloor\underline{\beta}/2\rfloor$ gives

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

Note that we used an L^2 version of Theorem 4 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 $B \gtrsim n^{\frac{2\beta-2}{d+2\beta}}$, we obtain

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

□

References

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

Cattaneo, M. D., Crump, R. K., and Jansson, M. (2013). Generalized jackknife estimators of weighted average derivatives. *Journal of the American Statistical Association*, 108(504):1243–1256.

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

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press.

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

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

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

Oprescu, M., Syrgkanis, V., and Wu, Z. S. (2019). Orthogonal random forest for causal inference. In *International Conference on Machine Learning*, pages 4932–4941. PMLR.

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

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

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