Electronic Journal of Statistics
Vol. 14 (2020) 3644–3671
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1758

Modeling of time series using random forests: Theoretical developments

Richard A. Davis and Mikkel S. Nielsen*

Department of Statistics, Columbia University, New York, USA
e-mail: rdavis@stat.columbia.edu; m.nielsen@columbia.edu

Abstract: In this paper we study asymptotic properties of random forests within the framework of nonlinear time series modeling. While random forests have been successfully applied in various fields, the theoretical justification has not been considered for their use in a time series setting. Under mild conditions, we prove a uniform concentration inequality for regression trees built on nonlinear autoregressive processes and, subsequently, we use this result to prove consistency for a large class of random forests. The results are supported by various simulations.

MSC2020 subject classifications: Primary 62G05; secondary 60G10, 60J05, 62G08, 62M05, 62M10.
Keywords and phrases: Markov processes, nonlinear autoregressive models, nonparametric regression, random forests.

Received August 2020.

1. Introduction

Random forests, originally introduced by Breiman [9], constitute an ensemble learning algorithm for classification and regression, which produces predictions by first growing a large number of randomized decision trees [10] and, then, aggregates the results. Since its introduction, the algorithm has been applied in various fields such as object recognition [27], bioinformatics [13], ecology [11, 24] and finance [17, 20], and the evidence is strong: with very little tuning, random forests are able to deliver a flexible tool for prediction which is fully comparable with other state-of-the-art algorithms. In fact, Howard and Bowles [19] claim that random forests have been the most successful general-purpose algorithm in recent times. While many successful applications indicate the wide applicability of random forests, only little theoretical work exists to support this impression. Among components, which make the random forests of Breiman [9] difficult to analyze, are the operation of bagging randomized predictors [8] as well as the highly data-dependent partitions associated to the so-called CART regression trees [10], which form the forest. Other types of random forests have been proposed; see, e.g., [2, 16].

While the bagging step is often discarded in theoretical work, or replaced by another resampling method, asymptotic results for random forests in the (classical) nonparametric regression setting, where $(X_1, Y_1), \ldots, (X_T, Y_T)$ are

*Corresponding author.

i.i.d. observations from the model

$$Y = f(X) + \varepsilon, \tag{1.1}$$

have been established under rather weak assumptions on the structure of the underlying regression trees. In (1.1), f is a suitable smooth function and ε is a mean-zero square integrable noise term which is independent of X. To mention a few significant results in this setup, Scornet, Biau and Vert [26] prove L^2 consistency of Breiman's random forests when f is additive (i.e., $f(x) = \sum_{i=1}^{p} f_i(x_i)$) and ε is Gaussian, Wager and Walther [29] establish pointwise consistency of similar forests with larger leaves in a high-dimensional setting, and Wager and Athey [28] prove (pointwise) asymptotic normality of a particular random forest algorithm. Although assumptions are more restrictive, valuable insights about performance (i.e., convergence rates) in sparse settings and lower bounds on mean squared error were provided by [5, 6, 21]. For a nice overview of existing theoretical work on random forests within the regression setting as well as further references, see the survey in Biau and Scornet [7].

In some applications, particularly financial, the underlying data correspond to observations from a time series, and the aim is to predict future values by feeding in a number of the most recent observations to the algorithm. While the problem is often treated precisely as in the regression setting from a practical point of view, by forming pairs $(X_1, Y_1), \ldots, (X_T, Y_T)$ where $X_t = (Y_{t-1}, \ldots, Y_{t-p})$ for some integer $p \geq 1$, things change dramatically on the theoretical side. Indeed, observations can no longer be assumed to be i.i.d. draws from (1.1) and, instead, the entire process $(Y_t)_{t \geq 1}$ is necessarily defined recursively by the equation

$$Y_t = f(X_t) + \varepsilon_t, \qquad t \geq 1, \tag{1.2}$$

given initial data $\xi = (Y_0, Y_{-1}, \ldots, Y_{1-p})$. Processes satisfying (1.2) are often referred to as nonlinear autoregressive processes of order p (or, in short, NLAR(p) processes). For further detail on these processes, see [3, 18]. In such a framework, the dependence structure, across pairs $(X_1, Y_1), \ldots, (X_T, Y_T)$ as well as between entries in X_t, is determined within the model. Consequently, in contrast to the regression setup, it is often only appropriate to impose assumptions on f and $(\varepsilon_t)_{t \geq 1}$. In fact, even if one accepts an implicit model assumption, e.g., the typical assumption that X_t admits a copula density which is bounded away from zero and infinity, it turns out to be rather restrictive. Indeed, if $(Y_t)_{t \geq 1}$ is Gaussian and $p \geq 2$, such an assumption is satisfied only if $f = 0$ almost everywhere. It follows that other types of assumptions and techniques are needed to guarantee the validity of random forests in the time series setting.

In this paper we rely on the principal ideas of [29] to obtain a uniform concentration inequality which applies simultaneously across all regression trees satisfying a mild condition on their minimum leaf size k, when data are generated by the NLAR(p) model (1.2). While it is required that k increases in the sample size, the growth rate may be very slow and trees are allowed to be grown adaptively (the partitions of the trees can be highly data-dependent). As an application of the established concentration inequality, we prove that all random

forests respecting a number of conditions are pointwise consistent estimators for f when the data generating process is (1.2). The assumptions we impose in the model (1.2) are explicit in terms of f and the distribution of $(\varepsilon_t)_{t\geq 1}$, and they are not difficult to check. For instance, all our results are applicable if f is bounded and Lipschitz continuous, and $(\varepsilon_t)_{t\geq 1}$ is an i.i.d. sequence with ε_1 having a suitably light-tailed distribution. (As pointed out in Section 2, the assumption of f being bounded is no stronger than what is usually imposed in the regression setting.) Our techniques rely on, among other things, the theory of Markov processes as well as various Bernstein type concentration inequalities. To the best of our knowledge, theoretical properties of random forests within the framework of time series have not been fully addressed.

The paper is laid out as follows. Section 2 introduces the model as well as the regression trees of interest and establishes uniform concentration of these around their so-called partition-optimal counterparts (Theorem 1). In Section 3 we translate this result into a concentration inequality for random forests (Corollary 1) and provide sufficient conditions ensuring that they are pointwise consistent estimators of f (Theorem 2). Subsequently, we carry out a simulation study in Section 4 which considers the performance of random forests within the NLAR(p) model for various specifications of f. Finally, Section 5 contains proofs of all statements as well as a number of auxiliary results.

2. Concentration of regression trees around partition-optimal counterparts

Let $(\varepsilon_t)_{t\geq 1}$ be a sequence of i.i.d. random variables with $\mathbb{E}[\varepsilon_1] = 0$ and $\mathbb{E}[\varepsilon_1^2] < \infty$, and fix an integer $p \geq 1$. Given a vector $\xi = (Y_0, Y_{-1}, \ldots, Y_{1-p})$ of initial data independent of $(\varepsilon_t)_{t\geq 1}$ and a measurable function $f\colon \mathbb{R}^p \to \mathbb{R}$, define the process $(Y_t)_{t\geq 1}$ recursively by

$$Y_t = f(X_t) + \varepsilon_t, \qquad t \geq 1, \tag{2.1}$$

where $X_t := (Y_{t-1}, \ldots, Y_{t-p})$. In addition to the initial data ξ, suppose that we have T observations $Y_1, \ldots, Y_T$ from the model (2.1) available and that we group them in input-output pairs,

$$\mathcal{D}_T = \{(X_1, Y_1), \ldots, (X_T, Y_T)\}.$$

The aim of this section is to establish uniform concentration inequalities for regression trees built on $\mathcal{D}_T$. We start by recalling the associated concept of recursive partitions [10], which is used to construct regression trees. Define a sequence of partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots$ by starting from $\mathcal{P}_1 = \{\mathbb{R}^p\}$ and then, for each $n \geq 1$, construct $\mathcal{P}_{n+1}$ from $\mathcal{P}_n$ by replacing one set (node) $A \in \mathcal{P}_n$ by $A_L := \{x \in A : x_i \leq \tau\}$ and $A_R := \{x \in A : x_i > \tau\}$, where the split direction $i \in \{1, \ldots, p\}$ and split position $\tau \in \{x_i : x \in A\}$ are chosen in accordance with some set of rules. Here x_i refers to the i-th entry of $x \in \mathbb{R}^p$. In this context, we will say that A is the *parent* node of A_L and A_R, while A_L and A_R are

the *child* nodes of A. A given partition Λ of $\mathbb{R}^p$ is called recursive if $\Lambda = \mathcal{P}_n$ for some $n \geq 1$, where $\mathcal{P}_1, \ldots, \mathcal{P}_n$ are obtained as above. Note that the rules determining how to choose node, direction and position of a split may depend on the data $\mathcal{D}_T$ as well as some injected randomness Θ. For instance, in Breiman's random forests a node is split as soon as it contains at least a certain number of observations, while the position and direction are determined by maximizing impurity decrease (or, equivalently, minimizing the total mean-corrected sum of squares of the outputs Y over the resulting two child nodes; see also [10]), but only over a randomly chosen subset of directions in $\{1, \ldots, p\}$. To any recursive partition Λ we define the corresponding regression tree T_Λ by

$$T_\Lambda(x) = \frac{1}{|\{t \in \{1, \ldots, T\} : X_t \in A_\Lambda(x)\}|} \sum_{t=1}^{T} Y_t \mathbb{1}_{A_\Lambda(x)}(X_t), \qquad x \in \mathbb{R}^p. \quad (2.2)$$

Here the notation $A_\Lambda(x)$ is used to refer to the unique set $A \in \Lambda$ with the property that $x \in A$. Our interest will be on regression trees defined by k-valid partitions ($k \geq 1$). We will say that a partition Λ is k-valid, and write $\Lambda \in \mathcal{V}_k$, if Λ is recursive and each set in Λ (sometimes called a leaf of the corresponding tree T_Λ) contains at least k data points. Note that, since Λ is recursive, it can depend on both the data $\mathcal{D}_T$ and a random mechanism Θ, while $\mathcal{V}_k$ depends only on $\mathcal{D}_T$. Setting a minimum number k of observations in each leaf of a tree is default in most practical implementations of random forests. Besides, such an assumption is natural since it ensures that $X_t \in A_\Lambda(x)$ for some $t \in \{1, \ldots, T\}$, and this implies that the regression tree (2.2) is well-defined for all $x \in \mathbb{R}^p$. In this section we will be working under the following set of assumptions:

(A1) The random variable ε_1 admits a density h_ε which is positive almost everywhere on $\mathbb{R}$ and, for some $c \in (0, \infty)$,

$$\mathbb{E}[|\varepsilon_1|^m] \leq m! c^{m-2}, \qquad m = 3, 4, \ldots \quad (2.3)$$

Moreover, the cumulative distribution function $F_\varepsilon(x) = \int_{-\infty}^{x} h_\varepsilon(y)\, \mathrm{d}y$ of ε_1 satisfies

$$\sup_{x \in \mathbb{R}} \frac{F_\varepsilon(x + \tau)}{F_\varepsilon(x)} < \infty \quad (2.4)$$

for any $\tau \in (0, \infty)$.

(A2) The function f in (2.1) is bounded,

$$M \coloneqq \sup_{x \in \mathbb{R}^p} |f(x)| < \infty. \quad (2.5)$$

(A3) The minimum number k of data points in each leaf satisfies $k/(\log T)^4 \to \infty$ as $T \to \infty$.

In contrast to k, the quantities c, M and p will be kept fixed, and hence we will not keep track of the dependence on these in the following results. In particular, the introduced constants can depend on c, M and p, but not on T and k. When ε_1 admits a density which is positive almost everywhere and f

satisfies (2.5) (in particular, when (A1) and (A2) are imposed), it follows by [3, Theorem 3.1] that the distribution of ξ can be chosen such that $(Y_t)_{t\geq 1}$ is strictly stationary and, thus, this will be assumed throughout the paper. This means that $(X_1, Y_1), \ldots, (X_T, Y_T)$ are identically distributed. Before turning to the results, let us attach some comments to the assumptions stated above. The assumption of (A1) that ε_1 has a positive density is convenient, since it ensures that the p-th order Markov chain $(Y_t)_{t\geq 1}$ can reach any state in one time step. In addition to strict stationarity of the chain, when combined with (A2), the assumption ensures geometrical ergodicity as well. While it is not required that f is bounded to prove such properties of $(Y_t)_{t\geq 1}$, we need boundedness to apply Bernstein type inequalities for strongly mixing processes and to obtain good estimates on the dependency between entries of the input vector X_1. The boundedness assumption (2.5) is implicitly assumed in essentially all theoretical work on random forests as one usually assumes that the input vector is transformed so that it belongs to the unit cube $[0, 1]^p$ and then requires continuity of f on this domain. The assumption on the moments of ε_1 in (2.3) implies that ε_1 is sub-exponential in the sense that

$$\mathbb{P}(|\varepsilon_1| > x) \leq \gamma_1 e^{-\gamma_2 x}, \qquad x > 0, \tag{2.6}$$

for suitably chosen $\gamma_1, \gamma_2 \in (0, \infty)$. It is a well-known assumption to impose when proving concentration inequalities and is often needed when ε_1 cannot be assumed bounded. Among distributions satisfying (2.3) are (sub-)Gaussian distributions, but also those with a slightly heavier tail such as the Laplace distribution. The assumption (2.4) is used in conjunction with (2.5) to estimate probabilities involving the input vector X_1 (see Lemma 1 for details). Ultimately, it is an assumption on the left tail of ε_1, and a sufficient condition for this to hold is that the limit

$$\lim_{x\to-\infty} \frac{h_\varepsilon(x)}{h_\varepsilon(x+\tau)}$$

exists and is non-zero for all $\tau \in (0, \infty)$. It is straightforward to verify that this, as well, is satisfied for both Gaussian and Laplace distributions. Together with (2.5), (2.4) ensures that we do not need to impose conditions on the copula density of the input vector X_1, as is usually done in the regression setting, and this is convenient since such conditions can be both difficult to verify and even rather restrictive in a time series setting. Finally, we impose (A3), which in particular implies that $k \to \infty$ as $T \to \infty$. Although it is allowed that $k \to \infty$ at a slow rate, the assumption contrasts the trees used in the random forests of Breiman [9], where k is some fixed and often small number. On the other hand, (A3) is very similar to assumptions imposed in most theoretical work within the regression setting (see, e.g., [5, 26, 29]). In fact, to the best of our knowledge, the only asymptotic result for random forests built on trees with fixed k is [26, Theorem 2]. The logarithmic factor $(\log T)^4$ is related to the fact that the established bound applies uniformly across all trees (see Remark 1) and that we use a Bernstein type inequality for strongly mixing processes which is slightly weaker than the classical one for the independent case.

While a couple of additional assumptions are needed to establish consistency of random forests in Section 3, (A1)–(A3) are sufficient to prove that regression trees of the form (2.2) concentrate around their so-called partition-optimal counterparts

$$T^*_\Lambda(x) \coloneqq \mathbb{E}_\Lambda[Y \mid X \in A_\Lambda(x)] \tag{2.7}$$

uniformly across $(x, \Lambda) \in \mathbb{R}^p \times \mathcal{V}_k$. Here (X, Y) is a copy of (X_1, Y_1) which is independent of $(\mathcal{D}_T, \Theta)$, and $\mathbb{E}_\Lambda$ denotes expectation with respect to the conditional probability measure $\mathbb{P}_\Lambda \coloneqq \mathbb{P}(\cdot \mid \mathcal{D}_T, \Theta)$. Conditional on $(\mathcal{D}_T, \Theta)$, the set $A_\Lambda(x)$ is non-random and, hence, the right-hand side of (2.7) simply means that the map $A \mapsto \mathbb{E}[Y \mid X \in A]$ is evaluated at $A_\Lambda(x)$. Our setting is very similar to that of Wager and Walther [29], but besides requiring partitions to be k-valid they impose an additional assumption that excludes too "unbalanced" splits (see also the trees constructed in Section 3).

Theorem 1. *Suppose that (A1)–(A3) are satisfied. Then there exists a constant* $\beta \in (0, \infty)$ *such that*

$$\sup_{(x,\Lambda)\in\mathbb{R}^p\times\mathcal{V}_k} |T_\Lambda(x) - T^*_\Lambda(x)| \leq \beta \frac{(\log T)^2}{\sqrt{k}} \tag{2.8}$$

with probability at least $1 - 4T^{-1}$ *for all sufficiently large* T.

Remark 1. For any given pair $(x, \Lambda) \in \mathbb{R}^p \times \mathcal{V}_k$, the quantity $|T_\Lambda(x) - T^*_\Lambda(x)|$ is the deviation of the sample average over at least k observations from its theoretical counterpart within a specific leaf L. Some of the leaves, which can be obtained by varying (x, Λ), contain only k observations and for these, the error is of order $1/\sqrt{k}$. This is almost the same upper bound as in (2.8) apart from the logarithmic factor $(\log T)^2$, which reflects the fact that the deviation is controlled simultaneously across all feasible pairs (x, Λ) as well as the sub-exponential tail of ε_1.

Remark 2. In Theorem 1, and the remaining results of this paper, it is assumed that one is able to select a suitable $p \geq 1$ such that (2.1) is correctly specified. If it is not possible to identify such p, one may consider a sequence of models (indexed by T) where p increases as more data become available. Eventually, if $(Y_t)_{t\geq 1}$ is an NLAR(p^*) process for some $p^* \geq 1$, this will ensure that the model is correctly specified for large samples. Under suitable assumptions, Theorem 1 can in fact be adjusted to allow for such setting by adapting the ideas of [29] and keeping track of how constants depend on p. However, the resulting upper bound on the uniform deviation of regression trees from their partition-optimal counterparts seems to be rather sensitive to the value of p and, thus, effectively demands that p increases very slowly in T. While this, at first glance, contrasts the results of [29] where the upper bound depends on p only through a $\sqrt{\log p}$ factor, they impose an assumption on the density of X which roughly means that the constant $\zeta \in (0, \infty)$ in (5.2) does not depend on p. Such an assumption, however, appears to be rather restrictive in a time series setting.

3. Concentration and consistency of forests

We start by translating the concentration inequality of Theorem 1 into the framework of random forests, which are constructed by averaging a number of trees. To this end, let $\mathcal{W}_k \coloneqq \{\mathbf{\Lambda} \subseteq \mathcal{V}_k : |\mathbf{\Lambda}| < \infty\}$ be the family of all finite collections of k-valid partitions. In line with Wager and Walther [29], given an element $\mathbf{\Lambda} = \{\Lambda_1, \ldots, \Lambda_B\}$ of $\mathcal{W}_k$, the corresponding k-valid random forest $H_{\mathbf{\Lambda}}$ is defined as

$$H_{\mathbf{\Lambda}}(x) = \frac{1}{B}\sum_{b=1}^{B} T_{\Lambda_b}(x), \qquad x \in \mathbb{R}^p. \tag{3.1}$$

The associated partition-optimal forest $H^*_{\mathbf{\Lambda}}$ is given by

$$H^*_{\mathbf{\Lambda}}(x) = \frac{1}{B}\sum_{b=1}^{B} T^*_{\Lambda_b}(x), \qquad x \in \mathbb{R}^p.$$

As an immediate consequence of Theorem 1, we obtain the following concentration inequality which applies uniformly across all k-valid forests (the result is stated without proof):

Corollary 1. *Suppose that (A1)–(A3) are satisfied. Then there exists a constant* $\beta \in (0, \infty)$ *such that*

$$\sup_{(x,\mathbf{\Lambda}) \in \mathbb{R}^p \times \mathcal{W}_k} |H_{\mathbf{\Lambda}}(x) - H^*_{\mathbf{\Lambda}}(x)| \leq \beta \frac{(\log T)^2}{\sqrt{k}}$$

with probability at least $1 - 4T^{-1}$ *for all sufficiently large* T.

Note that all trees $T_{\Lambda_1}, \ldots, T_{\Lambda_B}$ in (3.1) are based on the same data set $\mathcal{D}_T$ (the partitions $\Lambda_1, \ldots \Lambda_B$ as well as the averages within the relevant leaves $A_{\Lambda_1}(x), \ldots, A_{\Lambda_B}(x)$ depend on $\mathcal{D}_T$). In contrast, in the random forests of Breiman [9], an initial bootstrap step is performed before growing each tree, meaning that trees are built on a bootstrap sample from $\mathcal{D}_T$ (with replacement) rather than on $\mathcal{D}_T$ itself. Once we have a concentration inequality as in Theorem 1 (or Corollary 1) at our disposal, it is not difficult to design trees in such a way that the corresponding random forests are consistent estimators of f. Roughly speaking, given that f is smooth, and since each tree in a forest is close to its partition-optimal counterpart with high probability, it is sufficient to design the recursive partitioning scheme such that the maximal diameter of each leaf shrinks to zero as T becomes large. Below we demonstrate how to refine the collection of k-valid partitions $\mathcal{V}_k$ in a suitable way and, subsequently, prove consistency of the corresponding forests. The construction will be similar to those of [22, 28, 29]. We emphasize that the refinement considered here does not result in one particular random forest estimator; rather, a number of rules is outlined, and these will ensure consistency of any random forest estimator, which is built in line with them. For $\alpha \in (0, 1/2)$, $k \geq 1$ and $m \geq 2k$, we now define (α, k, m)-valid partitions, $\mathcal{V}_{\alpha,k,m}$. The first requirement for a partition Λ

to belong to $\mathcal{V}_{\alpha,k,m}$ is that it is recursive, meaning that $\Lambda = \mathcal{P}_n$ for some $n \geq 1$ where $\mathcal{P}_1, \ldots, \mathcal{P}_n$ are obtained as in Section 2. The second, and last, requirement is that the associated partitioning scheme used to obtain $\mathcal{P}_1, \ldots, \mathcal{P}_n$ obeys the following rules:

(i) Any currently unsplit node with at least m data points will be split.
(ii) The probability $\rho_i = \rho_i(\mathcal{D}_T)$ that a given (feasible) node is split along the i-th direction is bounded from below for all $i = 1, \ldots, p$ by a strictly positive constant.
(iii) The split position is chosen such that each child node contains at least a fraction $\alpha \in (0, 1/2)$ of the data points in its parent node.
(iv) All leaves of the tree contain at least k data points.

The corresponding (α, k, m)-valid forest is given by (3.1) with $\Lambda_1, \ldots, \Lambda_B \in \mathcal{V}_{\alpha,k,m}$. Let us now briefly address the rules outlined in (i)–(iv). Clearly, (iv) ensures $\mathcal{V}_{\alpha,k,m} \subseteq \mathcal{V}_k$, and thus (α, k, m)-valid forests form a subclass of k-valid forests. Rule (i) controls the maximal number of observations in each leaf of a tree, and $m = 2k$ corresponds to a situation where one keeps splitting until placing another split would violate (iv). In general, if m is not too large relative to T, this condition ensures that the number of leaves becomes large and, hence, the partition becomes fine. Concerning (ii), it ensures that, eventually, a split will be placed along any of the p (canonical) directions of the input space $\mathbb{R}^p$. Such a condition makes sense for us when p is thought of as being fixed and rather small, but will not be reasonable in sparse settings where $p \to \infty$, and one will instead design the algorithm in a way that detects important directions with high probability. On the other hand, ρ_i is indeed allowed to depend on $\mathcal{D}_T$, so one may use the data to identify which of the directions that are most important and then, based on this, form the probabilities $\rho_1, \ldots, \rho_p$. In a time series setting, it may be advantageous to favor splits along the first direction which corresponds to an observation that is likely to be highly dependent with the observed value of Y. Finally, (iii) is a balancing condition which prohibits "edge splits". This is a technical condition imposed to track the distribution of data points among leaves. In theoretical work on random forests within the regression setting, it is typical to impose assumptions similar to (i)–(iv), see [22, 28, 29]. On the other hand, standard implementations, such as the RandomForestRegressor from the **sklearn** library in Python and the **ranger** package in R, incorporate only (i), (ii) and (iv).

Since consistency will be established by relying on Theorem 1, we require that (A1)–(A3) are satisfied. Moreover, the following assumptions are imposed:

(A4) The function f in (2.1) is C-Lipschitz, i.e.,

$$|f(x') - f(x)| \leq C\|x' - x\| \qquad \text{for all } x, x' \in \mathbb{R}^p$$

with $C \in (0, \infty)$ being a suitable constant and $\|\cdot\|$ some norm on $\mathbb{R}^p$.
(A5) It holds that $\log(T/m)/\log(\alpha^{-1}) \to \infty$ as $T \to \infty$.

With assumptions (A1)–(A5) in hand, we can now state the following consistency result for (α, k, m)-forests applied to nonlinear autoregressive processes:

Theorem 2. *Let $\hat{f}_T$ be an (α, k, m)-forest and suppose that (A1)–(A5) are satisfied. Then the following statements hold:*

(a) $\hat{f}_T$ is a pointwise consistent estimator of f in the sense that

$$\hat{f}_T(x) \longrightarrow f(x) \qquad \text{in probability as } T \to \infty.$$

for any $x \in \mathbb{R}^p$.

(b) $\hat{f}_T(X)$ is a consistent estimator of the conditional mean $\mathbb{E}[Y \mid X]$ in the sense that

$$\hat{f}_T(X) \longrightarrow \mathbb{E}[Y \mid X] \qquad \text{in probability as } T \to \infty.$$

Remark 3. It should be emphasized that, since consistency is obtained through Theorem 1, the averaging effect gained by considering (3.1) rather than a single tree is not exploited in this setting. In particular, for the regression trees to concentrate around their partition-optimal counterparts, the number of observations in each leaf is required to approach infinity as T becomes large (cf. (A3)). If this is not the case, averages within leaves do not converge, meaning that individual trees will be inconsistent estimators for f. In this case, consistency of $\hat{f}_T$ must be caused by improved accuracy gained by averaging trees. It is generally recognized that random forests deliver much better performance than single trees in practice, and it has also been shown theoretically (in simplified settings) that both the bias and variance are smaller, see [4, 15].

4. A simulation study

In this section we consider a number of different specifications of f in (2.1) and illustrate through simulations the results of Theorem 2. In all examples, the distribution of ε_1 is assumed to have a standard Laplace distribution, so that $h_\varepsilon(x) = \frac{1}{2}e^{-|x|}$ for $x \in \mathbb{R}$. As already mentioned, this choice meets the conditions imposed in (A1). To keep things simple, we consider initially $p = 1$ so that f is one-dimensional and $(Y_t)_{t\geq 1}$ is a first order Markov chain. Within this setting, we choose four different specifications of f, namely

$$\begin{aligned} &f(x) = 0.5\,\mathrm{sign}(x)\min\{|x|, 10\}, && f(x) = -2xe^{-0.7x^2} + 3x^2e^{-0.95x^2}, \\ &f(x) = \cos(5x)e^{-x^2}, && \text{and} \quad f(x) = \min\{|x|, 0.75\}\min\{|x|, 10\}. \end{aligned} \tag{4.1}$$

The first specification of f satisfies $f(x) = 0.5x$ when $x \in [-10, 10]$, and is constant outside of $[-10, 10]$, and hence the corresponding process $(Y_t)_{t\geq 1}$ is intended to mimic the classical linear AR(1) process. Indeed, it is very unlikely that $|Y_t|$ exceeds 10, which means that there is only little practical difference between the two processes. The second specification is an example of an expo-

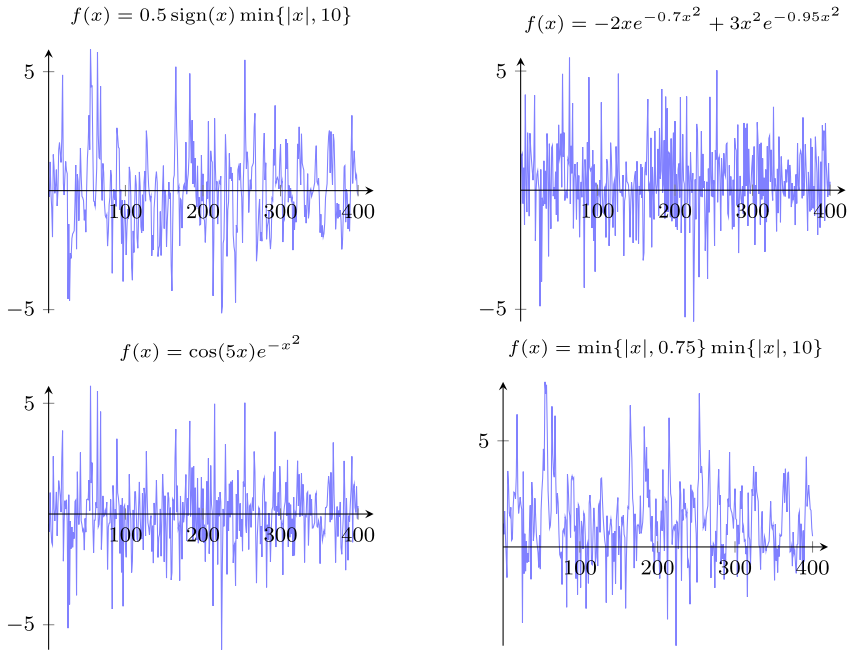

FIG 1. *Simulations of* $Y_1, \ldots, Y_{400}$ *from the model* (2.1) *for the four different specifications of* f *considered in* (4.1).

nential AR model (see, e.g., [3]), while the last two specifications of f correspond to an oscillating function and a particular spline, respectively. In Figure 1, we have simulated a sample path $Y_1, \ldots, Y_{400}$ for each of these specifications of f.

We consider estimation of f by a random forest $\hat{f}_T$ across different sample sizes T and we will be using the **ranger** package of R with $B = 500$ and $k = \lfloor 0.04(\log T)^4 \log\log T \rfloor$. To obtain diverse trees, we will use the extremely randomized trees of Geurts, Ernst and Wehenkel [16] which corresponds to setting the parameters `replace` = FALSE, `sample.fraction` = 1 and `splitrule` = "extratrees". Effectively, this means that split positions are chosen at random and that we build each tree using the entire sample $\mathcal{D}_T$ (no initial bootstrap step). Note that, while this implementation aligns with the (α, k, m)-valid forests treated in Section 3, α is not a prespecified parameter in the **ranger** package, yet in principle its value can be implicitly determined. In Figure 2 we compare $\hat{f}_T$ to f on the interval $[-2, 2]$ for each of the four different examples of f presented in (4.1). While the plots indicate the consistency of the random forest estimator in these examples (as should be the case), observations are in fact rather noisy, and hence the performance of the random forest is indeed remarkable. To support this, Figure 3 shows scatter plots of the data $\mathcal{D}_T = \{(Y_0, Y_1), \ldots, (Y_{T-1}, Y_T)\}$ for $T = 400$ and two specifications of f. Furthermore, we note that choosing the parameter k in finite samples is not a

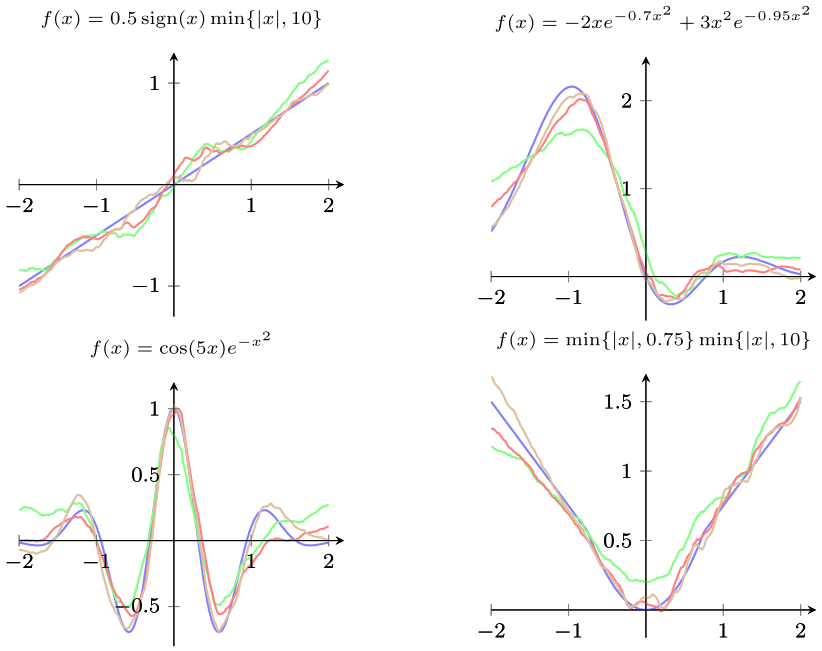

FIG 2. *The four specifications of f considered in (4.1) (blue) and the corresponding random forest estimator $\hat{f}_T$ based on sample sizes of $T = 400$ (green), $T = 1600$ (red) and $T = 6400$ (brown).*

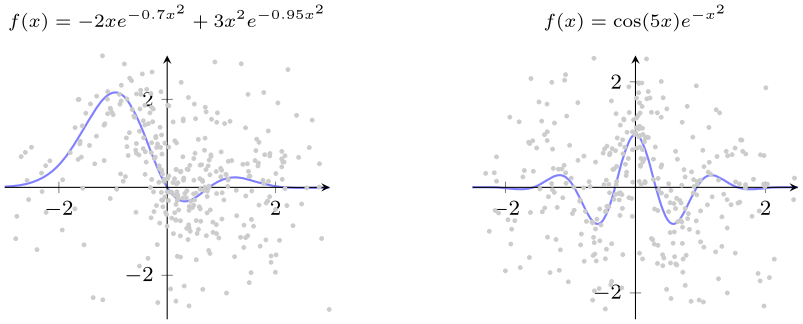

FIG 3. *Scatter plots of the data $\mathcal{D}_{400}$ under two of the specifications of f considered in (4.1).*

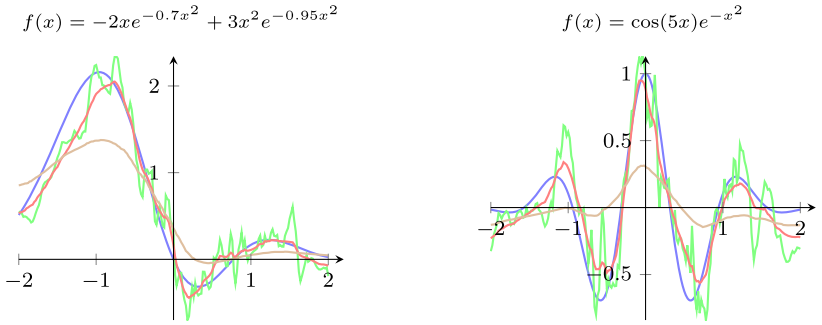

FIG 4. *Two of the specifications of f considered in (4.1) (blue) and the corresponding random forest estimator $\hat{f}_{1600}$ with $k = 40$ (green), $k = 160$ (red) and $k = 640$ (brown).*

trivial task, and the choice used above is rather arbitrary (the assumption of (A3) concerns only its asymptotic behavior). Nevertheless, its value can have a significant impact on performance as it controls the bias–variance tradeoff of the estimator. While optimal tuning of k is outside the scope of this paper, we illustrate its effect on $\hat{f}_T$ in Figure 4 where we estimate two of the functions in (4.1) for different values of k using a sample of size $T = 1600$. For comparison, the value used for k in Figure 2 when $T = 1600$ was 236.

We conclude this section by indicating consistency of random forests in a more challenging setting. In particular, we consider $p = 2$ and the following choice of f:

$$f(x_1, x_2) = x_1 e^{-0.6x_1^2} - 2(x_1^2 e^{-0.3x_1^2} + x_2 e^{-0.7x_2^2}) + 3x_2^2 e^{-0.95x_2^2}. \tag{4.2}$$

We rely on the **ranger** package again with the same specifications as were used to obtain Figure 2, but we set the additional parameter `split.select.weights` $= (1/2, 1/2)$ so that the probability of splitting along a given direction is the same for both directions (i.e., $\rho_1 = \rho_2 = 1/2$). To evaluate its performance, we compute the mean squared error

$$\mathrm{MSE} = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} (\hat{f}_T(x) - f(x))^2 \tag{4.3}$$

over the grid $\mathcal{X} := \{-2, -1.75, \ldots, 1.75, 2\}^2$ for different values of T. In Figure 5, the MSE is depicted as a function of $10^{-4}T$.

5. Proofs

It will be convenient to transform the input vector $X_t = (Y_{t-1}, \ldots, Y_{t-p})$ so that it takes values in $[0, 1]^p$. Effectively, this can be done by applying a cumulative distribution function

$$F_h(x) = \int_{-\infty}^{x} h(y)\, \mathrm{d}y, \qquad x \in \mathbb{R}, \tag{5.1}$$

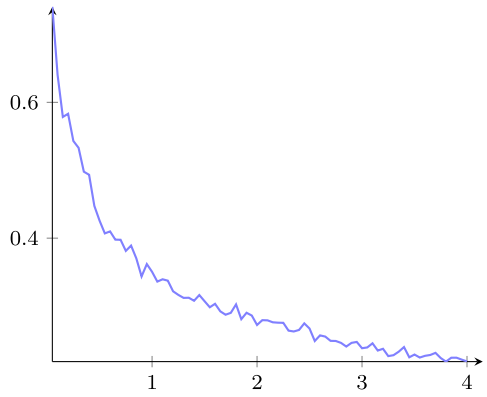

FIG 5. *The mean squared error* (4.3) *of the random forest estimator* $\hat{f}_T$ *as a function of* $10^{-4}T$ *when* f *is given by* (4.2).

with $h\colon \mathbb{R} \to [0,\infty)$ being a probability density which is strictly positive almost everywhere. We extend the domain of F_h to $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ by using the conventions $F_h(-\infty) = 0$ and $F_h(\infty) = 1$, so that the mapping

$$\iota_h\colon (x_1,\dots,x_p) \longmapsto (F_h(x_1),\dots,F_h(x_p))$$

is one-to-one between $\overline{\mathbb{R}}^p$ and $[0,1]^p$. The transformed input vector is defined by $Z_t = \iota_h(X_t)$. While there are no further restrictions on the choice of h, we will pick one that leads to good estimates on the density h_Z of Z_1.

Lemma 1. *Suppose that (A1) and (A2) hold. Then there exists a constant* $\zeta \in (1,\infty)$ *and a probability density* $h\colon \mathbb{R} \to [0,\infty)$*, which is strictly positive almost everywhere, such that the density* $h_Z\colon [0,1]^p \to [0,\infty)$ *of* $Z_1 = \iota_h(X_1)$ *satisfies*

$$\zeta^{-1} \le h_Z(z) \le \zeta \tag{5.2}$$

for almost all $z \in [0,1]^p$.

Proof. By (2.4) in (A1) it holds that

$$\bar{\zeta} := \sup_{x\in\mathbb{R}} \frac{F_\varepsilon(x+M)}{F_\varepsilon(x-M)} \in (1,\infty). \tag{5.3}$$

It follows as well from (A1) that ε_1 admits a density h_ε which is strictly positive almost everywhere, and hence

$$h(y) := \frac{1-\bar{\zeta}^{-1}}{\bar{\zeta}-\bar{\zeta}^{-1}} h_\varepsilon(y+M) + \frac{\bar{\zeta}-1}{\bar{\zeta}-\bar{\zeta}^{-1}} h_\varepsilon(y-M), \qquad y \in \mathbb{R},$$

is a valid density to use for defining $Z_t = \iota_h(X_t)$. To show that h_Z meets (5.2), it suffices to establish that

$$\zeta^{-1} \prod_{i=1}^p z_i \le \mathbb{P}(F_h(Y_1) \le z_1,\dots,F_h(Y_p) \le z_p) \le \zeta \prod_{i=1}^p z_i \tag{5.4}$$

for all $z_1, \dots, z_p \in [0,1]$, where F_h is the cumulative distribution function defined by (5.1) and $\zeta = \bar{\zeta}^p$. Since $\varepsilon_i - M \le Y_i \le \varepsilon_i + M$ by (A2), it follows immediately from the independence of $\varepsilon_1, \dots, \varepsilon_p$ and the monotonicity of F_h that

$$\prod_{i=1}^{p} F_\varepsilon(F_h^{-1}(z_i) - M) \le \mathbb{P}(F_h(Y_1) \le z_1, \dots, F_h(Y_p) \le z_p) \le \prod_{i=1}^{p} F_\varepsilon(F_h^{-1}(z_i) + M).$$

Consequently, we only need to show that both $F_\varepsilon(F_h^{-1}(z) + M) \le \bar{\zeta} z$ and $F_\varepsilon(F_h^{-1}(z) - M) \ge \bar{\zeta}^{-1} z$ for an arbitrary $z \in [0,1]$. Observe that, by (5.3) and the definition of h, $\bar{\zeta}^{-1} F_\varepsilon(x + M) \le F_h(x) \le \bar{\zeta} F_\varepsilon(x - M)$ for all $x \in \overline{\mathbb{R}}$. In particular, by choosing $x = F_h(z)^{-1}$ we obtain

$$\bar{\zeta}^{-1} F_\varepsilon(F_h^{-1}(z) + M) \le z \le \bar{\zeta} F_\varepsilon(F_h^{-1}(z) - M),$$

and this completes the proof. □

In all of the following, $Z_t = \iota_h(X_t)$ for some h such that (5.2) holds, and we will be using the notation $\#R := |\{t \in \{1, \dots T\} : Z_t \in R\}|$, $\mu(R) := \mathbb{P}(Z_1 \in R)$, and $\eta(R) := \mathbb{E}[Y_1 \mid Z_1 \in R]$ for any given measurable set $R \subseteq [0,1]^p$. Note that if $\Lambda \in \mathcal{V}_k$, the partition $\bar{\Lambda}$ of $\overline{\mathbb{R}}^p$ obtained by the exact same sequence of consecutive splits is again a k-valid partition and

$$T_\Lambda(x) - T_\Lambda^*(x) = T_{\bar{\Lambda}}(x) - T_{\bar{\Lambda}}^*(x) \tag{5.5}$$

for all $x \in \mathbb{R}^p$. Moreover, for any $x \in \mathbb{R}^p$ we have that

$$T_{\bar{\Lambda}}(x) - T_{\bar{\Lambda}}^*(x) = \frac{1}{\#L} \sum_{t:\ Z_t \in L} Y_t - \eta(L) =: G_T(L), \tag{5.6}$$

where $L = \iota_h(A_{\bar{\Lambda}}(x))$. Here, and in what follows, it is implicitly understood that a sum of the form $\sum_{t:\ Z_t \in R}$ runs over $\{t \in \{1, \dots, T\} : Z_t \in R\}$. Since A is a leaf of a k-valid partition of $\overline{\mathbb{R}}^p$ with respect to $(X_1, \dots, X_T)$ if and only if $\iota_h(A)$ is a leaf of a k-valid partition of $[0,1]^p$ with respect to $(Z_1, \dots, Z_T)$, it follows from (5.5) and (5.6) that

$$\sup_{L \in \mathcal{L}_k} |G_T(L)| = \sup_{(x,\Lambda) \in \mathbb{R}^p \times \mathcal{V}_k} |T_\Lambda(x) - T_\Lambda^*(x)|, \tag{5.7}$$

where $\mathcal{L}_k$ consists of all sets which are members of k-valid partitions of $[0,1]^p$. In particular, it suffices to prove uniform concentration inequalities for empirical averages over rectangles in $\mathcal{L}_k$. Still, there are infinitely many rectangles in $\mathcal{L}_k$, so one cannot simply analyze $|G_T(L)|$ and then rely on a union bound. We will follow the ideas of Wager and Walther [29] who demonstrated that one only needs to understand the concentration over a much smaller set of approximating rectangles. In particular, we will make use of one of their results which states that there exists a rather small collection of rectangles in $[0,1]^p$ containing good approximations to any non-negligible rectangle in terms of Lebesgue measure. Since their result is more general than what is needed here (e.g., it can be used in

situations where $p \to \infty$), we state a rather simplified version in Theorem 3 below. To avoid introducing too many non-informative constants in the following, we introduce some convenient notation. For two sequences $(a_t)_{t\geq 1}$ and $(b_t)_{t\geq 1}$ we will write $a_t \lesssim b_t$ if there exists a constant $c \geq 1$ such that $a_t \leq cb_t$ for all t. If both $a_t \lesssim b_t$ and $b_t \lesssim a_t$ we write $a_t \asymp b_t$.

Theorem 3 (Wager and Walther [29]). *Let $\varepsilon \asymp k^{-1/2}$ and $w \asymp k/T$. Then there exists a collection of rectangles $\mathcal{R}_{\varepsilon,w}$ with the following two properties:*

(i) For any rectangle $R \subseteq [0,1]^p$ with $\mathrm{Leb}(R) \geq w$, one can find $R_-, R_+ \in \mathcal{R}_{\varepsilon,w}$ satisfying

$$R_- \subseteq R \subseteq R_+ \quad \text{and} \quad e^{-\varepsilon}\,\mathrm{Leb}(R_+) \leq \mathrm{Leb}(R) \leq e^{\varepsilon}\,\mathrm{Leb}(R_-). \tag{5.8}$$

(ii) The cardinality $|\mathcal{R}_{\varepsilon,w}|$ of $\mathcal{R}_{\varepsilon,w}$ satisfies the bound $\log|\mathcal{R}_{\varepsilon,w}| \lesssim \log T$.

Let $\varepsilon, w \in (0,1)$ be given as in Theorem 3. It follows that any given leaf $L \in \mathcal{L}_k^w := \{L \in \mathcal{L}_k : \mathrm{Leb}(L) \geq w\}$ can be inner ε-approximated by a rectangle L_-^ε from $\mathcal{R}_{\varepsilon,w}$ in the sense of (5.8). Moreover,

$$\begin{aligned}\sup_{L\in\mathcal{L}_k^w}|G_T(L)| \leq{}& \sup_{L\in\mathcal{L}_k^w}|\eta(L_-^\varepsilon)-\eta(L)| + \sup_{L\in\mathcal{L}_k^w}|G_T(L_-^\varepsilon)| \\ &+ \sup_{L\in\mathcal{L}_k^w}\Big|\frac{1}{\#L}\sum_{t:\,Z_t\in L} Y_t - \frac{1}{\#L_-^\varepsilon}\sum_{t:\,Z_t\in L_-^\varepsilon} Y_t\Big|.\end{aligned} \tag{5.9}$$

Thus, to obtain a concentration inequality for (5.7) it suffices to show that, for all large T and with high probability, the three terms on the right-hand side of the inequality (5.9) are small and $\mathcal{L}_k = \mathcal{L}_k^w$. Bounding the first term of (5.9) is the easiest task.

Lemma 2. *Suppose that (A1) and (A2) are satisfied, and let $\varepsilon \asymp k^{-1/2}$ and $w \asymp k/T$. Then*

$$\sup_{L\in\mathcal{L}_k^w}|\eta(L_-^\varepsilon)-\eta(L)| \leq 2M\zeta^2\varepsilon,$$

where $\zeta \in (1,\infty)$ is given as in Lemma 1.

Proof. By (A2), we find for an arbitrary leaf $L \in \mathcal{L}_k^w$ that

$$\begin{aligned}|\eta(L_-^\varepsilon)-\eta(L)| &\leq \frac{1}{\mu(L)}\int_{L\setminus L_-^\varepsilon}|f(\iota_h^{-1}(z))|h_Z(z)\,\mathrm{d}z \\ &\quad + \frac{\mu(L)-\mu(L_-^\varepsilon)}{\mu(L)\mu(L_-^\varepsilon)}\int_{L_-^\varepsilon}|f(\iota_h^{-1}(z))|h_Z(z)\,\mathrm{d}z \\ &\leq 2M\frac{\mu(L)-\mu(L_-^\varepsilon)}{\mu(L)}.\end{aligned}$$

Moreover, Lemma 1 and (5.8) imply

$$\mu(L)-\mu(L_-^\varepsilon) \leq \zeta(1-e^{-\varepsilon})\lambda(L) \leq \zeta^2\varepsilon\mu(L),$$

and this concludes the proof. □

The key to obtain estimates of the second and third term of (5.9), as well as showing that $\mathcal{L}_k = \mathcal{L}_k^w$, with high probability is to establish good concentration inequalities for the counts $\#L$ and $\#L_-^\varepsilon$, which apply across all $L \in \mathcal{L}_k^w$. As we will see in later proofs, by relying on Theorem 3 and ideas similar to [29, Theorem 10 and Lemma 13], it suffices to understand the concentration of $\#R$ across all rectangles in $\mathcal{R}_{\varepsilon,w}$ of non-negligible volume. This is the motivation for the following result, which relies on a Bernstein type inequality for strongly mixing processes.

Lemma 3. *Suppose that (A1)–(A3) are satisfied, and let* $\varepsilon \asymp k^{-1/2}$ *and* $w \asymp k/T$. *Then there exists a constant* $\gamma \in (0,\infty)$ *such that*

$$\sup_{R\in\mathcal{R}_{\varepsilon,w}\colon \mu(R)\geq w} \frac{|\#R - T\mu(R)|}{\sqrt{T\mu(R)}} \leq \gamma \log T \tag{5.10}$$

with probability at least $1 - T^{-1}$ *for all sufficiently large* T.

Proof. Note that, by a union bound, it suffices to establish that for any $R \in \mathcal{R}_{\varepsilon,w}$ with $\mu(R) \geq w$,

$$\mathbb{P}\Big(\Big|\frac{\#R}{T} - \mu(R)\Big| > \gamma \log T \sqrt{\frac{\mu(R)}{T}}\Big) \leq \frac{1}{|\mathcal{R}_{\varepsilon,w}|T}. \tag{5.11}$$

To this end, observe that $(Y_t)_{t\geq 1}$ forms a stationary geometrically ergodic p-th order Markov chain (cf. [3, Theorem 3.1]). It is well-known that any such chain is exponentially α-mixing (see, e.g., [14, p. 89]). In particular, the t-th α-mixing coefficient $\alpha(t) := \sup_{A\in\sigma(X_1),\, B\in\sigma(X_{t+1})} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)|$ of $X_t = (Y_{t-1},\dots,Y_{t-p})$ satisfies

$$\log \alpha(t) \lesssim -t, \qquad t \geq 1. \tag{5.12}$$

Moreover, the α-mixing coefficients of $(\mathbb{1}_R(Z_t))_{t\geq 1}$ are obviously bounded by $(\alpha(t))_{t\geq 0}$ (which do not depend on R), and thus we can rely on a Bernstein type inequality for strongly mixing sequences [23, Theorem 2] to establish that

$$\log \mathbb{P}\Big(\Big|\frac{\#R}{T} - \mu(R)\Big| > x\Big) \lesssim -\frac{x^2 T}{\nu_R^2 + T^{-1} + x(\log T)^2}, \qquad x > 0, \tag{5.13}$$

where

$$\nu_R^2 := \mathrm{Var}(\mathbb{1}_R(Z_1)) + 2\sum_{t=1}^{\infty} |\mathrm{Cov}(\mathbb{1}_R(Z_{t+1}), \mathbb{1}_R(Z_1))|.$$

It is easy to see that $|\mathrm{Cov}(\mathbb{1}_R(Z_{t+1}), \mathbb{1}_R(Z_1))| \leq \min\{\alpha(t), \mu(R)\}$. From this inequality and the fact that $\sqrt{\alpha(t)} \leq \mu(R)$ as long as $t \gtrsim \log(T/k)$, which follows from (5.12) and $\mu(R) \gtrsim k/T$, we deduce that

$$\nu_R^2 \lesssim \mu(R)\Big(1 + \log(T/k) + \sum_{t=1}^{\infty} \sqrt{\alpha(t)}\Big) \lesssim \mu(R)\log T. \tag{5.14}$$

By combining this variance bound with inequality (5.13) and using that $\mu(R) \gtrsim 1/T$ we get

$$\log \mathbb{P}\Big(\Big|\frac{\#R}{T} - \mu(R)\Big| > x\Big) \lesssim -\frac{x^2 T}{\max\{\mu(R)\log T, x(\log T)^2\}}. \tag{5.15}$$

To put it differently, we may choose a sufficiently large constant $\bar{\gamma}$ such that for any fixed $\tau \in (0,\infty)$,

$$\mathbb{P}\Big(\Big|\frac{\#R}{T} - \mu(R)\Big| > x\Big) \le \frac{1}{\tau} \tag{5.16}$$

if

$$x \ge \bar{\gamma}\max\Big\{\frac{(\log T)^2 \log\tau}{T}, \sqrt{\frac{\mu(R)}{T}\log T \log\tau}\Big\}. \tag{5.17}$$

Since $\mu(R) \gtrsim k/T$, the maximum of (5.17) is equal to its last term if

$$k \ge \kappa(\log T)^3 \log\tau$$

for a suitable constant κ. Moreover, if $\tau = |\mathcal{R}_{\varepsilon,w}|T$, Theorem 3(ii) shows that $\log\tau \lesssim \log T$, so if k is chosen in accordance with (A3), the last term of the maximum in (5.17) is the dominating one when T is large. Consequently, by choosing x to be the right-hand side of (5.17) with $\tau = |\mathcal{R}_{\varepsilon,w}|T$, we obtain

$$\mathbb{P}\Big(\Big|\frac{\#R}{T} - \mu(R)\Big| > \bar{\gamma}\sqrt{\frac{\mu(R)}{T}\log T(\log|\mathcal{R}_{\varepsilon,w}| + \log T)}\Big) \le \frac{1}{|\mathcal{R}_{\varepsilon,w}|T}.$$

By using Theorem 3(ii) once again it follows that (5.11) is satisfied for a suitable constant γ and verifies that (5.10) holds with probability at least $1 - T^{-1}$ for all sufficiently large T. □

The next result shows how the inequality (5.10) impacts the magnitude of the third term of (5.9).

Lemma 4. *Suppose that (A1)–(A3) are satisfied, and let $\varepsilon = k^{-1/2}$ and $w = k/(4\zeta T)$ where $\zeta \in (1,\infty)$ is given as in Lemma 1. Then, the inequality (5.10) implies*

$$\sup_{L\in\mathcal{L}_k^w}\Big|\frac{1}{\#L}\sum_{t:\,Z_t\in L} Y_t - \frac{1}{\#L_-^\varepsilon}\sum_{t:\,Z_t\in L_-^\varepsilon} Y_t\Big| \le 6(M + \max_{t=1,\dots,T}|\varepsilon_t|)\frac{\zeta^2 + 2\gamma\log T}{\sqrt{k}}$$

for all sufficiently large T, where M is given by (2.5).

Proof. First, observe that

$$\begin{aligned}\sup_{L\in\mathcal{L}_k^w}\Big|\frac{1}{\#L}\sum_{t:\,Z_t\in L} Y_t - \frac{1}{\#L_-^\varepsilon}\sum_{t:\,Z_t\in L_-^\varepsilon} Y_t\Big| \\ \le 2\big(M + \max_{t=1,\dots,T}|\varepsilon_t|\big)\sup_{L\in\mathcal{L}_k^w}\frac{\#L - \#L_-^\varepsilon}{\#L}.\end{aligned} \tag{5.18}$$

It follows that we need to show how (5.10) implicitly restricts $\#L$ and $\#L_-^\varepsilon$. Initially, we will argue that (5.10) implies

$$\#L \leq e^{\zeta^2\varepsilon} T\mu(L) + \gamma e^{\frac{1}{2}\zeta^2\varepsilon} \log T \sqrt{T\mu(L)}, \tag{5.19}$$

$$\#L \geq \frac{T\mu(L_-^\varepsilon) - \gamma^2 (\log T)^2}{2}, \tag{5.20}$$

$$\text{and} \qquad \#L_-^\varepsilon \geq T\mu(L_-^\varepsilon) - \gamma \log T \sqrt{T\mu(L_-^\varepsilon)} \tag{5.21}$$

for all $L \in \mathcal{L}_k^w$. Consider any rectangle $R \subseteq [0,1]^p$ with $\mu(R) \geq \zeta w$, and note that such rectangle satisfies $\mathrm{Leb}(R) \geq w$ by Lemma 1. Consequently, Theorem 3 implies the existence of an outer approximation $R_+ \supseteq R$ from $\mathcal{R}_{\varepsilon,w}$ with $\mathrm{Leb}(R_+) \leq e^\varepsilon \mathrm{Leb}(R)$. The inequality (5.10) shows in particular that

$$\frac{\#R_+ - T\mu(R_+)}{\sqrt{T\mu(R_+)}} \leq \gamma \log T. \tag{5.22}$$

Obviously $\#R_+ \geq \#R$, and by Lemma 1 the μ-measure of R_+ is bounded in terms of that of R as

$$\mu(R_+) \leq \mu(R) + \zeta^2 (e^\varepsilon - 1)\mu(R) \leq e^{\zeta^2\varepsilon} \mu(R).$$

By combining this with (5.22) we conclude that

$$\sup_{R\colon\, \mu(R) \geq \zeta w} \frac{\#R - e^{\zeta^2\varepsilon} T\mu(R)}{\sqrt{T\mu(R)}} \leq \gamma e^{\frac{1}{2}\zeta^2\varepsilon} \log T, \tag{5.23}$$

where it is implicitly understood that the supremum only runs over rectangles in $[0,1]^p$. Now, if R is a rectangle with $\mu(R) < 2\zeta w$, we may expand it along one or more of the p directions to obtain a new rectangle $\widetilde{R}$ with $R \subseteq \widetilde{R} \subseteq [0,1]^p$ and $\mu(\widetilde{R}) = 2\zeta w$. Thus, by (5.23) this means that

$$\#R \leq \Big(\frac{e^{\zeta^2\varepsilon}}{2} + \frac{\gamma e^{\frac{1}{2}\zeta^2\varepsilon} \log T}{\sqrt{2k}} \Big) k. \tag{5.24}$$

By (A3), the last term in the parenthesis goes to zero and $e^{\zeta^2\varepsilon}$ goes to one as T approaches infinity, so we establish that $\#R < k$ as long as T exceeds a certain threshold (which does not depend on R). To put it differently, as long as T is sufficiently large and for any rectangle $R \subseteq [0,1]^p$, the following implication holds:

$$\#R \geq k \qquad \Longrightarrow \qquad \mu(R) \geq 2\zeta w. \tag{5.25}$$

Consider now any leaf $L \in \mathcal{L}_k$. By (5.25) it must be the case that $\mu(L) \geq 2\zeta w$, and thus (5.19) is an immediate consequence of (5.23). Moreover, the μ-measure of the inner ε-approximation L_-^ε of L is bounded from below as

$$\mu(L_-^\varepsilon) \geq (1 - \zeta^2(1 - e^{-\varepsilon}))\mu(L) \geq 2(1 - \zeta^2(1 - e^{-\varepsilon}))\zeta w \geq \zeta w, \tag{5.26}$$

where the last inequality applies as long as T is large enough. Thus, (5.21) is implied by (5.10). In order to prove (5.20), first note that

$$T\mu(L_-^\varepsilon) \le \#L + \gamma \log T \sqrt{T\mu(L_-^\varepsilon)} \tag{5.27}$$

by (5.21). By dividing both sides of (5.27) with $\sqrt{T\mu(L_-^\varepsilon)}$ and using that $T\mu(L_-^\varepsilon) \ge k/4$ when T is large (by (5.26)) we obtain

$$\sqrt{T\mu(L_-^\varepsilon)} \le \frac{2\,\#L}{\sqrt{k}} + \gamma \log T. \tag{5.28}$$

Now, by using the bound (5.28) for the last term in (5.27) and rearranging terms,

$$\#L \ge \frac{T\mu(L_-^\varepsilon) - \gamma^2 (\log T)^2}{1 + 2\gamma \log T/\sqrt{k}}.$$

By (A3), $2\gamma \log T/\sqrt{k} \le 1$ when T is sufficiently large, and this proves (5.20). Now we use (5.19)–(5.21) to bound $(\#L - \#L_-^\varepsilon)/\#L$ uniformly across $L \in \mathcal{L}_k^w$. For an arbitrary leaf $L \in \mathcal{L}_k^w$ it follows by (5.19) that

$$\big(e^{\frac{1}{2}\zeta^2\varepsilon}\sqrt{T\mu(L)} + \gamma \log T\big)^2 \ge \#L + \gamma^2 (\log T)^2,$$

and hence

$$\begin{aligned} e^{\zeta^2\varepsilon} T\mu(L) &\ge \big(\sqrt{\#L + \gamma^2(\log T)^2} - \gamma \log T\big)^2 \\ &= \#L + 2\gamma^2 (\log T)^2 - 2\gamma \log T \sqrt{\#L + \gamma^2 (\log T)^2} \\ &\ge \#L - 4\gamma \log T \sqrt{\#L}, \end{aligned} \tag{5.29}$$

where, due to (5.20), the last inequality applies as long as T exceeds a certain threshold (which does not depend on L). Moreover, (5.20) implies

$$\sqrt{T\mu(L_-^\varepsilon)} \le \sqrt{2\,\#L} + \gamma \log T \le 2\sqrt{\#L} \tag{5.30}$$

and, as in (5.26), the μ-measure of L_-^ε is bounded from below as

$$\mu(L_-^\varepsilon) \ge (1 - \zeta^2(1 - e^{-\varepsilon}))\mu(L) \ge e^{-2\zeta^2\varepsilon}\mu(L). \tag{5.31}$$

Both (5.30) and (5.31) require that T is large. By starting from (5.21), and then using (5.29)–(5.31), we get the estimate

$$\#L_-^\varepsilon \ge e^{-2\zeta^2\varepsilon} T\mu(L) - 2\gamma \log T \sqrt{\#L} \ge e^{-3\zeta^2\varepsilon}\#L - 6\gamma \log T \sqrt{\#L}$$

for large T. Thus, for such T,

$$\sup_{L \in \mathcal{L}_k^w} \frac{\#L - \#L_-^\varepsilon}{\#L} \le 1 - e^{-3\zeta^2\varepsilon} + \sup_{L \in \mathcal{L}_k^w} \frac{6\gamma \log T}{\sqrt{\#L}} \le \frac{3\zeta^2 + 6\gamma \log T}{\sqrt{k}}.$$

In view of (5.18), this finishes the proof. □

Remark 4. Suppose that we are in the setting of Lemma 4. In its proof it is in fact established that $\mathcal{L}_k = \mathcal{L}_k^w$ when (5.10) holds and T is large. For instance, this is an immediate consequence of (5.25).

In a similar way, we use (5.10) to bound the second term of (5.9); this is detailed in the following lemma.

Lemma 5. *Suppose that (A1)–(A3) are satisfied, and let* $\varepsilon = k^{-1/2}$ *and* $w = k/(4\zeta T)$ *where* $\zeta \in (1,\infty)$ *is given as in Lemma 1. Then, the inequality* (5.10) *implies*

$$\sup_{L\in\mathcal{L}_k^w} |G_T(L_-^\varepsilon)| \le \frac{4M\gamma\log T}{\sqrt{k}} + 2\sup_{R\in\mathcal{R}_{\varepsilon,w}\colon \mu(R)\ge\zeta w} \frac{1}{T\mu(R)}\Big|\sum_{t\colon Z_t\in R}\varepsilon_t\Big| + 2\sup_{R\in\mathcal{R}_{\varepsilon,w}\colon \mu(R)\ge\zeta w} \frac{\big|\frac{1}{T}\sum_{t\colon Z_t\in R} f(X_t) - \mathbb{E}[f(X)\mathbb{1}_R(Z)]\big|}{\mu(R)}$$

for all sufficiently large T, *where* M *is given by* (2.5).

Proof. For any given rectangle R we have the bound

$$|G_T(R)| \le M\frac{|\#R - T\mu(R)|}{\#R} + \frac{1}{\#R}\Big|\sum_{t\colon Z_t\in R}\varepsilon_t\Big| + \frac{T}{\#R}\Big|\frac{1}{T}\sum_{t\colon Z_t\in R} f(X_t) - \mathbb{E}[f(X)\mathbb{1}_R(Z)]\Big|. \tag{5.32}$$

When (5.10) is satisfied and T is large enough, it follows from (5.26) (which holds under (A1)–(A3)) that

$$\mathcal{R}' := \{R\in\mathcal{R}_{\varepsilon,w} \,:\, \mu(R)\ge\zeta w\} \supseteq \{L_-^\varepsilon \,:\, L\in\mathcal{L}_k^w\}. \tag{5.33}$$

Moreover, for any $R\in\mathcal{R}'$, (5.10) implies immediately that

$$\#R \ge T\mu(R)\Big(1 - \frac{2\gamma\log T}{\sqrt{k}}\Big) \ge \frac{T\mu(R)}{2}, \tag{5.34}$$

and hence also that

$$\frac{|\#R - T\mu(R)|}{\#R} \le \frac{4\gamma\log T}{\sqrt{k}} \tag{5.35}$$

as soon as T exceeds a certain threshold (which is independent of R). By combining (5.32)–(5.35) we obtain the result. □

Since $\mathcal{R}_{\varepsilon,w}$ is a rather small collection of sets, it is hinted by Lemmas 4 and 5 that the only missing part in order to prove Theorem 1 is to obtain bounds on

$$\max_{t=1,\dots,T}|\varepsilon_t|, \qquad \frac{1}{T}\Big|\sum_{t\colon Z_t\in R}\varepsilon_t\Big| \qquad \text{and} \qquad \Big|\frac{1}{T}\sum_{t\colon Z_t\in R} f(X_t) - \mathbb{E}[f(X)\mathbb{1}_R(Z)]\Big|$$

for any $R\in\mathcal{R}_{\varepsilon,w}$ with $\mu(R)\ge\zeta w$. The first term is easy to handle, since it is a maximum of i.i.d. random variables satisfying (2.3). The last two terms can be handled by relying on Bernstein type inequalities for martingale differences and strongly mixing sequences. We will go through the details below.

Proof of Theorem 1. The proof goes by defining four events $\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_3$ and $\mathcal{E}_4$ and arguing that (i) the inequality (2.8) holds true on $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4$, and (ii) each event $\mathcal{E}_i$ occurs with probability at least $1 - T^{-1}$. With $\varepsilon = k^{-1/2}$ and $w = k/(4\zeta T)$, $\zeta \in (1, \infty)$ given as in Lemma 1, the events that we will consider are the following:

$$\begin{aligned}
\mathcal{E}_1 &:= \Big\{ \sup_{R \in \mathcal{R}_{\varepsilon,w}\colon \mu(R) \geq \zeta w} \frac{|\#R - T\mu(R)|}{\sqrt{T\mu(R)}} \leq \gamma \log T \Big\}, \\
\mathcal{E}_2 &:= \big\{ \max_{t=1,\dots,T} |\varepsilon_t| \leq c_1 \log T \big\}, \\
\mathcal{E}_3 &:= \Big\{ \sup_{R \in \mathcal{R}_{\varepsilon,w}\colon \mu(R) \geq \zeta w} \frac{1}{T\mu(R)} \Big| \sum_{t\colon Z_t \in R} \varepsilon_t \Big| \leq c_2 \frac{\log T}{\sqrt{k}} \Big\} \cup \mathcal{E}_1^c, \\
\mathcal{E}_4 &:= \Big\{ \sup_{R \in \mathcal{R}_{\varepsilon,w}\colon \mu(R) \geq \zeta w} \frac{\big|\frac{1}{T}\sum_{t\colon Z_t \in R} f(X_t) - \mathbb{E}[f(X)\mathbb{1}_R(Z)]\big|}{\mu(R)} \leq c_3 \frac{\log T}{\sqrt{k}} \Big\}.
\end{aligned}$$

Here γ is the constant from Lemma 3, while c_1, c_2 and c_3 will be introduced during the proof. Moreover, $\mathcal{E}_1^c$ refers to the complement of $\mathcal{E}_1$.

Proof of (i). Suppose that the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4$ has occurred. Then, by using (5.9) and Lemmas 2, 4 and 5, it follows that

$$\begin{aligned}
\sup_{L \in \mathcal{L}_k} |G_T(L)| \leq \frac{2M\zeta^2}{\sqrt{k}} + \frac{4M\gamma \log T}{\sqrt{k}} + \frac{2c_2 \log T}{\sqrt{k}} + \frac{2c_3 \log T}{\sqrt{k}} \\
+ 6(M + c_1 \log T)\frac{\zeta^2 + 2\gamma \log T}{\sqrt{k}}.
\end{aligned}$$

In view of this inequality, (5.7) and Remark 4, we conclude that (2.8) is satisfied on $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4$ for a suitably chosen constant β.

Proof of (ii). The content of Lemma 3 is exactly that $\mathbb{P}(\mathcal{E}_1) \geq 1 - T^{-1}$. Since the moments of ε_1 meet (2.3), its distribution is sub-exponential and (2.6) holds. Moreover, $\varepsilon_1, \dots, \varepsilon_T$ are i.i.d. random variables, so by applying a union bound we obtain the estimate

$$\mathbb{P}\big(\max_{t=1,\dots,T} |\varepsilon_t| > x\big) \leq \gamma_1 T e^{-\gamma_2 x}, \qquad x > 0.$$

In other words, $\max_{t=1,\dots,T} |\varepsilon_t| \leq x$ with probability at least $1 - T^{-1}$ if $x \geq \log(\gamma_1 T^2)/\gamma_2$, and this shows $\mathbb{P}(\mathcal{E}_2) \geq 1 - T^{-1}$ for some c_1.

Next, consider any rectangle $R \in \mathcal{R}_{\varepsilon,w}$ with $\mu(R) \geq \zeta w$ and observe that $(\varepsilon_t \mathbb{1}_R(Z_t))_{t \geq 1}$ is a martingale difference sequence with respect to the filtration $\mathcal{F}_t = \sigma(Y_s \colon s \leq t)$. Since ε_t is independent of $\mathcal{F}_{t-1}$ and its moments satisfy (2.3),

$$\mathbb{E}[|\varepsilon_t \mathbb{1}_R(Z_t)|^m \mid \mathcal{F}_{t-1}] \leq m! c^{m-2} \mathbb{1}_R(Z_t), \qquad m \geq 3.$$

In particular, these observations show that we can rely on a Bernstein (Freedman) type inequality for unbounded summands to obtain

$$\log \mathbb{P}\Big(\frac{1}{T}\Big|\sum_{t\colon Z_t\in R}\varepsilon_t\Big| > x,\ \#R \le y\Big) \lesssim -\frac{x^2T}{y/T+x} \tag{5.36}$$

for any $x, y > 0$. Such a result can, e.g., be found in [12, Theorem 8.2.2]. Let γ be the constant from Lemma 3, consider specifically

$$y = T\mu(R) + \gamma \log T\sqrt{T\mu(R)}$$

and note that $y \le 2T\mu(R)$ when T is large (by (A3)). By using (5.36) with this choice of y it follows that

$$\log \mathbb{P}\Big(\Big\{\frac{1}{T}\Big|\sum_{t\colon Z_t\in R}\varepsilon_t\Big| > x\Big\}\cap\mathcal{E}_1\Big) \lesssim -\frac{x^2T}{\max\{\mu(R),x\}}.$$

From this inequality we deduce the existence of a constant κ, such that for any $\tau > 0$,

$$\mathbb{P}\Big(\Big\{\frac{1}{T}\Big|\sum_{t\colon Z_t\in R}\varepsilon_t\Big| > x\Big\}\cap\mathcal{E}_1\Big) \le \frac{1}{\tau} \tag{5.37}$$

when

$$x = \kappa \max\Big\{\frac{\log\tau}{T}, \sqrt{\frac{\mu(R)}{T}\log\tau}\Big\}. \tag{5.38}$$

The maximum in (5.38) is equal to its second term if

$$k \ge 4\log\tau. \tag{5.39}$$

It follows from Theorem 3(ii) and (A3) that (5.39) is satisfied if $\tau = |\mathcal{R}_{\varepsilon,w}|T$ and T is large, and thus (5.37) and (5.38) show that

$$\mathbb{P}\Big(\Big\{\frac{1}{T}\Big|\sum_{t\colon Z_t\in R}\varepsilon_t\Big| > c_2\log T\sqrt{\frac{\mu(R)}{4T}}\Big\}\cap\mathcal{E}_1\Big) \le \frac{1}{|\mathcal{R}_{\varepsilon,w}|T},$$

for a suitable constant c_2. By rearranging terms and using that $T\mu(R) \ge k/4$ it follows that

$$\mathbb{P}\Big(\Big\{\frac{1}{T\mu(R)}\Big|\sum_{t\colon Z_t\in R}\varepsilon_t\Big| > c_2\frac{\log T}{\sqrt{k}}\Big\}\cap\mathcal{E}_1\Big) \le \frac{1}{|\mathcal{R}_{\varepsilon,w}|T}.$$

Consequently, by relying on a union bound over all rectangles in $\mathcal{R}_{\varepsilon,w}$, we establish that $\mathbb{P}(\mathcal{E}_3) \ge 1 - T^{-1}$.

To show $\mathbb{P}(\mathcal{E}_4) \ge 1 - T^{-1}$ we consider again an arbitrary rectangle $R \in \mathcal{R}_{\varepsilon,w}$ with $\mu(R) \ge \zeta w$. The sequence $(f(X_t)\mathbb{1}_R(Z_t))_{t\ge1}$ is bounded and α-mixing, and its associated mixing coefficients are bounded by those of $(X_t)_{t\ge1}$, which we will denote by $(\alpha(t))_{t\ge1}$. In particular, by (5.12) it follows that the mixing coefficients

of $(f(X_t)\mathbb{1}_R(Z_t))_{t\geq 1}$ are bounded by an exponentially decaying sequence of numbers with a decay rate which does not depend on R. Consequently, as in the proof of Lemma 3, we can again rely on [23, Theorem 2] to obtain that

$$\log \mathbb{P}\Big(\Big|\frac{1}{T}\sum_{t\colon Z_t\in R} f(X_t) - \mathbb{E}[f(X)\mathbb{1}_R(Z)]\Big| > x\Big) \lesssim -\frac{x^2 T}{\nu_R^2 + T^{-1} + x(\log T)^2} \tag{5.40}$$

where

$$\nu_R^2 \coloneqq \operatorname{Var}(f(X_1)\mathbb{1}_R(Z_1)) + 2\sum_{t=1}^{\infty} |\operatorname{Cov}(f(X_{t+1})\mathbb{1}_R(Z_{t+1}), f(X_1)\mathbb{1}_R(Z_1))|.$$

Note that

$$\inf\{y\in[0,\infty)\,:\,\mathbb{P}(|f(X)|\mathbb{1}_R(Z) > y) \leq u\} \leq M\mathbb{1}_{\{u\leq \mu(R)\}},$$

so it follows by Rio's covariance inequality [25, Theorem 1.1] that

$$|\operatorname{Cov}(f(X_{t+1})\mathbb{1}_R(Z_{t+1}), f(X_1)\mathbb{1}_R(Z_1))| \leq 4M^2\min\{\alpha(t),\mu(R)\}.$$

Thus, by using the same arguments as in the proof of Lemma 3 (in relation to (5.14)) we establish $\nu_R^2 \lesssim \mu(R)\log T$, meaning that (5.40) implies

$$\log \mathbb{P}\Big(\Big|\frac{1}{T}\sum_{t\colon Z_t\in R} f(X_t) - \mathbb{E}[f(X)\mathbb{1}_R(Z)]\Big| > x\Big) \lesssim -\frac{x^2 T}{\max\{\mu(R)\log T, x(\log T)^2\}}. \tag{5.41}$$

Since the right-hand side of (5.41) is the same as in (5.15), we can use the exact same arguments to verify the existence of a constant c_3 such that

$$\mathbb{P}\left(\Big|\frac{1}{T}\sum_{t\colon Z_t\in R} f(X_t) - \mathbb{E}[f(X)\mathbb{1}_R(Z)]\Big| > c_3\log T\sqrt{\frac{\mu(R)}{4T}}\right) \leq \frac{1}{|\mathcal{R}_{\varepsilon,w}|T}$$

when T exceeds a certain threshold (not depending on R). In particular,

$$\mathbb{P}\left(\frac{\left|\frac{1}{T}\sum_{t\colon Z_t\in R} f(X_t) - \mathbb{E}[f(X)\mathbb{1}_R(Z)]\right|}{\mu(R)} > c_3\frac{\log T}{\sqrt{k}}\right) \leq \frac{1}{|\mathcal{R}_{\varepsilon,w}|T}$$

from which it follows by a union bound over rectangles in $\mathcal{R}_{\varepsilon,w}$ that $\mathbb{P}(\mathcal{E}_4) \geq 1 - T^{-1}$. We have now argued that both (i) and (ii) outlined in the beginning of the proof hold true, and hence we obtain the desired result. □

We now turn to the task of proving Theorem 2. To do so, we will make use of an auxiliary result which is presented in Lemma 6 below. In this formulation, $\operatorname{diam}(A) \coloneqq \sup_{x,x'\in A}\|x'-x\|$ is the diameter of $A\subseteq\mathbb{R}^p$.

Lemma 6. *Suppose (A1), (A2) and (A5) are satisfied and that* $\Lambda = \Lambda(\mathcal{D}_T,\Theta) \in \mathcal{V}_{\alpha,k,m}$ *for all* T. *Then, for any* $x\in\mathbb{R}^p$, $\operatorname{diam}(A_\Lambda(x)) \to 0$ *as* $T\to\infty$ *with probability one.*

Proof. Let us represent the rectangle $A_\Lambda(x)$ in Λ containing $x \in \mathbb{R}^p$ as $A_\Lambda(x) = A^1_\Lambda(x) \times \cdots \times A^p_\Lambda(x)$. Then, it suffices to show that

$$\mathrm{Leb}(A^i_\Lambda(x)) \longrightarrow 0, \qquad T \to \infty, \tag{5.42}$$

with probability one for $i = 1, \ldots, p$. To this end, imagine the tree illustrating how Λ is obtained by the recursive partitioning scheme and consider the path that x takes down the tree from its root to the leaf $A_\Lambda(x)$. Let d denote the depth of the tree at x (i.e., x traverses exactly $d-1$ nodes before it reaches $A_\Lambda(x)$), and let A^l be the node containing x at depth l. In particular, $(A^l)_l$ is a decreasing sequence of sets with $A^1 = \mathbb{R}^p$ and $A^d = A_\Lambda(x)$, and $A^l_j \neq A^{l+1}_j$ for exactly one j (with the notation $A = A_1 \times \cdots \times A_p$). We let $S^i_l = \mathbb{1}_{A^l_i \neq A^{l+1}_i}$ indicate whether the node containing x at depth l will be split along the i-th direction, and $\tau^i_l = \min\{j \in \{\tau^i_{l-1}+1, \ldots, d-1\} : S^i_j = 1\}$ the depth at which x will experience the l-th split along the i-th direction ($\tau^i_0 \equiv 0$ and, say, $\tau^i_l = \infty$ if the set is empty). For an illustration of these definitions, see Figure 6. By the construction of the tree (specifically, the rules (i) and (iii) outlined in Section 3) it holds that $m \geq T\alpha^{d-1}$, and hence

$$d \geq 1 + \frac{\log(T/m)}{\log(\alpha^{-1})}. \tag{5.43}$$

Recall also that the tree is constructed in such a way that there exists a strictly positive constant ρ which is a lower bound for the probability ρ_i of splitting along the i-th direction at any given node. Suppose for simplicity (but without loss of generality) that, in fact, $\rho_i = \rho$. Then, $(S^i_l)_{l \geq 1}$ is a sequence of i.i.d. Bernoulli random variables and thus, with probability one,

$$\sum_{l=1}^{n} S^i_l \longrightarrow \infty, \qquad n \to \infty.$$

Since the right-hand side of (5.43) tends to infinity by (A5), it follows that

$$|\{l \in \{1, \ldots, d-1\} : \tau^i_l < \infty\}| \longrightarrow \infty, \qquad T \to \infty, \tag{5.44}$$

almost surely. Consider an arbitrary number $l \in \{1, \ldots, d-1\}$ with $\tau^i_l < \infty$ and let h be any fixed density which aligns with Lemma 1. Then

$$\begin{aligned}
\frac{\mathrm{Leb}(F_h(A_i^{\tau^i_l+1}))}{\mathrm{Leb}(F_h(A_i^{\tau^i_l}))} &= 1 - \frac{\mathrm{Leb}(\iota_h(A^{\tau^i_l}) \setminus \iota_h(A^{\tau^i_l+1}))}{\mathrm{Leb}(\iota_h(A^{\tau^i_l}))} \\
&\leq 1 - \zeta^{-2} \frac{\mathbb{P}_\Lambda(X \in A^{\tau^i_l} \setminus A^{\tau^i_l+1})}{\mathbb{P}_\Lambda(X \in A^{\tau^i_l})} \\
&\leq 1 - \zeta^{-2}(1 - \mathbb{P}_\Lambda(X \in A^{\tau^i_l+1} \mid X \in A^{\tau^i_l})).
\end{aligned} \tag{5.45}$$

(Note that, for a given interval $A \subseteq \mathbb{R}$, $F_h(A) \subseteq [0,1]$ refers to the image of A under F_h.) Since the tree is grown with respect to the rule (iii),

$$\frac{|\{t \in \{1, \ldots, T\} : X_t \in A^{\tau^i_l+1}\}|}{|\{t \in \{1, \ldots, T\} : X_t \in A^{\tau^i_l}\}|} \leq 1 - \alpha,$$

so by a Glivenko–Cantelli theorem for ergodic processes (e.g., [1, Theorem 1]) we establish that

$$\limsup_{T\to\infty} \mathbb{P}_\Lambda(X \in A^{\tau_l^i+1} \mid X \in A^{\tau_l^i}) \le 1-\alpha \tag{5.46}$$

with probability one. By combining (5.45) and (5.46) it follows that we can fix $\delta \in (0,1)$ such that, with probability one,

$$\frac{\text{Leb}(F_h(A_i^{\tau_l^i+1}))}{\text{Leb}(F_h(A_i^{\tau_l^i}))} \le 1-\delta \tag{5.47}$$

for all sufficiently large T. By definition, $A_i^{\tau_{l-1}^i+1} = A_i^{\tau_l^i}$, so by repeated use of (5.47) we obtain, for any given $n \in \{1,\dots,d-1\}$ with $\tau_n^i < \infty$,

$$\text{Leb}(F_h(A_\Lambda^i(x))) \le \text{Leb}(F_h(A_i^{\tau_n^i+1})) = \prod_{l=1}^{n} \frac{\text{Leb}(F_h(A_i^{\tau_l^i+1}))}{\text{Leb}(F_h(A_i^{\tau_{l-1}^i+1}))} \le (1-\delta)^n \tag{5.48}$$

for all sufficiently large T almost surely. Thus, from (5.44) we deduce that $\text{Leb}(F_h(A_\Lambda^i(x))) \to 0$ almost surely as $T \to \infty$, and this completes the proof. □

Remark 5. While the arguments used to prove Lemma 6 are somewhat similar to those of [22, Lemma 2], they may appear slightly more complicated. The reason is that the proof of [22, Lemma 2] relies on an estimate of the form

$$|\{t \in \{1,\dots,T\} : X_t^i \in A_i^{\tau_n^i+1}\}| \le T(1-\alpha)^n, \tag{5.49}$$

from which one immediately deduces that $\text{Leb}(F_h(A_\Lambda^i(x))) \le \zeta(1-\alpha)^n$, which completes the proof. (Here $X_t^i = Y_{t-i}$ refers to the i-th entry of X_t.) However, the estimate (5.49) does not apply in general under the rules (i)–(iv) (of Section 3) upon which trees are grown, and hence a few additional arguments are needed to establish the alternative inequality (5.48).

Proof of Theorem 2. By Corollary 1,

$$|\hat{f}_T(x) - f(x)| \le \max_{b=1,\dots,B} |T_{\Lambda_b}^*(x) - f(x)| + o_p(1)$$

$$\text{and} \quad |\hat{f}_T(X) - f(X)| \le \max_{b=1,\dots,B} |T_{\Lambda_b}^*(X) - f(X)| + o_p(1)$$

for suitable $\Lambda_1,\dots,\Lambda_B \in \mathcal{V}_{\alpha,k,m}$. Thus, it suffices to show that $T_\Lambda^*(x) \to f(x)$ and $T_\Lambda^*(X) \to f(X)$ in probability as $T \to \infty$ when $\Lambda \in \mathcal{V}_{\alpha,k,m}$ for all T. By using Lemma 6 together with the inequality

$$|T_\Lambda^*(x) - f(x)| \le \frac{\mathbb{E}_\Lambda[|f(X)-f(x)|\mathbb{1}_{A_\Lambda(x)}(X)]}{\mathbb{P}_\Lambda(X \in A_\Lambda(x))} \le C\,\text{diam}(A_\Lambda(x)),$$

which holds by (A4), it follows that $T_\Lambda^*(x) \to f(x)$ almost surely and, in particular, in probability. Here, as in the proof of Lemma 6, subscript T indicates that we are conditioning on the randomness related to the partition Λ. The last part follows immediately from Tonelli's theorem as this implies that, on an event with probability one, $T_\Lambda^*(x) \to f(x)$ for (Lebesgue) almost all $x \in \mathbb{R}^p$. □

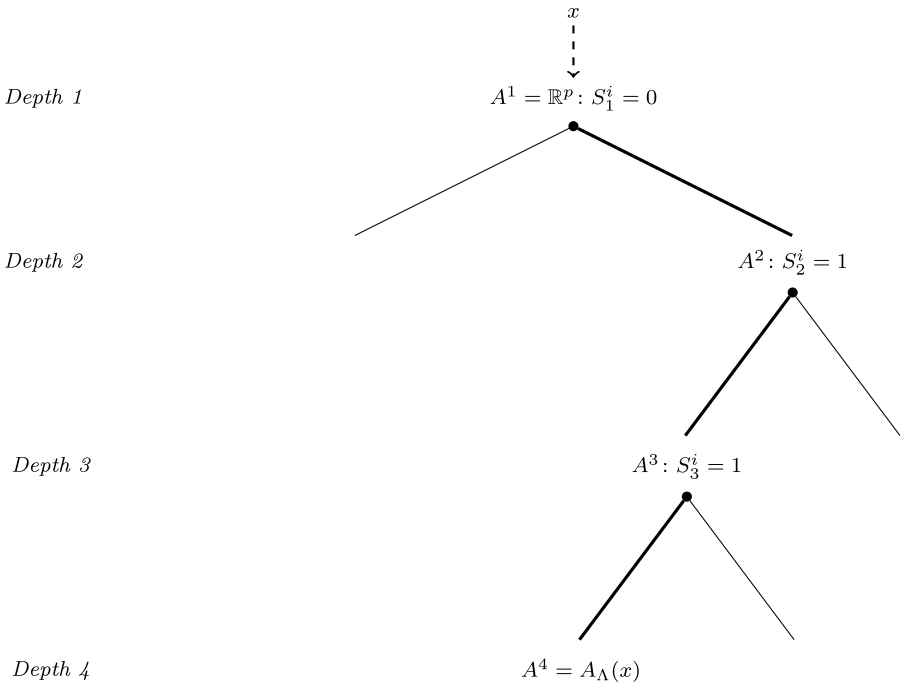

FIG 6. *An illustration of a tree with a depth of* $d = 4$ *at* x. *In this example* $\tau_1^i = 2$ *and* $\tau_2^i = 3$.

Acknowledgements

This work was supported by NSF grant DMS-2015379 for Davis and by Danish Council for Independent Research grant 9056-00011B for Nielsen. We wish to thank the referees for their constructive comments which led to an improvement of the manuscript.

References

[1] ADAMS, T. M. and NOBEL, A. B. (2010). Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling. *Ann. Probab.* **38** 1345–1367. MR2663629

[2] AMARATUNGA, D., CABRERA, J. and LEE, Y.-S. (2008). Enriched random forests. *Bioinformatics* **24** 2010–2014.

[3] AN, H. Z. and HUANG, F. C. (1996). The geometrical ergodicity of nonlinear autoregressive models. *Statist. Sinica* **6** 943–956. MR1422412

[4] ARLOT, S. and GENUER, R. (2014). Analysis of purely random forests bias. *arXiv preprint* arXiv:1407.3939.

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

[6] Biau, G. and Devroye, L. (2010). On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *J. Multivariate Anal.* **101** 2499–2518. MR2719877
[7] Biau, G. and Scornet, E. (2016). A random forest guided tour. *TEST* **25** 197–227. MR3493512
[8] Breiman, L. (1996). Bagging predictors. *Machine learning* **24** 123–140.
[9] Breiman, L. (2001). Random forests. *Machine learning* **45** 5–32. MR3874153
[10] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and regression trees. Wadsworth Statistics/Probability Series*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392
[11] Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J. and Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology* **88** 2783–2792.
[12] De la Pena, V. and Giné, E. (1999). *Decoupling: From Dependence to Independence*. Springer Science & Business Media. MR1666908
[13] Díaz-Uriarte, R. and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics* **7** 3.
[14] Doukhan, P. (2012). *Mixing: properties and examples* **85**. Springer Science & Business Media. MR1312160
[15] Genuer, R. (2012). Variance reduction in purely random forests. *J. Nonparametr. Stat.* **24** 543–562. MR2968888
[16] Geurts, P., Ernst, D. and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning* **63** 3–42.
[17] Gu, S., Kelly, B. and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies* **33** 2223–2273.
[18] Härdle, W., Lütkepohl, H. and Chen, R. (1997). A review of nonparametric time series analysis. *International Statistical Review* **65** 49–72.
[19] Howard, J. and Bowles, M. (2012). The two most important algorithms in predictive modeling today. In *Strata Conference presentation, February* **28**.
[20] Kumar, M. and Thenmozhi, M. (2006). Forecasting stock index movement: A comparison of support vector machines and random forest. In *Indian institute of capital markets 9th capital markets conference paper*.
[21] Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *J. Amer. Statist. Assoc.* **101** 578–590. MR2256176
[22] Meinshausen, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.* **7** 983–999. MR2274394
[23] Merlevède, F., Peligrad, M. and Rio, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume. Inst. Math. Stat. (IMS) Collect.* **5** 273–292. Inst. Math. Statist., Beachwood, OH. MR2797953
[24] Prasad, A. M., Iverson, L. R. and Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological

prediction. *Ecosystems* **9** 181–199.

[25] Rio, E. (1993). Covariance inequalities for strongly mixing processes. *Ann. Inst. H. Poincaré Probab. Statist.* **29** 587–597. MR1251142

[26] Scornet, E., Biau, G. and Vert, J.-P. (2015). Consistency of random forests. *Ann. Statist.* **43** 1716–1741. MR3357876

[27] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A. and Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *CVPR 2011* 1297–1304. IEEE.

[28] Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* **113** 1228–1242. MR3862353

[29] Wager, S. and Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint* arXiv:1503.06388.