Asymptotic confidence bands for centered purely random forests

Dissertation

zur Erlangung des Doktorgrades
an der Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Mathematik
der Universität Hamburg

vorgelegt von

Jan Rabe

Hamburg, 2025

Folgende Gutachter empfehlen die Annahme der Dissertation:
Prof. Dr. Natalie Neumeyer
Prof. Dr. Mathias Trabs
Dr. Katharina Proksch

Vorsitzender der Prüfungskommission:
Prof. Dr. Thomas Schmidt

Tag der Disputation: 09. Juli 2025

Danksagung

Ich möchte einigen Menschen danken, die mich während der letzten vier Jahre auf verschiedenste Weise unterstützt haben. Zuerst gilt mein Dank Natalie Neumeyer und Mathias Trabs für die Betreuung meiner Promotion. Ihr beide seid maßgeblich mitverantwortlich für die Fertigstellung meiner Dissertation.

Mathias, du hast mir regelmäßige Besuche in Karlsruhe ermöglicht, die nicht nur wichtig für meine Arbeit waren, sondern mir auch die Möglichkeit gegeben haben, viele nette Kollegen kennenzulernen. Du hast dir in diesen Wochen immer viel Zeit genommen, gemeinsam über die Problemstellen meiner Arbeit nachzudenken. In den meisten Fällen hat mich das einen entscheidenden Schritt vorangebracht und wenn es mal nicht so lief, hat deine Zuversicht über die Lösbarkeit der Probleme mich motiviert, weiter daran zu arbeiten. Für all das und nicht zuletzt für deinen Zuspruch, noch als Betreuer meiner Masterarbeit, den Weg zur Promotion einzuschlagen, möchte ich dir danken.

Natalie, du hattest immer spontan Zeit für mich, wenn ich mathematische oder auch andere Fragen hatte. Deine gelassene Art, die ich schon als Student in mündlichen Prüfungen geschätzt habe, und deine wertvollen Tipps haben immer dazu geführt, dass ich nach einem Gespräch über ein fachliches Problem zuversichtlicher war als davor. Auch die Zusammenarbeit bei den Lehrveranstaltungen, deren Übungen ich geleitet habe, oder bei mündlichen Prüfungen habe ich immer als sehr angenehm empfunden. Für all das danke ich dir von Herzen!

Ich möchte außerdem Katharina Proksch dafür danken, dass sie sich dazu bereit erklärt hat, meine Dissertation zu begutachten.

Allen (ehemaligen) Kollegen aus dem Bereich Mathematische Statistik und Stochastische Prozesse an der Universität Hamburg danke ich für eine angenehme Arbeitsatmosphäre, für gute Zusammenarbeit in der Lehre und bei einigen mündlichen Prüfungen sowie für viele schöne Gesprächsrunden in den Teepausen. Insbesondere geht ein großer Dank an Maximilian Steffen, Sebastian Neblung, Benedikt Lütke Schwienhorst und Niclas Jacobsen, die die Mühe auf sich genommen haben, meine Arbeit Korrektur zu lesen und mir wertvolles Feedback gegeben haben.

Zuletzt möchte ich meiner Familie und insbesondere meinen Eltern danken, die mich schon im Studium und vor allem auch während meiner Promotion auf jegliche Weise unterstützt haben. Danke für diese Unterstützung, für euer Interesse und für eure Aufmunterungen, wenn ich sie gebraucht habe.

Contents

List of Figures

List of Tables

List of Algorithms

Symbols, Notation and Conventions

$\|\cdot\|$	norm on a metric space, usually euclidean norm
$\|\cdot\|_\infty$	supremum norm
$\|\cdot\|_q$	q-norm on $\mathbb{R}^p$
$x = (x^{(1)}, \ldots, x^{(p)})$	vector in $\mathbb{R}^p$ with entries $x^{(l)} \in \mathbb{R}$ for $l \in \{1, \ldots, p\}$
$[n]$	$\{1, \ldots, n\}$, set of the first n natural numbers for $n \in \mathbb{N}$
$x \vee y$	minimum of x and y
$\|A\|$	number of elements in the set A
$\mathfrak{d}(A)$	diameter of a set $A \subset \mathbb{R}^p$, p.10
$\mathbb{V}(A)$	volume of a set $A \subset \mathbb{R}^p$, p.47
$\lfloor\cdot\rfloor$, $\lceil\cdot\rceil$	floor function, ceiling function
$\mathbb{I}$	indicator function
$\mathbb{E}$, Var, Cov	expectation, variance and covariance
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
$\mathrm{Bin}(n, p)$	Binomial distribution with n trials and success probability p
i.i.d.	independent and identically distributed
$\sim$	distributed as
$\stackrel{d}{=}$	equality in distribution
$\stackrel{d}{\to}$, $\stackrel{\mathbb{P}}{\to}$	convergence in distribution, convergence in probability
o, $\mathcal{O}$	small and big Landau symbol, usually for $n \to \infty$
$x_n = \Theta(y_n)$	$x_n = \mathcal{O}(y_n)$ and $y_n = \mathcal{O}(x_n)$
$o_\mathbb{P}$, $\mathcal{O}_\mathbb{P}$	small and big stochastic Landau symbol
$x_n \lesssim y_n$	$x_n = \mathcal{O}(y_n)$

$B_{r,n}$	set of subsets of $\{1,\dots,n\}$ with r elements, p.15
$U_{n,r_n,\omega}$	generalized complete U-statistic with kernel of order r_n, p.16
$U_{n,r_n,N,\omega}$	generalized incomplete U-statistic with kernel of order r_n, p.16
m	regression function, p.9
p	dimension of the feature space, p.9
$X \in [0,1]^p$	independent variable in regression model, p.9
f_X	density of X
F_X	distribution function (cdf) of X
$Y \in \mathbb{R}$	dependent variable in regression model, p.9
$\varepsilon \in \mathbb{R}$	error random variable in regression model, p.9
$\sigma^2 \in [0,\infty)$	variance of ε
$Z = (X,Y)$	observation pair, p.9
n	number of observations
RF	random forest
CART	classification and regression tree
CPRF	centered purely random forest, p.42
KeRF	kernel random forest, p.57
r_n	tuning parameter, number of observations used per regression tree
k	tuning parameter, tree depth, number of splits
N	tuning parameter, number of trees in a random forest
$U^{(\mathrm{RF})}_{n,r_n,N,\omega}(x_0)$	RF (usually CPRF) with generalized incomplete U-statistic form, p.42/p.44
$U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0)$	RF (usually CPRF) in generalized complete U-statistic form, p.42/p.44
ω	partition creating random variable, p.41/p.44
$A_k(x_0,\omega)$	partition cell around x_0 in a tree of a CPRF, p.44
$\mathcal{V}_{\cap,k}$	cell mixing coefficient, p.47

$\mathcal{F}_k$	function class of $f_{x_0,k}$, p.84
$N_f(k)$	number of undividable cells, p.49, also number of elements in $\mathcal{F}_k$, p.84
B_k	Gaussian process on $\mathcal{F}_k$, p.84
$\mathbf{S}_k$	random variable whose distribution is equal to $\sup_{f\in\mathcal{F}_k} \lvert B_k f\rvert$, p.84

Chapter 1

Introduction

Regression models provide a framework for one of the most prominent estimation problems in mathematical statistics, the prediction of the relationship between a dependent or response variable $Y \in \mathbb{R}$ and an independent variable or covariate $X \in \mathcal{X}$. The non-specific structure of this problem leads to a wide field of applications. An important class of regression models are nonparametric models that avoid specific assumptions on the relationship between X and Y. A nonparametric regression model with homoscedastic noise is described by

$$Y = m(X) + \varepsilon,$$

where m is the so called regression function and $\varepsilon \in \mathbb{R}$ is an error term that is independent of X. The codomain $\mathcal{X}$ of the random variable X is called feature space and usually is a subset of $\mathbb{R}^p$ for $p \in \mathbb{N}$. Throughout we will consider $\mathcal{X} = [0,1]^p$. The entries of X are called features or covariates. In this model, the goal is to estimate the function m based on a training sample consisting of i.i.d. copies of (X, Y). There are various types of estimators in this model, for instance classical smooth estimators such as kernel estimators, spline estimators or wavelet estimators. There are also non smooth estimators such as the histogram or regression trees, which are piecewise constant. Modern estimators that are frequently regarded as machine learning algorithms include neural networks and tree ensemble methods, such as boosted trees and, in particular, random forests. The latter are the central estimators considered in this dissertation.

In mathematical statistics, the focus is on statistical guarantees for the performance of these estimators. Often the performance is analyzed asymptotically in the sample size and important questions are the consistency and asymptotic distribution of the estimators. These aspects can both be analyzed pointwise for a fixed $x_0 \in \mathbb{R}^p$ or uniformly on the whole feature space. Pointwise consistency of an estimator $\hat{m}$ means that

$$m(x_0) - \hat{m}(x_0) \xrightarrow{\mathbb{P}} 0,$$

whereas the stronger uniform consistency means

$$\sup_{x_0 \in \mathcal{X}} |m(x_0) - \hat{m}(x_0)| \xrightarrow{\mathbb{P}} 0.$$

Both pointwise and uniform consistency are desirable for any estimator.

The asymptotic distribution of an estimator is useful for inference because it gives insight into the quality of a single estimator. Commonly it is used to construct asymptotic confidence sets for the regression function, which are a quantitative description of the quality of the estimator. Pointwise results allow for the construction of confidence intervals, but uniform results can be used for constructing confidence bands and goodness-of-fit tests. An asymptotic confidence interval of level $1-\beta \in (0,1)$ for $m(x_0)$ is a random interval $\mathcal{C}_n$ that satisfies

$$\liminf_{n\to\infty} \mathbb{P}(m(x_0) \in \mathcal{C}_n) \geq 1-\beta.$$

A confidence band $\mathcal{C}_n(x)$ is the uniform generalization of a confidence interval. For every $x \in \mathcal{X}$, it is an interval and it satisfies

$$\liminf_{n\to\infty} \mathbb{P}(m(x) \in \mathcal{C}_n(x),\ \forall x \in \mathcal{X}) \geq 1-\beta.$$

The statistical theory for classical methods is more developed than that of modern machine learning estimators. This work will address confidence intervals and bands derived from random forests, which are commonly referred to as machine learning algorithms, but are still nonparametric regression estimators.

Random forests

Random forests, initially proposed by Breiman (2001), are generally applicable to regression and classification problems. However, the focus of this thesis is on regression problems. Before explaining random forests, it is essential to understand regression trees since random forests aggregate an ensemble of regression trees by averaging. Each regression tree itself is an estimator of the regression function, which creates a partition of the feature space and performs a piecewise constant estimation on the partition. Breiman et al. (1984) provide a comprehensive overview of classification and regression trees (CART). In the aforementioned book, the authors introduce the CART-split criterion for partition construction, which is commonly employed in random forests. We call the random forest variant that utilizes this criterion the classical random forest.

For the averaging of the trees in a random forest to be useful, it is necessary to create diversity among the trees. Two important aspects lead to this diversity. First, each regression tree uses its own subset of the training sample. Second, the construction of the partition for each tree is randomized independently. This randomization and the aggregation of many trees leads to the name random forest. The partition of the feature space $[0,1]^p$ created by a randomized regression tree is hierarchical and its structure is linked to the tree structure. Each tree node corresponds to a set in the hierarchical partition. The root node of the tree corresponds to the entire feature space, which is at the top of the hierarchy. The leaf nodes correspond to the sets at the bottom of the partition hierarchy. The tree is constructed iteratively by splitting each set corresponding to a current leaf node that does not yet satisfy a termination criterion into two new sets. The two sets resulting from each split set correspond to the child nodes of the original node and form new leaves of the tree. The splitting of the sets is usually done orthogonal to an axis of the feature space. Thus, all the sets are hyperrectangles, also called cells. An example of a partition and the corresponding tree structure in the case $p = 2$ can be seen

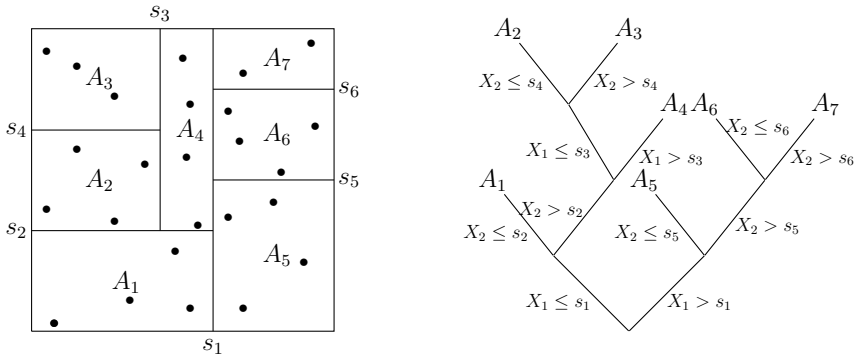

Figure 1.1: Hierarchical partition of a regression tree with corresponding tree structure.

in Figure 1.1. The first splitting in the figure is done orthogonal to the horizontal axis at s_1, the second splitting is for the left cell orthogonal to the second axis at s_2 and so on. For the trees forming a random forest, the rule for choosing the feature and its splitting point involves randomness and may also depend on observations from the training sample. The partitioning method is the central and most complex tuning parameter of a random forest.

After termination of the tree construction, the partition of the feature space formed by the leaves of the tree is used for the estimation of m. In a regression tree each leaf usually contains at least one of the data points. The estimate of the regression function at an arbitrary point x_0 is the average response value of the observations in the hyperrectangle/leaf that x_0 falls into. We recall that each tree only uses a subset of the training sample for the piecewise constant estimation in the partition.

Let $I \subset \{1, \ldots, n\}$ denote an index set of a subsample of the training data. Further, let $A_n(x_0)$ denote the cell in the final partition of the tree that contains a fixed x_0. We omit the dependence of this set on the data and the randomization for simplicity. The regression tree estimator at x_0 is defined as

$$\tilde{m}_n(x_0) := \sum_{j \in I} Y_j \frac{\mathbb{I}\{X_j \in A_n(x_0)\}}{\sum_{l \in I} \mathbb{I}\{X_l \in A_n(x_0)\}}.$$

The sum in the denominator is equal to the number of observations in I that fall into the cell $A_n(x_0)$. Thus, the estimator takes the average of all Y_i for which the X_i is in the same cell of the feature space partition as the argument x_0. One obtains the final random forest estimator by averaging multiple regression trees.

A good starting point to discover the literature on random forests is provided in the review article by Biau and Scornet (2016). In general, the literature covers different aspects of the method. There are extensions of the original algorithm, such as quantile regression forests covered by Meinshausen (2006), random survival forests introduced by Ishwaran et al. (2008), generalized random forests covered by Athey et al. (2019), or neural random forests by Biau et al. (2019). Alternative data dependent splitting procedures involving more or less randomness have been studied, for instance, by Zhang et al. (2024).

There are several consistency results for random forests in the literature. Some are for the standard random forest by Breiman (2001), and others are more general or apply

to different types of random forests. Examples are the results by Biau et al. (2008) or Scornet et al. (2015), who prove consistency of Breiman's forest in an additive regression model. Chi et al. (2022) show consistency in a high-dimensional regression model. Scornet (2016a) analyzes the relationship between random forests and their infinite version, i.e., a random forest with infinitely many trees, and links the consistency of the finite forest to the infinite.

Multiple results in the literature regarding the asymptotic distribution of random forests provide asymptotic normality, see for example in the work of Mentch and Hooker (2016), Wager and Athey (2018) or Peng et al. (2022). All these results have in common that they rely on the interpretation of a random forest as a generalized U-statistic, so that the Hoeffding-decomposition and the Hájek projection can be utilized to analyze the asymptotic behavior. Furthermore, in all the results, the variance of the asymptotic normal distribution is either unknown or described abstractly. Mentch and Hooker (2016), as well as Peng et al. (2022), provide results that are actually for generalized U-statistics and are not applied to a specific random forest. Thus, for all the results, there is no formula that describes the effect of the density of the covariates and the variance of the errors on the asymptotic variance of the estimator. Nevertheless, the aforementioned results allow to infer the asymptotic distribution, if an estimator for the asymptotic variance is provided. This is done, for instance, by Mentch and Hooker (2016) or Wager and Athey (2018), and more recently by Xu et al. (2024), who introduce an unbiased estimator of the variance of a random forest in U-statistic form. The estimator is based on the Hoeffding-decomposition and includes terms of different orders, not just the first order terms, as is usually done when analyzing the asymptotic distribution without variance estimation.

A substantial amount of literature on random forests deals with purely random forests, in which the partitions of the trees do not depend on the training sample. A popular version are centered purely random forests that are characterized by always splitting each hyperrectangle in its center orthogonal to a random coordinate. Breiman (2004) introduced centered purely random forests and proved their pointwise consistency. Later results dealing with this type of random forest are by Biau (2012) or Klusowski (2021) and give explicit rates for the mean squared error. Note that the three aforementioned articles on purely random forests do not incorporate subsampling for the individual regression trees, but instead use the entire training sample for each tree, which is a noteworthy distinction from our work. The main reason to consider purely random forests is that one usually faces less technical difficulties when proving statistical guarantees.

This is the reason why we will consider this type of random forest in our work. We already mentioned the lack of results regarding an explicit limit distribution of any random forest type. Further there are also no results that allow for the construction of uniform asymptotic confidence bands that we are aware of. We want to close this gap and prove both these results for centered purely random forests.

Confidence bands in nonparametric regression

Despite the common tendency to categorize a random forest as a machine learning algorithm, it is a nonparametric regression estimator. This is especially true for purely random forests, that do not exploit the training data in their partition creation. A substantial amount of literature exists on the subject of confidence bands for different nonparametric

density and regression estimators. The literature on density estimation is also relevant because it shares similar proof techniques with the literature on regression. Most of the early results in the literature were for the univariate case. One of the first results is by Smirnov (1950) and gives confidence bands for a probability density based on a histogram estimator.

Bickel and Rosenblatt (1973) study confidence bands for kernel density estimators. The article is widely cited and their proof technique has subsequently been applied to construct confidence bands in density and regression estimation. Giné and Nickl (2010) investigated confidence bands for general density estimators that adapt to the degree of smoothness of the density function. Their results can for instance be applied to wavelet or kernel density estimators.

There are several articles that construct confidence bands in regression models. Johnston (1982) constructed them for the Nadaraya-Watson kernel estimator, Härdle (1989) showed asymptotic uniform confidence bands for a wider class of regression estimators, the M-smoothers. Typically, the confidence bands rely on an undersmoothing of the estimator to reduce the bias relative to the stochastic error. An alternative is a direct bias correction, which is used, for instance, by Eubank and Speckman (1993), who considered a deterministic, uniform design for local constant regression estimation, and in the article by Xia (1998), who considered a random design under dependence and used local linear estimation. There are also bootstrap confidence bands for nonparametric regression, see for instance Hall (1993), Neumann and Polzehl (1998), Claeskens and Van Keilegom (2003) or Hall and Horowitz (2013). Sabbah (2014) showed confidence bands for quantile estimators which is an alternative to a regression model and Birke et al. (2010) proved confidence bands in an inverse regression model.

The aforementioned articles only considered univariate regression. In the multivariate case Konakov and Piterbarg (1984) analyzed the asymptotic distribution of the maximal deviation for the Nadaraya-Watson estimate in a random design setting. Proksch (2016) proved confidence bands in a multivariate regression model with fixed, deterministic design for a general class of estimators that includes, for example, local polynomial estimators. Chao et al. (2017) covered confidence bands for multivariate quantile regression.

Own Contributions

The central contribution of this thesis are explicit asymptotic confidence intervals and bands for centered purely random forests. Explicit means that one does not need an estimator for the variance of the random forest to get the radius of the confidence intervals or bands. Instead, the only objects that might be unknown, dependent on the assumptions in the regression model, and affect the radius are the density of the covariates and the variance of the errors. If these two quantities are assumed to be unknown in the nonparametric regression model, they can be estimated independently of the estimation of the regression function. Knowing the density of the covariates and the variance of the errors, the radius and the asymptotic variance can be calculated numerically by means of a Monte Carlo simulation. To the best of our knowledge, there are no explicit results regarding the asymptotic normal distribution and no uniform confidence bands for any random forest type in the literature. The explicit form of the results can be useful in practice because it does not rely on the direct estimation of the variance of the estimator.

Another noteworthy aspect of the asymptotic confidence bands is the fact that they do not depend on an extreme value limit distribution. Instead, the quantiles used depend on the depth of the trees in the random forest. This is advantageous for two reasons. The first is that we do not need to know the limit distribution and the second is that the convergence to the limit distribution may be slow and thereby would come with an additional error term.

As explained above, uniform confidence bands are desirable for inference in practical applications. Thus, the lack of confidence bands for random forests is a gap in the literature that is important to fill. The contribution of confidence bands for a simple random forest type is a first step towards filling this gap. An extension of the results to more sophisticated, data dependent types of random forests applied in practice can hopefully be provided by future research.

Our asymptotic results are valid for two specific types of centered purely random forests. One of them is the classically considered version with uniformly distributed splitting coordinates, see Breiman (2004). The other type is newly proposed in our work and is called the Ehrenfest centered purely random forest, because it is based on the Ehrenfest model for diffusion. The novelty of this algorithm is to link the partition construction, which is the main tuning parameter of the forest, to the Ehrenfest model. This new partition construction method leads to better results than the classical case with uniformly distributed splits. In particular, we can prove that the Ehrenfest centered purely random forest reaches the minimax optimal rate of the mean squared error in nonparametric regression problems with Hölder continuous regression functions, which is not possible for the classical uniform centered purely random forest, see Klusowski (2021).

The uniform confidence band result, or rather the auxiliary results used to prove it, further allow us to prove uniform convergence of both centered purely random forests. As far as we are aware, all previous consistency results for random forests are pointwise. In addition, we construct confidence bands for the histogram regression estimator based on the same proof technique that we used for the confidence bands for the purely random forest. We are not aware of a similar result for the histogram estimator in multivariate regression in the literature, especially one that does not depend on an extreme value limit distribution.

The pointwise asymptotic normality leading to the confidence intervals can be proved by using a central limit theorem for generalized incomplete U-statistics by Peng et al. (2022). However, our results leading to the construction of confidence bands yield the same asymptotic distribution under less restrictive assumptions. Namely, that the subsample size r_n of the trees can have the same rate as the sample size, which is usually not possible in the central limit theorems for random forests in the literature.

The uniform results, and in particular the confidence bands, require a more sophisticated proof technique than the pointwise results. One component of the proof that demanded a novel and more complex approach is the uniform bound for the remainder terms that emerge from the Hájek projection of a generalized U-statistic.

The dominating term of the asymptotic distribution also requires a uniform proof technique that handles the distribution of the supremum of an empirical process asymptotically. A typical proof technique used for nonparametric regression estimators relies on the uniform approximation of an empirical process by a Gaussian process and has been

applied, for instance, by Johnston (1982), Claeskens and Van Keilegom (2003), Proksch (2016), or Chao et al. (2017). Our proof technique is more direct because, instead of uniformly approximating the entire process, it approximates the supremum of the empirical process by the supremum of a Gaussian process owing to a result by Chernozhukov et al. (2014b). This more direct approximation of the supremum has been employed, for instance, by Patschkowski and Rohde (2019), to construct adaptive confidence bands for probability densities.

The direct approximation of the supremum by the supremum of a Gaussian process has a smaller error than the uniform approximation of the entire process, at least with the uniform approximation results currently available in the literature that we are aware of. Utilizing the available uniform approximation results, the confidence results in this thesis would not be possible. The larger error of the Gaussian approximation would contradict the assumptions on the bias term, which we will later call the approximation error. This prevents the construction of confidence bands based on a classical undersmoothing assumption.

Substantial parts of this thesis were published in the preprints Neumeyer et al. (2025) and Neumeyer et al. (2026) which are available on arXiv and have been submitted to peer reviewed journals.

Structure of the thesis

This dissertation is structured as follows. First, in Chapter 2 some foundations for the rest of the work are given. These include the nonparametric regression model accompanied by an introduction of the histogram estimator, a selection of key definitions and relevant results from the literature. The most important of these is a result for the Gaussian approximation of the supremum of an empirical process by Chernozhukov et al. (2014b). Further, the chapter covers an introduction into generalized incomplete U-statistics and some results regarding the Binomial distribution. Chapter 3 introduces random forest regression estimators with a focus on centered purely random forests and the connection to generalized incomplete U-statistics. With these foundations at hand, we derive a bound for the mean squared error and a pointwise central limit theorem for centered purely random forests in Chapter 4. Uniform results are addressed in Chapter 5, most notably confidence band construction and uniform convergence for centered purely random forests, so that Chapter 4 and Chapter 5 contain the main contributions of the thesis. The confidence band results are underlined by a simulation study in Chapter 6 that investigates the empirical performance of the asymptotic confidence bands in a selection of settings. In Chapter 7, we conclude the thesis with an approach for the extension of the results to a data dependent centered random forest, a description of the classical proof structure for confidence bands compared to the proof structure in Chapter 5 and a discussion of future research.

Chapter 2

Foundations

The sections in this chapter cover different topics that will all be used throughout the thesis. In Section 2.1 we introduce the nonparametric regression model and discuss the approximation-variance tradeoff in nonparametric regression using the histogram estimator as an example. Section 2.2 contains several important results from the literature and in Section 2.3 we give a brief introduction to U-statistics, including results we will need in Chapter 4. We conclude the chapter with a collection of results on the moments of Binomial distributed denominators in Section 2.4.

2.1 The nonparametric regression model

We introduce the nonparametric regression framework which we will consider throughout the thesis. On some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ which we will consider throughout, we observe the training sample $\mathcal{D}_n = (X_i, Y_i)_{i=1}^n$ consisting of n observations which are i.i.d. copies of the random variable (X, Y) with $X \in [0,1]^p$ and $Y \in \mathbb{R}$. The goal is to estimate the regression function $m : [0,1]^p \to \mathbb{R}$ with

$$Y = m(X) + \varepsilon \tag{2.1}$$

for an observation error ε that is independent of X and satisfies $\mathbb{E}[\varepsilon] = 0$. We note that $m(x) = \mathbb{E}[Y|X = x]$. We denote $Z_i = (X_i, Y_i)$ and $Z = (X, Y)$ for a generic random variable with the joint distribution F_Z. It holds that $(X, \varepsilon) = (X, Y - m(X))$ which is a function of Z. We call $[0,1]^p$ the feature space and each $x \in [0,1]^p$ is a vector with p entries called coordinates or features.

Standard estimators are smooth kernel or spline estimators. A relatively simple estimator in the above regression model is the histogram estimator. It is not smooth but piecewise constant and usually other estimators are preferred in practice. Nonetheless it is of interest for our work because it has some similarities to random forests, for example the piecewise constancy. Later, we will see that the proof technique that gives us confidence bands for random forests can also be applied to the histogram in a less technical and complex way.

The histogram estimator uses the cell wise average of the response variables in a grid of the feature space as a piece-wise constant estimation of m. The spacing of the grid is usually chosen dependent on n and we will assume that the grid is equidistant. For

$\delta^{-1} \in \mathbb{N}$ we consider the grid with edge length δ on $[0,1]^p$. For $x_0 \in [0,1]^p$ let $A_\delta(x_0)$ be the hypercube, alternatively referred to as a cell, that contains x_0. We define the histogram estimator

$$\hat{m}_H(x_0) = \sum_{j=1}^{n} Y_j \frac{\mathbb{I}\{X_j \in A_\delta(x_0)\}}{\sum_{i=1}^{n} \mathbb{I}\{X_i \in A_\delta(x_0)\}}. \tag{2.2}$$

If there is no observation in $A_\delta(x_0)$ we define the estimator as zero on this cell. Let

$$\hat{m}_H^{(m)}(x_0) = \sum_{j=1}^{n} m(X_j) \frac{\mathbb{I}\{X_j \in A_\delta(x_0)\}}{\sum_{i=1}^{n} \mathbb{I}\{X_i \in A_\delta(x_0)\}}$$

and

$$\hat{m}_H^{(\varepsilon)}(x_0) = \sum_{j=1}^{n} \varepsilon_j \frac{\mathbb{I}\{X_j \in A_\delta(x_0)\}}{\sum_{i=1}^{n} \mathbb{I}\{X_i \in A_\delta(x_0)\}}.$$

Using these definitions and (2.1) the error of the estimator can be decomposed as

$$\hat{m}_H(x_0) - m(x_0) = \hat{m}_H^{(m)}(x_0) - m(x_0) + \hat{m}_H^{(\varepsilon)}(x_0). \tag{2.3}$$

We call $\hat{m}_H^{(m)} - m$ the approximation error and $\hat{m}_H^{(\varepsilon)}$ the stochastic error. A similar decomposition of the error can be done for different regression estimators, for example kernel estimators and random forests as we will see in Chapter 3. It holds that

$$\hat{m}_H^{(m)}(x_0) = \mathbb{E}\left[\hat{m}_H(x_0) \mid (X_j)_{j=1}^n\right] \tag{2.4}$$

which illustrates that given the independent observations $(X_j)_{j=1}^n$ the approximation error is indeed caused by the ability of the piecewise constant estimator to approximate m. Further note that the expectation of the approximation error is the bias of the estimator. For any $A \subset \mathbb{R}^p$ let us denote its diameter

$$\mathfrak{d}(A) := \sup_{x_1, x_2 \in A} \|x_1 - x_2\|, \tag{2.5}$$

where $\|\cdot\|$ is the Euclidean norm. Suppose that there exists at least one observation in $A_\delta(x_0)$ and that m is α-Hölder continuous for $\alpha \in (0,1]$ and Hölder constant C_H. For any $q \geq 1$ we obtain the upper bound

$$\begin{aligned}
\mathbb{E}\left[|\hat{m}_H^{(m)}(x_0) - m(x_0)|^q\right] &= \mathbb{E}\left[\Big|\sum_{j=1}^{n} (m(X_j) - m(x_0)) \frac{\mathbb{I}\{X_j \in A_\delta(x_0)\}}{\sum_{i=1}^{n} \mathbb{I}\{X_i \in A_\delta(x_0)\}}\Big|^q\right] \\
&\leq C_H^q \mathbb{E}\left[\Big|\sum_{j=1}^{n} \|X_j - x_0\|^\alpha \frac{\mathbb{I}\{X_j \in A_\delta(x_0)\}}{\sum_{i=1}^{n} \mathbb{I}\{X_i \in A_\delta(x_0)\}}\Big|^q\right] \\
&\leq C_H^q \mathfrak{d}(A_\delta(x_0))^{q\alpha},
\end{aligned} \tag{2.6}$$

which demonstrates that the approximation error profits from a smaller diameter of the grid cells. Let $\mathbb{E}[\varepsilon^2] = \sigma^2 < \infty$ and assume that the density f_X of X is bounded away from

zero and from above. For the variance of the stochastic error we will prove in Chapter 5, within the proof of Theorem 5.10, that

$$\operatorname{Var}\left(\hat{m}_H^{(\varepsilon)}(x_0)\right) = \Theta(n^{-1}\delta^{-p})$$

if $n\delta^p \to \infty$. This shows that the stochastic error profits from a larger δ, which is caused by larger cells leading to more ε_j being averaged and reducing the variance. The direct relation between the diameter and δ is $\mathfrak{d}(A_\delta(x_0)) = \delta\sqrt{p}$. Thus, we face a tradeoff between the two error parts, where the approximation error decreases and the stochastic error increases as δ decreases. This phenomenon is common in nonparametric regression problems. For smooth kernel estimators the tradeoff is in the bandwidth which is a comparable parameter to δ for the histogram.

2.2 Empirical process and supplementary results

In this section we gather some results from the literature that will be of use throughout the thesis. The first result by Chernozhukov et al. (2014b) will be important for the proof of our main results. Before we give the result we need several definitions, starting with empirical processes.

Definition 2.1 (Empirical process)**.** Let $(S, \mathcal{S})$ be a measurable space and for $j = 1, \dots, n$ let $Z_j : (\Omega, \mathcal{A}, \mathbb{P}) \to (S, \mathcal{S})$ be i.i.d. random variables with distribution P. For a function class $\mathcal{F}$ of measurable functions $f : S \to \mathbb{R}$,

$$\mathbb{G}_n f = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (f(Z_j) - \mathbb{E}[f(Z_1)]), \qquad f \in \mathcal{F}$$

is an empirical process in its standardized form.

Most results for empirical processes need assumptions on the structure of the function class $\mathcal{F}$. In our work we will always consider a finite function class that trivially satisfies the necessary assumptions. Nevertheless, we include the following definitions for completeness.

Definition 2.2 (Covering and packing number)**.** On an arbitrary semimetric space (T, d), the covering number $N(T, d, \varepsilon)$ is the minimal number of balls $B_\varepsilon(s) = \{t \in T \mid d(s,t) < \varepsilon\}$ that covers T. The packing number $D(T, d, \varepsilon)$ is the maximum number of points whose pairwise distance is strictly larger than ε. In the context of empirical processes the semimetric space is $(\mathcal{F}, d)$ for a semimetric d on the function class $\mathcal{F}$. For the balls $B_\varepsilon(g) = \{f \in \mathcal{F} \mid d(g,f) < \varepsilon\}$ it is not always necessary that $g \in \mathcal{F}$ but it is sufficient here.

For any probability measure Q on a measurable space $(S, \mathcal{S})$ we denote $Qf := \int f dQ$. Further $\mathcal{L}^q(Q)$ denotes the space of all measurable functions with $\|f\|_{Q,q} := (Q|f|^q)^{1/q} < \infty$ for any $q \geq 1$. Let us denote the $\mathcal{L}^2(Q)$-semimetric

$$e_Q(f,g) := \|f - g\|_{Q,2} = \sqrt{Q(f-g)^2}.$$

For a function class $\mathcal{F}$, an envelope is a non-negative function $F : S \to \mathbb{R}$ with $F(x) \geq \sup_{f \in \mathcal{F}} |f(x)|$ for all $x \in S$.

Definition 2.3 (VC type class)**.** On a measurable space $(S, \mathcal{S})$, let $\mathcal{F}$ be a class of measurable functions with a measurable envelope F. Let $\mathcal{Q}$ denote the set of all finitely discrete probability measures on $(S, \mathcal{S})$. The function class $\mathcal{F}$ is called VC (Vapnik-Chervonenkis) type with envelope F if there exist constants $A, v > 0$ such that

$$\sup_{Q \in \mathcal{Q}} N(\mathcal{F}, e_Q, \varepsilon \|F\|_{Q,2}) \leq (A/\varepsilon)^v \qquad \text{for all } \varepsilon \in (0, 1].$$

The next theorem is Corollary 2.2 by Chernozhukov et al. (2014b). It is used to approximate the supremum of an empirical process.

Theorem 2.4. *Suppose that $\mathcal{F}$ is a pointwise measurable function class with envelope F that is VC type for constants $A \geq e, v \geq 1$. Let P denote the distribution of Z_1 from Definition 2.1 and let G_P be a centered Gaussian process indexed in $\mathcal{F}$ with*

$$\mathbb{E}[G_P(f)G_P(g)] = P(fg) = \mathbb{E}[f(Z_1)g(Z_1)].$$

Suppose also that for some $b \geq \tilde{\sigma} > 0$, and $\nu \in [4, \infty]$, we have $\sup_{f \in \mathcal{F}} P|f|^k \leq \tilde{\sigma}^2 b^{k-2}$ for $k = 2, 3$ and $\|F\|_{P,\nu} \leq b$. Let $\mathbb{S}_n = \sup_{f \in \mathcal{F}} \mathbb{G}_n f$. Then for every $\gamma \in (0, 1)$ there exists a random variable $\mathbf{S} \stackrel{d}{=} \sup_{f \in \mathcal{F}} G_P f$ such that

$$\mathbb{P}\left(|\mathbb{S}_n - \mathbf{S}| > \frac{bK_n}{\gamma^{1/2} n^{1/2 - 1/\nu}} + \frac{(b\tilde{\sigma})^{1/2} K_n^{3/4}}{\gamma^{1/2} n^{1/4}} + \frac{(b\tilde{\sigma}^2 K_n^2)^{1/3}}{\gamma^{1/3} n^{1/6}} \right) \leq C\left(\gamma + \frac{\log n}{n}\right)$$

where $K_n = cv(\log n \vee \log(Ab/\tilde{\sigma}))$, and $c, C > 0$ are constants that depend only on ν ("$1/\nu$" is interpreted as "0" when $\nu = \infty$).

The following result by van der Vaart (1998) will be used in Section 2.3. It allows to prove that the Hájek projection of a U-statistic is the asymptotically leading term if its variance divided by the variance of the U-statistic converges to one.

Theorem 2.5 (van der Vaart (1998, Theorem 11.2))**.** *Let $\mathcal{S}_n$ be spaces of random variables with finite second moments that satisfy the following two conditions. For all $a, b \in \mathbb{R}$ and $S_1, S_2 \in \mathcal{S}_n$ it holds that $aS_1 + bS_2 \in \mathcal{S}_n$ and $a \in \mathcal{S}_n$. Let T_n be random variables and suppose*

$$\hat{S}_n := \operatorname*{arg\,min}_{S \in \mathcal{S}_n} \mathbb{E}\left[(T_n - S)^2\right].$$

If $\mathrm{Var}T_n / \mathrm{Var}\hat{S}_n \to 1$ then

$$\frac{T_n - \mathbb{E}T_n}{\sqrt{\mathrm{Var}T_n}} - \frac{\hat{S}_n - \mathbb{E}\hat{S}_n}{\sqrt{\mathrm{Var}\hat{S}_n}} \xrightarrow{\mathbb{P}} 0.$$

We need the Khintchine inequality below, see for instance Chow and Teicher (1997, Section 10.3, Theorem 1), to prove the last result of this section.

Theorem 2.6 (Khintchine Inequality)**.** *If $(V_i)_{i \in \mathbb{N}}$ are i.i.d. random variables with $\mathbb{P}(V_1 = 1) = \mathbb{P}(V_1 = -1) = 1/2$ and c_i are any real numbers, then for every $q \in (0, \infty)$ there exist positive, finite constants A_q and B_q such that*

$$A_q \left(\sum_{i=1}^n c_i^2 \right)^{1/2} \leq \mathbb{E}\left[\left| \sum_{i=1}^n c_i V_i \right|^q \right]^{1/q} \leq B_q \left(\sum_{i=1}^n c_i^2 \right)^{1/2}.$$

The Marcinkiewicz-Zygmund inequality below will be employed repeatedly in this thesis. In particular, we will frequently use an implication of the result which we add as a second statement of the proposition. The first statement and its proof can be found as Theorem 2 in Chow and Teicher (1997, Section 10.3).

Proposition 2.7 (Marcinkiewicz-Zygmund Inequality)**.** *Let $q \in [1, \infty)$, for centered and independent random variables $(X_i)_{i=1}^n$ it holds that there exist positive constants $C_{q,1}$ and $C_{q,2}$ only dependent on q such that*

$$C_{q,1}\mathbb{E}\left[\left(\sum_{i=1}^n X_i^2\right)^{q/2}\right] \leq \mathbb{E}\left[\left|\sum_{i=1}^n X_i\right|^q\right] \leq C_{q,2}\mathbb{E}\left[\left(\sum_{i=1}^n X_i^2\right)^{q/2}\right].$$

Let $(Y_i)_{i=1}^n$ be random variables that are independent of $(X_i)_{i=1}^n$ but not necessarily independent of each other. Assume that $\mathbb{E}[g(X_i, Y_i) \mid Y_i] = 0$, then it holds that

$$\mathbb{E}\left[\left|\sum_{i=1}^n g(X_i, Y_i)\right|^q\right] \leq C_{q,2}\mathbb{E}\left[\left(\sum_{i=1}^n g^2(X_i, Y_i)\right)^{q/2}\right]. \tag{2.7}$$

Remark 2.8. If we additionally assume that the $(X_i)_{i=1}^n$ are identically distributed and $q \geq 2$ we further have

$$\mathbb{E}\left[\left(\sum_{i=1}^n X_i^2\right)^{q/2}\right] \leq n^{q/2}\mathbb{E}\left[|X_1|^q\right].$$

Hölder's inequality yields

$$\sum_{i=1}^n 1 \cdot X_i^2 \leq n^{1-2/q}\left(\sum_{i=1}^n |X_i|^q\right)^{2/q} = \left(n^{q/2-1}\sum_{i=1}^n |X_i|^q\right)^{2/q}$$

and thus the above statement follows directly.

Proof of Proposition 2.7. Let $\mathcal{L}^q$ denote the space of all random variables X with $\mathbb{E}[|X|^q] < \infty$. Due to the equivalence

$$X_i \in \mathcal{L}^q \; \forall i \in [n] \Leftrightarrow \sum_{i=1}^n X_i \in \mathcal{L}^q \Leftrightarrow \left(\sum_{i=1}^n X_i^2\right)^{1/2} \in \mathcal{L}^q$$

we assume that $X_i \in \mathcal{L}^q$ for all $i \in [n]$ for the rest of the proof. Otherwise the assertion holds trivially since all expectations are equal to infinity. Let $\tilde{X}_i$ be i.i.d. copies of the X_i. Further let $(V_i)_{i=1}^n$ be i.i.d. random variables independent of $(X_i)_{i=1}^n$ and $(\tilde{X}_i)_{i=1}^n$ with $\mathbb{P}(V_1 = 1) = \mathbb{P}(V_1 = -1) = 1/2$. We note that

$$\mathbb{E}\left[\sum_{i=1}^n V_i(X_i - \tilde{X}_i) \,\middle|\, (V_i)_{i=1}^n, (X_i)_{i=1}^n\right] = \sum_{i=1}^n V_i X_i$$

because $\mathbb{E}[\tilde{X}_i] = 0$. This statement, Jensen's inequality for the conditional expectation, the triangle inequality and the convexity of $|\cdot|^q$ for $q \geq 1$ yield

$$\begin{aligned}
\mathbb{E}\left[\Big|\sum_{i=1}^n V_i X_i\Big|^q\right] &= \mathbb{E}\left[\Big|\mathbb{E}\left[\sum_{i=1}^n V_i(X_i - \tilde{X}_i) \;\Big|\; (V_i)_{i=1}^n, (X_i)_{i=1}^n\right]\Big|^q\right] \\
&\leq \mathbb{E}\left[\Big|\sum_{i=1}^n V_i(X_i - \tilde{X}_i)\Big|^q\right] \\
&\leq 2^q \mathbb{E}\left[\left(\frac{1}{2}\Big|\sum_{i=1}^n V_i X_i\Big| + \frac{1}{2}\Big|\sum_{i=1}^n V_i \tilde{X}_i\Big|\right)^q\right] \\
&\leq 2^{q-1}\mathbb{E}\left[\Big|\sum_{i=1}^n V_i X_i\Big|^q + \Big|\sum_{i=1}^n V_i \tilde{X}_i\Big|^q\right] \\
&= 2^q \mathbb{E}\left[\Big|\sum_{i=1}^n V_i X_i\Big|^q\right]. \qquad (2.8)
\end{aligned}$$

By conditioning on $(X_i)_{i=1}^n$ Theorem 2.6 implies

$$A_q^q \left(\sum_{i=1}^n X_i^2\right)^{q/2} \leq \mathbb{E}\left[\Big|\sum_{i=1}^n V_i X_i\Big|^q \;\Big|\; (X_i)_{i=1}^n\right] \leq B_q^q \left(\sum_{i=1}^n X_i^2\right)^{q/2}.$$

Taking the expectation and using (2.8) we obtain

$$\begin{aligned}
A_q^q \mathbb{E}\left[\left(\sum_{i=1}^n X_i^2\right)^{q/2}\right] &\leq \mathbb{E}\left[\Big|\sum_{i=1}^n V_i(X_i - \tilde{X}_i)\Big|^q\right] \\
&\leq 2^q \mathbb{E}\left[\Big|\sum_{i=1}^n V_i X_i\Big|^q\right] \leq 2^q B_q^q \mathbb{E}\left[\left(\sum_{i=1}^n X_i^2\right)^{q/2}\right]. \qquad (2.9)
\end{aligned}$$

Similar to the calculation in (2.8) we obtain

$$\begin{aligned}
\mathbb{E}\left[\Big|\sum_{i=1}^n X_i\Big|^q\right] &= \mathbb{E}\left[\Big|\mathbb{E}\left[\sum_{i=1}^n (X_i - \tilde{X}_i) \;\Big|\; (X_i)_{i=1}^n\right]\Big|^q\right] \\
&\leq \mathbb{E}\left[\Big|\sum_{i=1}^n (X_i - \tilde{X}_i)\Big|^q\right] \\
&\leq 2^q \mathbb{E}\left[\Big|\sum_{i=1}^n X_i\Big|^q\right]
\end{aligned}$$

The symmetry of $X_i - \tilde{X}_i$ implies that $(X_i - \tilde{X}_i) \stackrel{d}{=} V_i(X_i - \tilde{X}_i)$ for all $i \in \mathbb{N}$ and hence we get

$$\mathbb{E}\left[\Big|\sum_{i=1}^n X_i\Big|^q\right] \leq \mathbb{E}\left[\Big|\sum_{i=1}^n V_i(X_i - \tilde{X}_i)\Big|^q\right] \leq 2^q \mathbb{E}\left[\Big|\sum_{i=1}^n X_i\Big|^q\right].$$

The above and (2.9) yield

$$A_q^q \mathbb{E}\left[\left(\sum_{i=1}^n X_i^2\right)^{q/2}\right] \leq 2^q \mathbb{E}\left[\left|\sum_{i=1}^n X_i\right|^q\right] \leq 2^{2q} B_q^q \mathbb{E}\left[\left(\sum_{i=1}^n X_i^2\right)^{q/2}\right]$$

and dividing by 2^q the first assertion holds for $C_{q,1} = A_q/2$ and $C_{q,2} = 2B_q$.

It remains to prove the second assertion. Conditioned on $(Y_i)_{i=1}^n$ the random variables $g(X_i, Y_i)$ are independent and centered. We use the first inequality and get

$$\begin{aligned}
\mathbb{E}\left[\left|\sum_{i=1}^n g(X_i, Y_i)\right|^q\right] &= \mathbb{E}\left[\mathbb{E}\left[\left|\sum_{i=1}^n g(X_i, Y_i)\right|^q \,\middle|\, (Y_i)_{i=1}^n\right]\right] \\
&\leq C_{q,2}\mathbb{E}\left[\mathbb{E}\left[\left(\sum_{i=1}^n g^2(X_i, Y_i)\right)^{q/2} \,\middle|\, (Y_i)_{i=1}^n\right]\right] \\
&= C_{q,2}\mathbb{E}\left[\left(\sum_{i=1}^n g^2(X_i, Y_i)\right)^{q/2}\right]. \qquad \square
\end{aligned}$$

2.3 U-statistics

An important concept that is essential for our theory and many proofs in the thesis are U-statistics. In this section, we will introduce them and also cover some extensions of the classical case that are necessary for our applications. Peng et al. (2022) previously considered the same combination of extensions, as they enable the interpretation of a random forest as a U-statistic. The individual extensions have been introduced in earlier literature. The proofs of Theorem 2.12 and Theorem 2.13 below can be found in the aforementioned article. However, they are reproduced here for the sake of completeness and to allow for a comparison with subsequent proof methods.

Suppose there are i.i.d. random variables $Z_1, \ldots, Z_n$ from distribution F_Z. Let $\vartheta(F_Z)$ be a parameter of this distribution such that there is a function h that is permutation-invariant in its r arguments and satisfies

$$\mathbb{E}[h(Z_1, \ldots, Z_r)] = \vartheta(F_Z).$$

We use $B_{r,n}$ to denote the set of subsets of $\{1, \ldots, n\}$ with r elements. The U-statistic of order r

$$\begin{aligned}
U_{n,r} &= \binom{n}{r}^{-1} \sum_{\{i_1,\ldots,i_r\}\in B_{r,n}} h(Z_{i_1}, \ldots, Z_{i_r}) \\
&= \binom{n}{r}^{-1} \sum_{I\in B_{r,n}} h((Z_i)_{i\in I})
\end{aligned}$$

is an unbiased estimator for $\vartheta(F_Z)$ and its theory goes back to Halmos (1946) and Hoeffding (1948). Under additional assumptions on the class of distributions, the U-statistic

is the minimum variance unbiased estimator. The function h is the so called kernel of the U-statistic and r is its order. There is a lot of theory for U-statistics, a good introduction is the book by Lee (1990).

We will use a more general version of the standard U-statistic. The first generalization we need is that the order r_n depends on the sample size n and $r_n \to \infty$. This was introduced by Frees (1989) and is called infinite order U-statistic. This implies that the kernel depends on n at least by the number of arguments. If the sample size n is large it can be computationally infeasible to include all $\binom{n}{r_n}$ subsamples in the average. In this case, one usually uses a smaller number of terms or subsets, respectively. An introduction to these incomplete U-statistics can be found in Lee (1990, Section 4.3), where different options to select the included subsets are discussed. Lastly we assume that the kernel depends on another random variable ω that is independent of the Z_i. This additional randomization will be necessary to get a U-statistic interpretation of a random forest. For each summand in the U-statistic or for each subset of size r_n respectively one uses an i.i.d. copy of ω. These three generalizations are separate concepts and in principle it is not required to combine them, but it is nevertheless necessary for our purpose. The following definition follows Definition 1 by Peng et al. (2022), but the same object under a different name is for instance used by Mentch and Hooker (2016).

Definition 2.9 (Generalized U-statistic)**.** Let $(Z_i)_{i=1}^n$ and ω_I for $I \in B_{r_n,n}$ be i.i.d. sequences of random variables. Let $N \in \mathbb{N}$ with $N \leq \binom{n}{r_n}$ and for $I \in B_{r_n,n}$ let ρ_I be i.i.d. Bernoulli random variables with $\mathbb{P}(\rho_I = 1) = N/\binom{n}{r_n}$. Assume that all i.i.d. sequences are independent of each other. Let h_n denote a real valued function utilizing r_n of the Z_i and one of the ω_I. For

$$\hat{N} = \sum_{I \in B_{r_n,n}} \rho_I \tag{2.10}$$

a generalized U-statistic is an estimator of the form

$$U_{n,r_n,N,\omega} = \frac{1}{\hat{N}} \sum_{I \in B_{r_n,n}} \rho_I h_n((Z_i)_{i \in I}, \omega_I). \tag{2.11}$$

If $N < \binom{n}{r_n}$ the generalized U-statistic is incomplete and if $N = \binom{n}{r_n}$ the estimator is a generalized complete U-statistic equal to

$$U_{n,r_n,\omega} = \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} h_n((Z_i)_{i \in I}, \omega_I).$$

$\hat{N}$ is the actual number of subsamples that are included in the U-statistic. It is a random variable with $\mathbb{E}[\hat{N}] = N$. This ensures that the random weights sum up to one. We have

$$\begin{aligned} \mathbb{E}\left[\frac{\hat{N}}{N} U_{n,r_n,N,\omega} \,\middle|\, (Z_i)_{i=1}^n, (\omega_I)_{I \in B_{r_n,n}}\right] &= \frac{1}{N} \sum_{I \in B_{r_n,n}} \mathbb{E}\left[\rho_I\right] h_n((Z_i)_{i \in I}, \omega_I) \\ &= \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} h_n((Z_i)_{i \in I}, \omega_I) = U_{n,r_n,\omega}. \end{aligned}$$

We use ω to denote a single independent copy of the ω_I. Let us denote $\vartheta_n := \mathbb{E}[U_{n,r_n,\omega}]$, it holds that

$$\mathbb{E}[h_n((Z_i)_{i=1}^{r_n}, \omega)] = \vartheta_n.$$

An important result for U-statistics is the Hoeffding-decomposition, first introduced by Hoeffding (1961). We present the result for generalized complete U-statistics with infinite order, but it reduces to the classical result without these extensions.

Theorem 2.10 (Hoeffding-decomposition)**.** *Suppose $U_{n,r_n,\omega}$ is a generalized complete U-statistic of order r_n with kernel h_n. For $j = 1, \dots, r_n$ denote*

$$h_{n,j}(z_1, \dots, z_j) := \mathbb{E}[h_n(z_1, \dots, z_j, Z_{j+1}, \dots, Z_{r_n}, \omega)] - \vartheta_n \tag{2.12}$$

then it holds that $\mathbb{E}[h_{n,j}(Z_1, \dots, Z_j)] = 0$. Further, for $j = 1, \dots, r_n - 1$ let

$$h_n^{(j)}(z_1, \dots, z_j) = h_{n,j}(z_1, \dots, z_j) - \sum_{l=1}^{j-1} \sum_{\{i_1,\dots,i_l\} \in B_{l,j}} h_n^{(l)}(z_{i_1}, \dots, z_{i_l}) \tag{2.13}$$

and

$$h_n^{(r_n)}(z_1, \dots, z_{r_n}, \omega) = h_n(z_1, \dots, z_{r_n}, \omega) - \sum_{l=1}^{r_n-1} \sum_{\{i_1,\dots,i_l\} \in B_{l,r_n}} h_n^{(l)}(z_{i_1}, \dots, z_{i_l}). \tag{2.14}$$

It holds that

$$H_n^{(j)} = \binom{n}{j}^{-1} \sum_{I \in B_{n,j}} h_n^{(j)}((Z_i)_{i \in I})$$

and

$$H_{n,\omega}^{(r_n)} = \binom{n}{r_n}^{-1} \sum_{I \in B_{n,r_n}} h_n^{(r_n)}((Z_i)_{i \in I}, \omega_I)$$

are U-statistics of order j and r_n with kernels $h_n^{(j)}$ and $h_n^{(r_n)}$, respectively, satisfying

$$\begin{aligned} U_{n,r_n,\omega} &= \vartheta_n + \sum_{j=1}^{r_n-1} \binom{r_n}{j} H_n^{(j)} + H_{n,\omega}^{(r_n)} \\ &= \vartheta_n + \sum_{j=1}^{r_n-1} \binom{r_n}{j} \binom{n}{j}^{-1} \sum_{I \in B_{n,j}} h_n^{(j)}((Z_i)_{i \in I}) \\ &\quad + \binom{n}{r_n}^{-1} \sum_{I \in B_{n,r_n}} h_n^{(r_n)}((Z_i)_{i \in I}, \omega_I). \end{aligned} \tag{2.15}$$

An important property of the Hoeffding-decomposition is that the kernels $h_n^{(j)}$ and thus the U-statistics $H_n^{(j)}$ are uncorrelated. More precisely it holds that

$$\mathrm{Cov}(h_n^{(j_1)}((Z_i)_{i \in I}), h_n^{(j_2)}((Z_l)_{l \in L})) = 0 \tag{2.16}$$

if $j_1 \neq j_2$ or if $j_1 = j_2$ and $I \neq L$. This property includes the kernel $h_n^{(r_n)}$, which would have an additional argument ω. The Hoeffding-decomposition can be especially useful for analyzing the asymptotic properties of a U-statistic. An important characteristic of the U-statistic are the variances of the different kernels from above. We define

$$\begin{aligned}
\zeta_{r_n}^n &:= \mathrm{Var}(h_n((Z_i)_{i=1}^{r_n}, \omega)), \\
\zeta_{j,\omega}^n &:= \mathrm{Cov}(h_n(Z_1, \ldots, Z_j, Z_{j+1}, \ldots, Z_{r_n}, \omega_1), h_n(Z_1, \ldots, Z_j, \tilde{Z}_{j+1}, \ldots, \tilde{Z}_{r_n}, \omega_2)) \\
&= \mathrm{Var}\left(\mathbb{E}\left[h_n(Z_1, \ldots, Z_{r_n}, \omega) | Z_1, \ldots, Z_j\right]\right) \\
&= \mathrm{Var}\left(h_{n,j}(Z_1, \ldots, Z_j)\right) \qquad (2.17)
\end{aligned}$$

for $j \in [r_n]$, where ω_1, ω_2 are i.i.d. and the $\tilde{Z}_i$ are i.i.d. copies of the Z_i. Further let

$$V_{n,\omega}^{(j)} := \mathrm{Var}\left(h_n^{(j)}(Z_1, \ldots, Z_j)\right) \qquad \text{and} \qquad (2.18)$$

$$V_n^{(r_n)} := \mathrm{Var}\left(h_n^{(r_n)}((Z_i)_{i=1}^{r_n}, \omega)\right) \qquad (2.19)$$

for $h_n^{(j)}$ and $h_n^{(r_n)}$ from (2.13) and (2.14). We note that we could analogously define $V_{r_n,\omega}$ if we would define a kernel similar to $h_n^{(r_n)}$ that has no argument ω. With (2.14) and (2.16) we get a useful relation between these variances. It holds that

$$\begin{aligned}
\zeta_{r_n}^n &= \mathrm{Var}(h_n((Z_i)_{i=1}^{r_n}, \omega)) \\
&= \mathrm{Var}\left(h_n^{(r_n)}((Z_i)_{i=1}^{r_n}, \omega)) + \sum_{j=1}^{r_n-1} \sum_{I \in B_{j,r_n}} h_n^{(j)}((Z_i)_{i \in I})\right) \\
&= V_n^{(r_n)} + \sum_{j=1}^{r_n-1} \binom{r_n}{j} V_{n,\omega}^{(j)}. \qquad (2.20)
\end{aligned}$$

Similarly (2.13) implies

$$\zeta_{j,\omega}^n = V_{n,\omega}^{(j)} + \sum_{l=1}^{j-1} \binom{j}{l} V_{n,\omega}^{(l)}.$$

In particular, we have $\zeta_{1,\omega}^n = V_{n,\omega}^{(1)}$ and with (2.20) we obtain

$$\zeta_{r_n}^n = V_n^{(r_n)} + \sum_{j=2}^{r_n-1} \binom{r_n}{j} V_{n,\omega}^{(j)} + r_n \zeta_{1,\omega}^n \geq r_n \zeta_{1,\omega}^n \qquad (2.21)$$

because the variances are non-negative. Similarly, but with a slightly more sophisticated argument for the non-negativity one can prove that

$$\frac{1}{l}\zeta_{l,\omega}^n \leq \frac{1}{j}\zeta_{j,\omega}^n$$

for $1 \leq l \leq j \leq r_n$. Jensen's inequality for the conditional expectation directly implies $\zeta_{r_n,\omega}^n \leq \zeta_{r_n}^n$. Applying the covariance property of the kernels from (2.16) to the Hoeffding-decomposition (2.15) and using the definitions (2.18) and (2.19), we further obtain the following formula for the variance of a U-statistic

$$\mathrm{Var}(U_{n,r_n,\omega}) = \sum_{j=1}^{r_n-1} \binom{r_n}{j}^2 \binom{n}{j}^{-1} V_{n,\omega}^{(j)} + \binom{n}{r_n}^{-1} V_n^{(r_n)}. \qquad (2.22)$$

It is also possible to express the variance through the $\zeta_{j,\omega}^n$. A random variable related to the U-statistic $U_{n,r_n,\omega}$ that will be useful for its asymptotic analysis is the so-called Hájek projection of $U_{n,r_n,\omega}$, defined as

$$\begin{aligned}
\hat{U}_{n,r_n,\omega} &:= \sum_{j=1}^{n} \mathbb{E}\left[U_{n,r_n,\omega}|Z_j\right] - (n-1)\vartheta_n \\
&= \sum_{j=1}^{n} \mathbb{E}\left[\binom{n}{r_n}^{-1} \sum_{\{i_1,\dots,i_{r_n}\}\in B_{r_n,n}} h_n(Z_{i_1},\dots,Z_{i_{r_n}},\omega_I)|Z_j\right] - (n-1)\vartheta_n \\
&= \sum_{j=1}^{n} \binom{n}{r_n}^{-1} \sum_{\{i_1,\dots,i_{r_n}\}\in B_{r_n,n}} \mathbb{E}[h_n(Z_{i_1},\dots,Z_{i_{r_n}},\omega_I)|Z_j] - (n-1)\vartheta_n \\
&= \sum_{j=1}^{n} \binom{n}{r_n}^{-1} \left(\binom{n-1}{r_n-1}(h_{n,1}(Z_j)+\vartheta_n) + \binom{n-1}{r_n}\vartheta_n\right) - (n-1)\vartheta_n \\
&= \sum_{j=1}^{n} \binom{n}{r_n}^{-1} \left(\binom{n-1}{r_n-1}h_{n,1}(Z_j) + \binom{n}{r_n}\vartheta_n\right) - (n-1)\vartheta_n \\
&= \sum_{j=1}^{n} \binom{n}{r_n}^{-1} \binom{n-1}{r_n-1}h_{n,1}(Z_j) + \vartheta_n \\
&= \frac{r_n}{n}\sum_{j=1}^{n} h_{n,1}(Z_j) + \vartheta_n. \qquad (2.23)
\end{aligned}$$

It is called Hájek projection because it is the projection onto a subspace of random variables of the form $\sum_{j=1}^n g_j(Z_j)$. The function $h_{n,1}$ is centered due to (2.12), and thus it holds that $\mathbb{E}[\hat{U}_{n,r_n,\omega}] = \vartheta_n$. Further (2.13) implies that $h_n^{(1)} = h_{n,1}$ and

$$\hat{U}_{n,r_n,\omega} = \vartheta_n + r_n H_n^{(1)},$$

where $H_n^{(1)}$ is the U-statistic of order one from the Hoeffding-decomposition. This means, the Hájek projection consists of the first order terms from the Hoeffding-decomposition.

The Lemma below is useful to bound moments of generalized U-statistics. The result and the proof are a generalization of an analogue result for standard U-statistics, see Lee (1990, Section 1.5, Theorem 1). The lemma will be used in the proof of Theorem 2.13 below and later in Chapter 5.

Lemma 2.11. *Let $U_{n,r_n,\omega}$ be a generalized U-statistic with kernel $h_n(Z_1,\dots,Z_{r_n},\omega)$ satisfying $\mathbb{E}[h_n(Z_1,\dots,Z_{r_n},\omega)] = 0$. For $q \in [2,\infty)$ there exists a constant C such that*

$$\mathbb{E}\left[|U_{n,r_n,\omega}|^q\right] \leq C\left(\frac{r_n}{n}\right)^{q/2} \mathbb{E}\left[|h_n(Z_1,\dots Z_{r_n},\omega)|^q\right].$$

The remaining two results in this section are from the article by Peng et al. (2022). They state asymptotic normality of the complete and incomplete generalized U-statistics. Notably, the results differ from those presented in the aforementioned article due to the

inclusion of the additional assumption (2.25). We explain the reason for this disparity directly after the first theorem. The results and more detailed versions of their proofs are included here for completeness and to discuss and compare different proof techniques.

Theorem 2.12. *Let $Z_1, \dots, Z_n$ be i.i.d. from F_Z and $U_{n,r_n,\omega}$ be a generalized complete U-statistic with kernel $h_n = h_n(Z_1, \dots, Z_{r_n}, \omega)$ such that $\zeta_{r_n}^n < \infty$. Suppose that*

$$\frac{r_n}{n} \frac{\zeta_{r_n}^n}{r_n \zeta_{1,\omega}^n} \to 0 \qquad \textit{and} \tag{2.24}$$

$$\frac{\mathbb{E}\left[|h_{n,1}(Z_1)|^{2+\delta}\right]^2}{n^\delta (\zeta_{1,\omega}^n)^{2+\delta}} \to 0, \tag{2.25}$$

for some $\delta > 0$. For $\vartheta_n = \mathbb{E}[h_n(Z_1, \dots, Z_{r_n}, \omega)]$, it holds that

$$\frac{U_{n,r_n,\omega} - \vartheta_n}{\sqrt{r_n^2 \zeta_{1,\omega}^n / n}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

The assumption (2.25) is a consequence of the Lyapunov condition that we use to prove the convergence in distribution of the Hájek projection. Peng et al. (2022) claim in their proof of Theorem 2.12 that the convergence holds directly with the Lindeberg condition. However, the kernels $h_{n,1}$ from the Hájek projection are dependent on n, meaning that the $h_{n,1}(Z_j)$ form a triangular array. Consequently, the Lindeberg condition is not satisfied without additional assumptions on $h_{n,1}$. The assumption in (2.25) is sufficient because it implies the Lyapunov condition. The Lindeberg condition holds if

$$\mathbb{E}\left[h_{n,1}(Z_1)^2 / \zeta_{1,\omega}^n \mathbb{I}\{h_{n,1}(Z_1)^2 / \zeta_{1,\omega}^n > \varepsilon^2 n\}\right] \to 0.$$

Therefore, an alternative assumption would be that $h_{n,1}(Z_1)^2/\zeta_{1,\omega}^n$ is dominated by some integrable function $g(z_1)$ and $h_{n,1}(z_1)^2/(\zeta_{1,\omega}^n n) \to 0$ for every z_1. In this case, the dominated convergence theorem yields the Lindeberg condition. In our application in Chapter 4 both options would lead to the same assumption.

The main assumption of this result is

$$\frac{r_n}{n} \frac{\zeta_{r_n}^n}{r_n \zeta_{1,\omega}^n} = \frac{\zeta_{r_n}^n}{n \zeta_{1,\omega}^n} \to 0.$$

It is formulated this way because $\zeta_{1,\omega}^n \le \frac{1}{r_n} \zeta_{r_n,\omega}^n \le \frac{1}{r_n} \zeta_{r_n}^n$. The assumption depends on the concrete kernel of the U-statistic. The final theorem in this section covers incomplete U-statistics and will be applied in Chapter 4.

Theorem 2.13. *Let $Z_1, \dots, Z_n$ be i.i.d. from F_Z and $U_{n,r_n,N,\omega}$ be a generalized incomplete U-statistic with kernel $h_n = h_n(Z_1, \dots, Z_{r_n}, \omega)$ such that $\mathbb{E}[|h_n - \vartheta_n|^{2k}] / \mathbb{E}[|h_n - \vartheta_n|^k]^2 \le C$ for $k = 2, 3$ for some constant C and all r_n. Let $\vartheta_n = \mathbb{E}[h_n(Z_1, \dots, Z_{r_n}, \omega)]$ and assume that (2.24) and (2.25) from Theorem 2.12 hold for some $\delta > 0$. If $N \to \infty$, then*

$$\frac{U_{n,r_n,N,\omega} - \vartheta_n}{\sqrt{r_n^2 \zeta_{1,\omega}^n / n + \zeta_{r_n}^n / N}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

2.3.1 Proofs

We start with the proof of Lemma 2.11, followed by the proofs of Theorem 2.12 and Theorem 2.13.

2.3.1.1 Proof of Lemma 2.11

We denote $[n] := \{1, \dots, n\}$ and

$$\Pi(n) := \{\pi : [n] \to [n] : \pi([n]) = [n]\}$$

which is the set of all permutations of $\{1, \dots, n\}$. Let $n_b = \lfloor n/r_n \rfloor$, we further denote

$$g(z_1, \dots, z_n) = \frac{1}{n_b} \sum_{l=0}^{n_b-1} h_n(z_{lr_n+1}, \dots, z_{(l+1)r_n}, \omega_{\{lr_n+1,\dots,(l+1)r_n\}})$$

By counting the appropriate permutations it can be seen that

$$\begin{aligned}
&n_b \sum_{\pi\in\Pi(n)} g(Z_{\pi(1)}, \dots, Z_{\pi(n)}) \\
&= \sum_{\pi\in\Pi(n)} \sum_{l=0}^{n_b-1} h_n(Z_{\pi(lr_n+1)}, \dots, Z_{\pi((l+1)r_n)}, \omega_{\{\pi(lr_n+1),\dots,\pi((l+1)r_n)\}}) \\
&= n_b r_n!(n-r_n)! \sum_{I\in B_{r_n,n}} h_n((Z_i)_{i\in I}, \omega_I).
\end{aligned}$$

Hence we can write a generalized U-statistic as

$$U_{n,r_n,\omega} = \frac{1}{n!} \sum_{\pi\in\Pi(n)} g(Z_{\pi(1)}, \dots, Z_{\pi(n)}).$$

For any convex function f we can now use

$$f(U_{n,r_n,\omega}) = f\left(\frac{1}{n!} \sum_{\pi\in\Pi(n)} g(Z_{\pi(1)}, \dots, Z_{\pi(n)})\right) \le \frac{1}{n!} \sum_{\pi\in\Pi(n)} f\left(g(Z_{\pi(1)}, \dots, Z_{\pi(n)})\right)$$

and thus

$$\mathbb{E}\big[f(U_{n,r_n,\omega})\big] \le \mathbb{E}\big[f(g(Z_1, \dots, Z_n))\big].$$

For this inequality we do not need the centered kernel yet. We choose $f(\xi) = |\xi|^q$ and obtain

$$\begin{aligned}
\mathbb{E}\left[|U_{n,r_n,\omega}|^q\right] &\le \mathbb{E}\left[|g(Z_1, \dots, Z_n)|^q\right] \\
&= \mathbb{E}\left[\left|\frac{1}{n_b} \sum_{l=0}^{n_b-1} h_n(Z_{lr_n+1}, \dots, Z_{(l+1)r_n}, \omega_{\{lr_n+1,\dots,(l+1)r_n\}})\right|^q\right]
\end{aligned}$$

$$= \frac{1}{n_b^q}\mathbb{E}\left[\left|\sum_{l=0}^{n_b-1} h_n(Z_{lr_n+1},\dots,Z_{(l+1)r_n},\omega_{\{lr_n+1,\dots,(l+1)r_n\}})\right|^q\right].$$

The terms

$$\left(h_n(Z_{lr_n+1},\dots Z_{(l+1)r_n},\omega_{\{lr_n+1,\dots,(l+1)r_n\}})\right)_{l=0}^{n_b-1}$$

are independent because h_n is applied to blocks of independent Z_j and the ω_I are independent as well. We can apply Proposition 2.7 and Remark 2.8 to obtain

$$\begin{aligned}
\mathbb{E}\left[|U_{n,r_n,\omega}|^q\right] &\leq \frac{1}{n_b^q}\mathbb{E}\left[\left|\sum_{l=0}^{n_b-1} h_n(Z_{lr_n+1},\dots,Z_{(l+1)r_n},\omega_{\{lr_n+1,\dots,(l+1)r_n\}})\right|^q\right] \\
&\leq C_{q,2}\frac{1}{n_b^q}\mathbb{E}\left[\left(\sum_{l=0}^{n_b-1} h_n(Z_{lr_n+1},\dots Z_{(l+1)r_n},\omega_{\{lr_n+1,\dots,(l+1)r_n\}})^2\right)^{q/2}\right] \\
&\leq C_{q,2}\frac{1}{n_b^{q/2}}\mathbb{E}\left[|h_n(Z_1,\dots Z_{r_n},\omega)|^q\right] \\
&= C_{q,2}\frac{1}{(\lfloor n/r_n\rfloor)^{q/2}}\mathbb{E}\left[|h_n(Z_1,\dots Z_{r_n},\omega)|^q\right] \\
&\leq C\left(\frac{r_n}{n}\right)^{q/2}\mathbb{E}\left[|h_n(Z_1,\dots Z_{r_n},\omega)|^q\right]
\end{aligned}$$

for a suitable constant C. □

2.3.1.2 Proof of Theorem 2.12

This proof is by Peng et al. (2022). It is included in a more detailed version for completeness and for comparison of proof methods. First we consider the Hájek projection $\hat{U}_{n,r_n,\omega}$. We have

$$\hat{U}_{n,r_n,\omega} = \frac{r_n}{n}\sum_{j=1}^{n} h_{n,1}(Z_j) + \vartheta_n.$$

We want to show asymptotic normality of this term. We keep in mind that $h_{n,1}(Z_i)$ depends on r_n since the order of the kernel is not assumed to be fixed. We consider $\hat{U}_{n,r_n,\omega} - \vartheta_n$ to get a term with expectation zero. We get

$$\begin{aligned}
\mathrm{Var}(\hat{U}_{n,r_n,\omega}) &= r_n^2\mathrm{Var}\left(\frac{1}{n}\sum_{j=1}^{n} h_{n,1}(Z_j)\right) \\
&= \frac{r_n^2}{n}\mathrm{Var}\big(h_{n,1}(Z_1)\big) \\
&= \frac{r_n^2}{n}\zeta_{1,\omega}^n.
\end{aligned}$$

Since $h_{n,1}(Z_i)$ depends on r_n we effectively have a triangular array of random variables. The random variables in the array are $r_n h_{n,1}(Z_j)$. Assumption (2.25) implies that the Lyapunov condition holds for some $\delta > 0$ because

$$\frac{1}{(nr_n^2\zeta_{1,\omega}^n)^{(2+\delta)/2}}\sum_{j=1}^{n}\mathbb{E}\left[|r_n h_{n,1}(Z_j)|^{2+\delta}\right] = \frac{\mathbb{E}\left[|h_{n,1}(Z_1)|^{2+\delta}\right]}{n^{\delta/2}(\zeta_{1,\omega}^n)^{(2+\delta)/2}} \overset{n\to\infty}{\longrightarrow} 0.$$

This directly implies

$$\frac{\hat{U}_{n,r_n,\omega} - \vartheta_n}{\sqrt{r_n^2 \zeta_{1,\omega}^n / n}} = \frac{r_n}{n} \sum_{i=1}^n h_{n,1}(Z_i) \frac{1}{\sqrt{r_n^2 \zeta_{1,\omega}^n / n}} = \frac{\sum_{i=1}^n r_n h_{n,1}(Z_i)}{\sqrt{r_n^2 \zeta_{1,\omega}^n n}} \xrightarrow{d} \mathcal{N}(0,1).$$

We want to show that

$$\frac{\mathrm{Var}(U_{n,r_n,\omega})}{\mathrm{Var}(\hat{U}_{n,r_n,\omega})} \to 1.$$

To get this we need the following two observations. For $n \geq 3$ and $r_n \geq 2$ we obtain

$$\begin{aligned}\binom{r_n}{j}\binom{n}{j}^{-1} = \frac{r_n!}{j!(r_n-j)!}\frac{j!(n-j)!}{n!} &= \frac{r_n!}{(r_n-j)!}\frac{(n-j)!}{n!} \\ &= \prod_{i=0}^{j-1} \frac{r_n - i}{n-i} \\ &= \frac{r_n}{n}\frac{r_n-1}{n-1}\prod_{i=2}^{j-1}\frac{r_n-i}{n-i} \leq \frac{r_n^2}{n^2}\end{aligned}$$

as well as

$$\begin{aligned}\binom{n}{r_n}^{-1} = \frac{r_n!(n-r_n)!}{n!} &= \prod_{i=0}^{r_n-1}\frac{r_n-i}{n-i} \\ &= \frac{r_n}{n}\frac{r_n-1}{n-1}\prod_{i=2}^{r_n-1}\frac{r_n-i}{n-i} \leq \frac{r_n^2}{n^2}. \qquad (2.26)\end{aligned}$$

We use (2.22), together with these observations, $V_{n,\omega}^{(1)} = \zeta_{1,\omega}^n$ and (2.20) to get

$$\begin{aligned}\frac{\mathrm{Var}(U_{n,r_n,\omega})}{\mathrm{Var}(\hat{U}_{n,r_n,\omega})} &= \left(\frac{r_n^2}{n}\zeta_{1,\omega}^n\right)^{-1}\left(\sum_{j=1}^{r_n-1}\binom{r_n}{j}^2\binom{n}{j}^{-1}V_{n,\omega}^{(j)} + \binom{n}{r_n}^{-1}V_n^{(r_n)}\right) \\ &= \left(\frac{r_n^2}{n}V_{n,\omega}^{(1)}\right)^{-1}\left(r_n^2 n^{-1}V_{n,\omega}^{(1)} + \sum_{j=2}^{r_n-1}\binom{r_n}{j}^2\binom{n}{j}^{-1}V_{n,\omega}^{(j)} + \binom{n}{r_n}^{-1}V_n^{(r_n)}\right) \\ &= 1 + \frac{n}{r_n^2\zeta_{1,\omega}^n}\left(\sum_{j=2}^{r_n-1}\binom{r_n}{j}^2\binom{n}{j}^{-1}V_{n,\omega}^{(j)} + \binom{n}{r_n}^{-1}V_n^{(r_n)}\right) \\ &\leq 1 + \frac{n}{r_n^2\zeta_{1,\omega}^n}\frac{r_n^2}{n^2}\left(\sum_{j=2}^{r_n-1}\binom{r_n}{j}V_{n,\omega}^{(j)} + V_n^{(r_n)}\right) \\ &\leq 1 + \frac{1}{n\zeta_{1,\omega}^n}\zeta_{r_n}^n = 1 + \frac{r_n}{n}\frac{\zeta_{r_n}^n}{r_n\zeta_{1,\omega}^n} \to 1. \qquad (2.27)\end{aligned}$$

The above calculation further implies $\mathrm{Var}(U_{n,r_n,\omega})/\mathrm{Var}(\hat{U}_{n,r_n,\omega}) \geq 1$ because the remaining variance terms in the third line are non-negative. Thus, we can apply Theorem 11.2 by van der Vaart (1998) which is given as Theorem 2.5 to get the claim. □

2.3.1.3 Proof of Theorem 2.13

Without loss of generality, let $\vartheta_n = 0$. Within this proof let us denote $p = N/\binom{n}{r_n}$. We use the following decomposition

$$\begin{aligned} U_{n,r_n,N,\omega} &= \frac{1}{\hat{N}} \sum_{I \in B_{r_n,n}} \rho_I h_n((Z_i)_{i\in I}, \omega_I) \\ &= \frac{1}{\hat{N}} \frac{N}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} h_n((Z_i)_{i\in I}, \omega_I) \\ &\quad + \frac{1}{\hat{N}} \sum_{I \in B_{r_n,n}} (\rho_I - p) h_n((Z_i)_{i\in I}, \omega_I) \\ &= \frac{N}{\hat{N}} \left[U_{n,r_n,\omega} + \frac{1}{N} \sum_{I \in B_{r_n,n}} (\rho_I - p) h_n((Z_i)_{i\in I}, \omega_I) \right] \\ &=: \frac{N}{\hat{N}} (A_n + B_n). \end{aligned} \tag{2.28}$$

We note that $N/\hat{N} \xrightarrow{\mathbb{P}} 1$. Using $\vartheta_n = 0$ we get for the covariance of A_n and B_n that

$$\begin{aligned} \mathrm{Cov}(A_n, B_n) &= \mathbb{E}\left[\frac{1}{N} \sum_{I \in B_{r_n,n}} (\rho_I - p) h_n((Z_i)_{i\in I}, \omega_I) \frac{1}{\binom{n}{r_n}} \sum_{J \in B_{r_n,n}} h_n((Z_i)_{i\in J}, \omega_J) \right] \\ &= \frac{1}{N} \sum_{I \in B_{r_n,n}} \mathbb{E}\big[\rho_I - p\big] \mathbb{E}\left[h_n((Z_i)_{i\in I}, \omega_I) \frac{1}{\binom{n}{r_n}} \sum_{J \in B_{r_n,n}} h_n((Z_i)_{i\in J}, \omega_J) \right] \\ &= 0. \end{aligned}$$

Using the independence of the ρ_I we obtain

$$\begin{aligned} \mathrm{Var}(B_n) &= \frac{1}{N^2} \sum_{I \in B_{r_n,n}} \mathbb{E}\left[(\rho_I - p)^2\right] \mathbb{E}\left[h_n^2((Z_i)_{i\in I}, \omega_I)\right] \\ &= \frac{1}{N^2} \sum_{I \in B_{r_n,n}} p(1-p)\zeta_{r_n}^n \\ &= \frac{1}{N^2} \binom{n}{r_n} p(1-p)\zeta_{r_n}^n \\ &= \frac{1}{N}(1-p)\zeta_{r_n}^n. \end{aligned} \tag{2.29}$$

First let us consider the case $p = N/\binom{n}{r_n} \not\to 0$. Theorem 2.12 yields

$$\frac{A_n}{\sqrt{r_n^2 \zeta_{1,\omega}^n / n}} \xrightarrow{d} \mathcal{N}(0,1), \tag{2.30}$$

since (2.24) and (2.25) are also assumed in Theorem 2.13. The fact that $p \not\to 0$ and (2.26) from the proof of Theorem 2.12 imply

$$\frac{n^2}{r_n^2 N} \leq \frac{\binom{n}{r_n}}{N} = p^{-1} = \mathcal{O}(1). \tag{2.31}$$

Using this in conjunction with (2.29) we obtain

$$\mathbb{E}\left[\frac{B_n^2}{r_n^2\zeta_{1,\omega}^n/n}\right] = \frac{\mathrm{Var}(B_n)}{r_n^2\zeta_{1,\omega}^n/n} = \frac{(1-p)\zeta_{r_n}^n n}{N r_n^2 \zeta_{1,\omega}^n} = \frac{(1-p)n^2}{r_n^2 N}\frac{r_n}{n}\frac{\zeta_{r_n}^n}{r_n\zeta_{1,\omega}^n} \to 0$$

due to assumption (2.24). The convergence in L_2 implies convergence in distribution and thus

$$\frac{A_n + B_n}{\sqrt{r_n^2\zeta_{1,\omega}^n/n}} \xrightarrow{d} \mathcal{N}(0,1)$$

with Slutsky's Theorem and (2.30). Again, using (2.31) together with (2.24), we get

$$\frac{\sqrt{r_n^2\zeta_{1,\omega}^n/n}}{\sqrt{r_n^2\zeta_{1,\omega}^n/n + \zeta_{r_n}^n/N}} = \left(1 + \frac{\zeta_{r_n}^n n}{r_n^2\zeta_{1,\omega}^n N}\right)^{-1/2} = \left(1 + \frac{n^2}{r_n^2 N}\frac{r_n}{n}\frac{\zeta_{r_n}^n}{r_n\zeta_{1,\omega}^n}\right)^{-1/2} \to 1,$$

and hence (2.28) and another application of Slutsky's Theorem yield

$$\frac{U_{n,r_n,N,\omega}}{\sqrt{r_n^2\zeta_{1,\omega}^n/n + \zeta_{r_n}^n/N}} = \frac{N}{\hat{N}}\frac{A_n + B_n}{\sqrt{r_n^2\zeta_{1,\omega}^n/n + \zeta_{r_n}^n/N}} \xrightarrow{d} \mathcal{N}(0,1).$$

From here on we consider the case $p \to 0$. Let us denote the σ-algebra

$$\mathcal{A}_{Z,\omega} := \sigma(\{Z_1, \ldots, Z_n, (\omega_I)_{I \in B_{r_n,n}}\}).$$

For any fixed t, we define

$$\begin{aligned}
\phi_{A_n+B_n}(t) &:= \mathbb{E}\left[\exp\left(it\left(\frac{r_n^2\zeta_{1,\omega}^n}{n} + \frac{\zeta_{r_n}^n}{N}\right)^{-1/2}(A_n + B_n)\right)\right] \\
\hat{\phi}_{A_n}(t) &:= \exp\left(it\left(\frac{r_n^2\zeta_{1,\omega}^n}{n} + \frac{\zeta_{r_n}^n}{N}\right)^{-1/2}A_n\right) \qquad (2.32) \\
\hat{\phi}_{B_n}(t) &:= \mathbb{E}\left[\exp\left(it\left(\frac{r_n^2\zeta_{1,\omega}^n}{n} + \frac{\zeta_{r_n}^n}{N}\right)^{-1/2}B_n\right) \,\middle|\, \mathcal{A}_{Z,\omega}\right] \qquad (2.33)
\end{aligned}$$

such that

$$\phi_{A_n+B_n}(t) = \mathbb{E}\left[\hat{\phi}_{A_n}(t)\hat{\phi}_{B_n}(t)\right] \tag{2.34}$$

because A_n is $\mathcal{A}_{Z,\omega}$ measurable. We denote

$$U_{n,2} = \binom{n}{r_n}^{-1} \sum_{I \in B_{r_n,n}} h_n^2((Z_i)_{i\in I}, \omega_I),$$

which is a complete U-statistic with kernel h_n^2. Let $\varepsilon > 0$ be arbitrary. When denoting the moments of h_n, we frequently omit its arguments for convenience. Using Chebyshev's inequality and Lemma 2.11 (with constant denoted by C_1), we get

$$\begin{aligned}\mathbb{P}\left(|U_{n,2} - \mathbb{E}[h_n^2]| \geq \varepsilon \mathbb{E}[h_n^2]\right) &\leq \frac{\operatorname{Var}(U_{n,2})}{\varepsilon^2 \mathbb{E}[h_n^2]^2} \\ &\leq C_1 \frac{r_n}{n} \frac{\mathbb{E}\left[\left(h_n^2 - \mathbb{E}[h_n^2]\right)^2\right]}{\varepsilon^2 \mathbb{E}[h_n^2]^2} \\ &\leq C_1 \frac{r_n}{n} \frac{\mathbb{E}[h_n^4]}{\varepsilon^2 \mathbb{E}[h_n^2]^2} \leq \frac{r_n}{n} \frac{C_1 C}{\varepsilon^2}\end{aligned}$$

for C from the claim. Assumption (2.24) and the fact that $\zeta_{1,\omega}^n \leq \frac{1}{r_n}\zeta_{r_n}^n$ from (2.21) imply that $r_n/n \to 0$ and thus $U_{n,2}/\zeta_{r_n}^n \xrightarrow{\mathbb{P}} 1$. Due to the assumptions in the theorem the analogue result $U_{n,3}/\mathbb{E}[|h_n|^3] \xrightarrow{\mathbb{P}} 1$ holds for

$$U_{n,3} := \binom{n}{r_n}^{-1} \sum_{I \in B_{r_n,n}} |h_n((Z_i)_{i\in I}, \omega_I)|^3.$$

We define the set

$$D := \{U_{n,2}/\mathbb{E}[h_n^2] \in [1-\delta, 1+\delta]\} \cap \{U_{n,3}/\mathbb{E}[|h_n|^3] \in [1-\delta, 1+\delta]\}. \tag{2.35}$$

For any $\delta, \varepsilon > 0$ this event holds with probability at least $1 - \varepsilon$, if n is sufficiently large. We define $\hat{d}_{n,r_n,N} := ((1-p)U_{n,2}/N)^{1/2}$ and consider

$$B_n/\hat{d}_{n,r_n,N} = \frac{1}{\sqrt{N(1-p)U_{n,2}}} \sum_{I \in B_{r_n,n}} (\rho_I - p) h_n((Z_i)_{i\in I}, \omega_I)$$

conditioned on $\mathcal{A}_{Z,\omega}$. The remaining random variables are the ρ_I, thus the term is a sum of the random variables

$$X_{n,I} := \frac{1}{\sqrt{N(1-p)U_{n,2}}} (\rho_I - p) h_n((Z_i)_{i\in I}, \omega_I), \tag{2.36}$$

that are independent conditioned on $\mathcal{A}_{Z,\omega}$. It holds that $\mathbb{E}[X_{n,I} \mid \mathcal{A}_{Z,\omega}] = 0$ and we further denote

$$\begin{aligned}\sigma_{n,I}^2 := \mathbb{E}\left[X_{n,I}^2 \mid \mathcal{A}_{Z,\omega}\right] &= \frac{1}{N(1-p)U_{n,2}} h_n^2((Z_i)_{i\in I}, \omega_I) p(1-p) \\ &= \frac{1}{U_{n,2}\binom{n}{r_n}} h_n^2((Z_i)_{i\in I}, \omega_I),\end{aligned} \tag{2.37}$$

such that

$$\operatorname{Var}\left(B_n/\hat{d}_{n,r_n,N} \mid \mathcal{A}_{Z,\omega}\right) = \sum_{I \in B_{r_n,n}} \sigma_{n,I}^2 = 1.$$

We bound the difference of the characteristic function of $B_n/\hat{d}_{n,r_n,N}$ conditional $\mathcal{A}_{Z,\omega}$ to the characteristic function of a standard normal distribution. Using that

$$\Big|\prod_{j=1}^{n}\xi_j-\prod_{j=1}^{n}\tilde{\xi}_j\Big|\leq\sum_{j=1}^{n}|\xi_j-\tilde{\xi}_j|$$

for all $\xi_1,\tilde{\xi}_1,\ldots,\xi_n,\tilde{\xi}_n\in\mathbb{C}$ with $|\xi_j|\leq 1$ and $|\tilde{\xi}_j|\leq 1$, we obtain

$$\begin{aligned}
&\Big|\mathbb{E}\left[\exp\left(iuB_n/\hat{d}_{n,r_n,N}\right)\mid\mathcal{A}_{Z,\omega}\right]-e^{-\frac{u^2}{2}}\Big|\\
&\quad=\Big|\prod_{I\in B_{r_n,n}}\mathbb{E}\left[\exp\left(iuX_{n,I}\right)\mid\mathcal{A}_{Z,\omega}\right]-\prod_{I\in B_{r_n,n}}e^{-\frac{u^2\sigma_{n,I}^2}{2}}\Big|\\
&\quad\leq\sum_{I\in B_{r_n,n}}\Big|\mathbb{E}\left[\exp\left(iuX_{n,I}\right)\mid\mathcal{A}_{Z,\omega}\right]-e^{-\frac{u^2\sigma_{n,I}^2}{2}}\Big|\\
&\quad\leq\sum_{I\in B_{r_n,n}}\Big|\mathbb{E}\left[\exp\left(iuX_{n,I}\right)\mid\mathcal{A}_{Z,\omega}\right]-\left(1-\frac{u^2}{2}\sigma_{n,I}^2\right)\Big|\\
&\qquad+\sum_{I\in B_{r_n,n}}\Big|e^{-\frac{u^2\sigma_{n,I}^2}{2}}-\left(1-\frac{u^2}{2}\sigma_{n,I}^2\right)\Big|.
\end{aligned}\tag{2.38}$$

The Taylor expansion of e^{it} yields

$$\Big|e^{it}-\sum_{j=0}^{k}\frac{(it)^j}{j!}\Big|\leq\min\left\{\frac{2|t|^k}{k!},\frac{|t|^{k+1}}{(k+1)!}\right\}.\tag{2.39}$$

Applying this to the first term from (2.38) for $k=2$ and using the definition of $X_{n,I}$ in (2.36) we get

$$\begin{aligned}
&\sum_{I\in B_{r_n,n}}\Big|\mathbb{E}\left[\exp\left(iuX_{n,I}\right)\mid\mathcal{A}_{Z,\omega}\right]-\left(1-\frac{u^2}{2}\sigma_{n,I}^2\right)\Big|\\
&\qquad=\sum_{I\in B_{r_n,n}}\left|\mathbb{E}\left[\exp\left(iuX_{n,I}\right)-\left(1+iuX_{n,I}-\frac{u^2}{2}X_{n,I}^2\right)\mid\mathcal{A}_{Z,\omega}\right]\right|\\
&\qquad\leq\sum_{I\in B_{r_n,n}}\mathbb{E}\left[\min\left\{|uX_{n,I}|^2,\frac{1}{6}|uX_{n,I}|^3\right\}\mid\mathcal{A}_{Z,\omega}\right]\\
&\qquad\leq\frac{|u|^3}{6}\sum_{I\in B_{r_n,n}}\mathbb{E}\left[|X_{n,I}|^3\mid\mathcal{A}_{Z,\omega}\right]\\
&\qquad=\frac{|u|^3}{6}\frac{1}{(N(1-p)U_{n,2})^{3/2}}\sum_{I\in B_{r_n,n}}\mathbb{E}\left[|\rho_I-p|^3\right]h_n^3((Z_i)_{i\in I},\omega_I)\\
&\qquad=\frac{|u|^3}{6}\frac{(p(1-p)^3+(1-p)p^3)}{(N(1-p)U_{n,2})^{3/2}}\sum_{I\in B_{r_n,n}}h_n^3((Z_i)_{i\in I},\omega_I)
\end{aligned}$$

$$= \frac{|u|^3}{6} \frac{(1-p)^2+p^2}{\sqrt{N(1-p)}} \frac{p}{N} \binom{n}{r_n} \frac{U_{n,3}}{U_{n,2}^{3/2}}$$
$$= \frac{|u|^3}{6} \frac{(1-p)^2+p^2}{\sqrt{N(1-p)}} \frac{U_{n,3}}{U_{n,2}^{3/2}}. \tag{2.40}$$

We proceed with the second term from (2.38). Let $X_* \sim \mathcal{N}(0,1)$ be independent of all the other random variables. With (2.39) and the definition of $\sigma_{n,I}^2$ in (2.37) we obtain

$$\begin{aligned}
&\sum_{I \in B_{r_n,n}} \left| e^{-\frac{u^2\sigma_{n,I}^2}{2}} - \left(1 - \frac{u^2}{2}\sigma_{n,I}^2\right) \right| \\
&\quad = \sum_{I \in B_{r_n,n}} \left| \mathbb{E}\left[\exp(iu\sigma_{n,I}X_*) - \left(1 + iuX_* - \frac{1}{2}u^2\sigma_{n,I}^2 X_*^2\right) \mid \mathcal{A}_{Z,\omega} \right] \right| \\
&\quad \leq \sum_{I \in B_{r_n,n}} \mathbb{E}\left[\min\left\{ |u\sigma_{n,I}X_*|^2, \frac{1}{6}|u\sigma_{n,I}X_*|^3 \right\} \mid \mathcal{A}_{Z,\omega} \right] \\
&\quad \leq \frac{|u|^3}{6} \mathbb{E}\left[|X_*|^3\right] \sum_{I \in B_{r_n,n}} |\sigma_{n,I}|^3 \\
&\quad \leq \frac{|u|^3}{6} \mathbb{E}\left[|X_*|^4\right]^{3/4} \sum_{I \in B_{r_n,n}} \frac{1}{U_{n,2}^{3/2}\binom{n}{r_n}^{3/2}} h_n^3((Z_i)_{i\in I}, \omega_I) \\
&\quad = 3^{3/4} \frac{|u|^3}{6} \frac{U_{n,3}}{U_{n,2}^{3/2}\binom{n}{r_n}^{1/2}}.
\end{aligned}$$

Together with (2.38) and (2.40) we obtain on the event D (see (2.35)) that

$$\begin{aligned}
&\left| \mathbb{E}\left[\exp\left(iuB_n/\hat{d}_{n,r_n,N} \right) \mid \mathcal{A}_{Z,\omega} \right] - e^{-\frac{u^2}{2}} \right| \mathbb{I}_D \\
&\quad \leq \frac{|u|^3}{6} \frac{U_{n,3}}{U_{n,2}^{3/2}} \left(\frac{(1-p)^2+p^2}{\sqrt{N(1-p)}} + 3^{3/4}\binom{n}{r_n}^{-1/2} \right) \mathbb{I}_D \\
&\quad \leq \frac{|u|^3}{6} \frac{1+\delta}{(1-\delta)^{3/2}} \frac{\mathbb{E}[|h_n|^3]}{\mathbb{E}^{3/2}[|h_n|^2]} \left(\frac{(1-p)^2+p^2}{\sqrt{N(1-p)}} + 3^{3/4}\binom{n}{r_n}^{-1/2} \right) \mathbb{I}_D \\
&\quad \leq \frac{|u|^3}{6} \frac{1+\delta}{(1-\delta)^{3/2}} C^{3/4} \left(\frac{(1-p)^2+p^2}{\sqrt{N(1-p)}} + 3^{3/4}\binom{n}{r_n}^{-1/2} \right) \mathbb{I}_D.
\end{aligned} \tag{2.41}$$

For

$$u = t \left(\frac{(1-p)\zeta_{r_n}^n/N}{r_n^2\zeta_{1,\omega}^n/n + \zeta_{r_n}^n/N} \right)^{1/2} \left((\zeta_{r_n}^n)^{-1} U_{n,2} \right)^{1/2}$$

we use the definition of $\hat{\phi}_{B_n}(t)$ from (2.33) to obtain

$$\hat{\phi}_{B_n}(t) = \mathbb{E}\left[\exp\left(it \left(\frac{r_n^2\zeta_{1,\omega}^n}{n} + \frac{\zeta_{r_n}^n}{N} \right)^{-1/2} \left((1-p)U_{n,2}/N\right)^{1/2} B_n/\hat{d}_{n,r_n,N} \right) \mid \mathcal{A}_{Z,\omega} \right]$$

$$= \mathbb{E}\left[\exp\left(it\left(\frac{(1-p)\zeta_{r_n}^n/N}{r_n^2\zeta_{1,\omega}^n/n+\zeta_{r_n}^n/N}\right)^{1/2}\left((\zeta_{r_n}^n)^{-1}U_{n,2}\right)^{1/2}B_n/\hat{d}_{n,r_n,N}\right)\;\middle|\;\mathcal{A}_{Z,\omega}\right]$$
$$= \mathbb{E}\left[\exp\left(iuB_n/\hat{d}_{n,r_n,N}\right)\;\middle|\;\mathcal{A}_{Z,\omega}\right].$$

Noting that $|u|\mathbb{I}_D \leq |t|\sqrt{(1+\delta)}$ we get with (2.41) that

$$\left|\hat{\phi}_{B_n}(t) - \exp\left(-\frac{t^2}{2}\left(\frac{(1-p)\zeta_{r_n}^n/N}{r_n^2\zeta_{1,\omega}^n/n+\zeta_{r_n}^n/N}\right)(\zeta_{r_n}^n)^{-1}U_{n,2}\right)\right|\mathbb{I}_D$$
$$\leq \frac{|t|^3}{6}\frac{(1+\delta)^{5/2}}{(1-\delta)^{3/2}}C^{3/4}\left(\frac{(1-p)^2+p^2}{\sqrt{N(1-p)}}+3^{3/4}\binom{n}{r_n}^{-1/2}\right)\mathbb{I}_D \leq \varepsilon. \tag{2.42}$$

if n is large enough, because $N\to\infty$ and $p\to 0$. Let us denote

$$\phi_B(t) := \exp\left(-\frac{t^2}{2}\left(\frac{(1-p)\zeta_{r_n}^n/N}{r_n^2\zeta_{1,\omega}^n/n+\zeta_{r_n}^n/N}\right)\right). \tag{2.43}$$

The definition of the event D in (2.35) yields

$$\left|\frac{t^2}{2}\left(\frac{(1-p)\zeta_{r_n}^n/N}{r_n^2\zeta_{1,\omega}^n/n+\zeta_{r_n}^n/N}\right)(\zeta_{r_n}^n)^{-1}U_{n,2} - \frac{t^2}{2}\left(\frac{(1-p)\zeta_{r_n}^n/N}{r_n^2\zeta_{1,\omega}^n/n+\zeta_{r_n}^n/N}\right)\right|\mathbb{I}_D \leq \delta\frac{t^2}{2}.$$

Since the exponential function is uniformly continuous on any finite interval we conclude that there exists a $\delta' = \mathcal{O}(\delta)$, such that

$$\left|\exp\left(-\frac{t^2}{2}\left((1-p)\cdot\frac{\zeta_{r_n}^n/N}{r_n^2\zeta_{1,\omega}^n/n+\zeta_{r_n}^n/N}\right)(\zeta_{r_n}^n)^{-1}U_{n,2}\right) - \phi_B(t)\right|\mathbb{I}_D \leq \delta'.$$

Combined with (2.42) we get

$$|\hat{\phi}_{B_n}(t) - \phi_B(t)|\mathbb{I}_D \leq (\varepsilon+\delta')\mathbb{I}_D. \tag{2.44}$$

We continue with $\hat{\phi}_{A_n}$. Since the assumptions of Theorem 2.12 also hold here, it yields that $A_n/\sqrt{r_n^2\zeta_{1,\omega}^n/n} \xrightarrow{d} \mathcal{N}(0,1)$. The convergence in distribution implies that the corresponding characteristic functions converge uniformly on every bounded set, see for instance Kallenberg (2021, Theorem 6.3). Thus, for each $0 < K < \infty$ it holds that

$$\sup_{u\leq K}\left|\mathbb{E}\left[\exp\left(iuA_n/\sqrt{r_n^2\zeta_{1,\omega}^n/n}\right)\right] - e^{-\frac{u^2}{2}}\right| \to 0. \tag{2.45}$$

For

$$\nu := \frac{r_n^2\zeta_{1,\omega}^n/n}{r_n^2\zeta_{1,\omega}^n/n+\zeta_{r_n}^n/N} \leq 1,$$

we define $\phi_A(t) := e^{-\frac{t^2}{2}\nu}$. With (2.32), (2.45) and the upper bound for ν we get

$$\left|\mathbb{E}\left[\hat{\phi}_{A_n}(t)\right] - \phi_A(t)\right|$$

$$
\begin{aligned}
&= \left| \mathbb{E}\left[\exp\left(it \left(\frac{r_n^2 \zeta_{1,\omega}^n}{n} + \frac{\zeta_{r_n}^n}{N} \right)^{-1/2} A_n \right) \right] - \phi_A(t) \right| \\
&= \left| \mathbb{E}\left[\exp\left(it\nu^{1/2} A_n / \sqrt{r_n^2 \zeta_{1,\omega}^n / n} \right) \right] - e^{-\frac{t^2}{2}\nu} \right| \\
&\leq \sup_{u \leq t} \left| \mathbb{E}\left[\exp\left(iuA_n / \sqrt{r_n^2 \zeta_{1,\omega}^n / n} \right) \right] - e^{-\frac{u^2}{2}} \right| \leq \varepsilon, \qquad (2.46)
\end{aligned}
$$

if n is sufficiently large. Recalling the definitions of ϕ_B in (2.43) and ϕ_A above we use (2.34), (2.46), (2.44), $\mathbb{P}(D) \geq 1 - \varepsilon$ and $|e^{iu}| \leq 1$ to obtain

$$
\begin{aligned}
&|\phi_{A_n+B_n}(t) - \phi_A(t)\phi_B(t)| \\
&\quad = \left| \mathbb{E}\left[\hat{\phi}_A(t)\hat{\phi}_{B_n}(t) \right] - \phi_A(t)\phi_B(t) \right| \\
&\quad \leq \left| \mathbb{E}\left[\hat{\phi}_A(t)\hat{\phi}_{B_n}(t) - \hat{\phi}_A(t)\hat{\phi}_{B_n}(t)\mathbb{I}_D \right] \right| \\
&\qquad + \left| \mathbb{E}\left[\hat{\phi}_A(t)\hat{\phi}_{B_n}(t)\mathbb{I}_D - \hat{\phi}_A(t)\phi_B(t)\mathbb{I}_D \right] \right| \\
&\qquad + \left| \mathbb{E}\left[\hat{\phi}_A(t)\phi_B(t)\mathbb{I}_D - \hat{\phi}_A(t)\phi_B(t) \right] \right| \\
&\qquad + \left| \mathbb{E}\left[\hat{\phi}_A(t)\phi_B(t) \right] - \phi_A(t)\phi_B(t) \right| \\
&\quad \leq 2\mathbb{P}(D^C) + \mathbb{E}\left[|\hat{\phi}_{B_n}(t) - \phi_B(t)|\mathbb{I}_D \right] + \left| \mathbb{E}\left[\hat{\phi}_A(t) \right] - \phi_A(t) \right| \\
&\quad \leq 2\varepsilon + (\varepsilon + \delta') + \varepsilon \\
&\quad = 4\varepsilon + \delta'. \qquad (2.47)
\end{aligned}
$$

Further we can obtain

$$
\begin{aligned}
\phi_A(t)\phi_B(t) &= \exp\left(-\frac{t^2}{2} \frac{r_n^2 \zeta_{1,\omega}^n / n}{r_n^2 \zeta_{1,\omega}^n / n + \zeta_{r_n}^n / N} \right) \exp\left(-\frac{t^2}{2} \left((1-p) \cdot \frac{\zeta_{r_n}^n / N}{r_n^2 \zeta_{1,\omega}^n / n + \zeta_{r_n}^n / N} \right) \right) \\
&= \exp\left(-\frac{t^2}{2} \left(\frac{r_n^2 \zeta_{1,\omega}^n / n}{r_n^2 \zeta_{1,\omega}^n / n + \zeta_{r_n}^n / N} + \left(1 - \frac{N}{\binom{n}{r_n}} \right) \frac{\zeta_{r_n}^n / N}{r_n^2 \zeta_{1,\omega}^n / n + \zeta_{r_n}^n / N} \right) \right) \\
&= \exp\left(-\frac{t^2}{2} \right) \exp\left(-\frac{t^2}{2} \left(p \frac{\zeta_{r_n}^n / N}{r_n^2 \zeta_{1,\omega}^n / n + \zeta_{r_n}^n / N} \right) \right) \to \exp\left(-\frac{t^2}{2} \right) \qquad (2.48)
\end{aligned}
$$

because

$$
p \frac{\zeta_{r_n}^n / N}{r_n^2 \zeta_{1,\omega}^n / n + \zeta_{r_n}^n / N} \leq p \to 0.
$$

Combining (2.47) and (2.48) we get that

$$
\phi_{A_n+B_n}(t) \to \exp\left(-\frac{t^2}{2} \right),
$$

which completes the proof. □

2.4 Moments of Binomial denominators

Throughout the work we will need different results for the binomial distribution, especially for its moments and for moments of terms with a Binomial distributed denominator. These results are gathered in this section. Throughout the section B will denote a Binomial random variable. Its parameters will vary and will be stated in the results. The first lemma is similar to Lemma 4.1 by Györfi et al. (2002). The first statement in the lemma is a slight variation that gives equality instead of an upper bound.

Lemma 2.14. *Let $B \sim Bin(n, p)$, it holds that*

$$\mathbb{E}\left[\frac{1}{1+B}\right] = \frac{1}{(n+1)p}\left(1-(1-p)^{n+1}\right) \leq \frac{1}{(n+1)p}$$

and $$\mathbb{E}\left[\frac{\mathbb{I}\{B>0\}}{B}\right] \leq \frac{2}{(n+1)p}.$$

The next lemma gives us bounds for higher moments of this kind.

Lemma 2.15. *Let $B \sim Bin(n-1, p)$ and let $q \in \mathbb{N}$. The following inequalities hold*

$$\mathbb{E}\left[(1+B)^{-2}\right] \geq \frac{1}{p^2 n(n+1)}\left(1-(n+1)p(1-p)^n-(1-p)^{n+1}\right)$$
$$\mathbb{E}\left[(1+B)^{-q}\right] \leq \frac{q!}{p^q}\frac{(n-1)!}{(n+q-1)!} \leq \frac{q!}{p^q n^q}.$$

The proposition below is used to prove the two results that follow subsequently. It is given as Theorem 4 in the work by Skorski (2025).

Proposition 2.16. *Let $B \sim Bin(n, p)$ and $\sigma^2 = p(1-p)$. Then for any integer $q > 1$ we have*

$$\mathbb{E}\left[(B-np)^q\right]^{1/q} = C(n,p,q)\max\{k^{1-k/q}(n\sigma^2)^{k/q} : k = 1, \ldots, \lfloor q/2 \rfloor\},$$

where $C(n, p, q)$ is uniformly bounded by $(3e)^{-1} \leq C(n, p, q) \leq (5/2)^{1/5}e^{1/2}$.

Remark 2.17. Let $np > 1$, it holds that

$$\max\{k^{1-k/q}(n\sigma^2)^{k/q} : k = 1, \ldots, \lfloor q/2 \rfloor\} \leq (np)^{\lfloor q/2 \rfloor/q}\max\{k^{1-k/q} : k = 1, \ldots, \lfloor q/2 \rfloor\}.$$

Hence $\mathbb{E}[(B-np)^q] = \mathcal{O}((np)^{q/2})$ if q is fixed for all n.

The next two lemmas consider the moments of terms like

$$\frac{1}{B} - \frac{1}{np}.$$

We need upper bounds for the rate of these moments in n and p because in the proofs of the main results we will need to replace binomial random variables in the denominator by their expectation.

Lemma 2.18. *Let $B \sim Bin(n, p)$ such that $p^{-1} = \mathcal{O}(n)$. For any fixed integer $q > 1$ it holds that*

$$\mathbb{E}\left[\mathbb{I}\{B>0\}\left(\frac{1}{B}-\frac{1}{np}\right)^q\right] = \mathcal{O}\left((np)^{-3q/2}\right).$$

The last lemma is similar to the above, but instead of the indicator an added positive term ensures that the denominator is positive.

Lemma 2.19. *For any fixed integer $q > 1$ and $h \in \mathbb{N}$ with $1 \leq h \leq q$ let $B \sim Bin(n-h, p)$. If $p^{-1} = \mathcal{O}(n)$ it holds that*

$$\mathbb{E}\left[\left(\frac{1}{h+B} - \frac{1}{np}\right)^q\right] = \mathcal{O}\left((np)^{-3q/2}\right).$$

2.4.1 Proofs

We start with the proof of Lemma 2.14, which is from Györfi et al. (2002). We continue with the proofs of Lemma 2.15, Lemma 2.18 and Lemma 2.19.

2.4.1.1 Proof of Lemma 2.14

Using that $p \leq 1$, we obtain the first claim by

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{1+B}\right] &= \sum_{i=0}^{n} \frac{1}{i+1}\binom{n}{i} p^i (1-p)^{(n-i)} \\
&= \sum_{i=1}^{n+1} \frac{1}{i}\binom{n}{i-1} p^{i-1} (1-p)^{(n+1-i)} \\
&= \sum_{i=1}^{n+1} \frac{1}{i} \frac{n!}{(i-1)!(n+1-i)!} p^{i-1} (1-p)^{(n+1-i)} \\
&= \frac{1}{n+1} \sum_{i=1}^{n+1} \frac{(n+1)!}{i!(n+1-i)!} p^{i-1} (1-p)^{(n+1-i)} \\
&= \frac{1}{(n+1)p} \sum_{i=1}^{n+1} \binom{n+1}{i} p^i (1-p)^{(n+1-i)} \\
&= \frac{1}{(n+1)p} \left(1 - (1-p)^{n+1}\right) \\
&\leq \frac{1}{(n+1)p}.
\end{aligned}$$

Using the fact that

$$\frac{1}{l} \leq \frac{2}{1+l}$$

for every $l \in \mathbb{N}$ and the first part we obtain

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathbb{I}\{B>0\}}{B}\right] \leq \mathbb{E}\left[\frac{2}{1+B}\right] &= \frac{2}{(n+1)p}\left(1 - (1-p)^{n+1}\right) \\
&\leq \frac{2}{(n+1)p},
\end{aligned}$$

which completes the proof. □

2.4.1.2 Proof of Lemma 2.15

We get the lower bound by

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{(1+B)^2}\right] &= \sum_{i=0}^{n-1} \frac{1}{(1+i)^2}\binom{n-1}{i}p^i(1-p)^{n-1-i} \\
&\geq \sum_{i=0}^{n-1} \frac{1}{(1+i)(2+i)}\frac{(n-1)!}{i!(n-1-i)!}p^i(1-p)^{n-1-i} \\
&= \frac{1}{n(n+1)}\sum_{i=0}^{n-1} \frac{(n+1)!}{(i+2)!(n-1-i)!}p^i(1-p)^{n-1-i} \\
&= \frac{1}{p^2n(n+1)}\sum_{i=0}^{n-1} \binom{n+1}{i+2}p^{i+2}(1-p)^{n+1-(i+2)} \\
&= \frac{1}{p^2n(n+1)}\sum_{i=2}^{n+1} \binom{n+1}{i}p^{i}(1-p)^{n+1-i} \\
&= \frac{1}{p^2n(n+1)}\left(1-(n+1)p(1-p)^n-(1-p)^{n+1}\right).
\end{aligned}$$

For the upper bound we have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{(1+B)^q}\right] &= \sum_{i=0}^{n-1} \frac{1}{(1+i)^q}\binom{n-1}{i}p^i(1-p)^{n-1-i} \\
&= \sum_{i=0}^{n-1} \frac{1}{(1+i)^q}\frac{(n-1)!}{i!(n-1-i)!}p^i(1-p)^{n-1-i}.
\end{aligned}$$

For $j \geq 1$ we note that

$$\frac{1}{1+i} = 1-\frac{i}{i+1} \leq 1-\frac{i}{i+j} = \frac{j}{j+i}.$$

This implies

$$\frac{1}{(1+i)^q} = \prod_{j=1}^{q}\frac{1}{(1+i)} \leq \prod_{j=1}^{q}\frac{j}{j+i} = \prod_{j=1}^{q} j \prod_{j=1}^{q}\frac{1}{j+i} = q!\frac{i!}{(i+q)!}.$$

We get

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{(1+B)^q}\right] \\
&= \sum_{i=0}^{n-1} \frac{1}{(1+i)^q}\frac{(n-1)!}{i!(n-1-i)!}p^i(1-p)^{n-1-i} \\
&\leq \sum_{i=0}^{n-1} \frac{q!i!}{(i+q)!}\frac{(n-1)!}{i!(n-1-i)!}p^i(1-p)^{n-1-i}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{q!(n-1)!}{p^q(n+q-1)!} \sum_{i=0}^{n-1} \frac{i!}{(i+q)!} \frac{(n+q-1)!}{(i+q)!(n+q-1-(i+q))!} p^{i+q}(1-p)^{n+q-1-(i+q)} \\
&= \frac{q!(n-1)!}{p^q(n+q-1)!} \sum_{i=0}^{n-1} \binom{n+q-1}{i+q} p^{i+q}(1-p)^{n+q-1-(i+q)} \\
&= \frac{q!(n-1)!}{p^q(n+q-1)!} \sum_{i=q}^{n+q-1} \binom{n+q-1}{i} p^{i}(1-p)^{n+q-1-i} \\
&\leq \frac{q!(n-1)!}{p^q(n+q-1)!} \\
&= \frac{q!}{p^q} \prod_{j=0}^{q-1} \frac{1}{n+j} \\
&\leq \frac{q!}{p^q n^q}.
\end{aligned}
$$

The second inequality holds due to the binomial theorem and the last inequality is implied by $n + j \geq n$ if $j \geq 0$. □

2.4.1.3 Proof of Lemma 2.18

We rewrite the difference

$$
\begin{aligned}
\frac{1}{B} - \frac{1}{np} &= \frac{np - B}{Bnp} \\
&= \frac{np - B}{(np)^2} + \frac{np - B}{Bnp} - \frac{np - B}{(np)^2} \\
&= \frac{np - B}{(np)^2} + \frac{np(np - B)}{B(np)^2} - \frac{B(np - B)}{B(np)^2} \\
&= \frac{np - B}{(np)^2} + \frac{(np - B)^2}{B(np)^2} \\
&= \frac{np - B}{(np)^2} + \frac{(np - B)^2}{(np)^3} + \frac{(np - B)^2}{B(np)^2} - \frac{(np - B)^2}{(np)^3} \\
&= \frac{np - B}{(np)^2} + \frac{(np - B)^2}{(np)^3} + \frac{(np - B)^3}{B(np)^3}. \qquad (2.49)
\end{aligned}
$$

Hence we get

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{I}\{B > 0\} \left(\frac{1}{B} - \frac{1}{np}\right)^q\right] \\
&\lesssim \mathbb{E}\left[\left(\frac{np - B}{(np)^2}\right)^q\right] + \mathbb{E}\left[\left(\frac{(np - B)^2}{(np)^3}\right)^q\right] + \mathbb{E}\left[\mathbb{I}\{B > 0\} \left(\frac{(np - B)^3}{B(np)^3}\right)^q\right] \\
&\leq (np)^{-2q}\mathbb{E}\left[(np - B)^q\right] + (np)^{-3q}\mathbb{E}\left[(np - B)^{2q}\right] + (np)^{-3q}\mathbb{E}\left[(np - B)^{3q}\right].
\end{aligned}
$$

We apply Proposition 2.16 in the case where $p^{-1} = \mathcal{O}(n)$. We obtain

$$
\mathbb{E}\left[\mathbb{I}\{B > 0\} \left(\frac{1}{B} - \frac{1}{np}\right)^q\right]
$$

$$\begin{aligned}
&\lesssim (np)^{-2q}\mathbb{E}\left[(np-B)^q\right] + (np)^{-3q}\mathbb{E}\left[(np-B)^{2q}\right] + (np)^{-3q}\mathbb{E}\left[(np-B)^{3q}\right] \\
&\lesssim (np)^{-3q/2} + (np)^{-2q} + (np)^{-3q/2} \\
&\lesssim (np)^{-3q/2},
\end{aligned}$$

which yields the claim. □

2.4.1.4 Proof of Lemma 2.19

Using (2.49) from the previous proof and replacing B by $h+B$ we get

$$\begin{aligned}
\frac{1}{h+B} - \frac{1}{np} &= \frac{np-h-B}{(np)^2} + \frac{(np-h-B)^2}{(np)^3} + \frac{(np-h-B)^3}{(h+B)(np)^3} \\
&= \frac{(n-h)p-B-(1-p)h}{(np)^2} + \frac{((n-h)p-B-(1-p)h)^2}{(np)^3} \\
&\quad + \frac{((n-h)p-B-(1-p)h)^3}{(h+B)(np)^3}.
\end{aligned}$$

Hence

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{h+B} - \frac{1}{np}\right)^q\right] \\
&\lesssim \mathbb{E}\left[\left(\frac{(n-h)p-B-(1-p)h}{(np)^2}\right)^q\right] + \mathbb{E}\left[\left(\frac{((n-h)p-B-(1-p)h)^2}{(np)^3}\right)^q\right] \\
&\quad + \mathbb{E}\left[\left(\frac{((n-h)p-B-(1-p)h)^3}{(h+B)(np)^3}\right)^q\right] \\
&\leq (np)^{-2q}\mathbb{E}\left[((n-h)p-B-(1-p)h)^q\right] + (np)^{-3q}\mathbb{E}\left[((n-h)p-B-(1-p)h)^{2q}\right] \\
&\quad + (np)^{-3q}\mathbb{E}\left[((n-h)p-B-(1-p)h)^{3q}\right] \\
&\lesssim (np)^{-2q}\left(\mathbb{E}\left[((n-h)p-B)^q\right] + h^q\right) + (np)^{-3q}\left(\mathbb{E}\left[((n-h)p-B)^{2q}\right] + h^{2q}\right) \\
&\quad + (np)^{-3q}\left(\mathbb{E}\left[((n-h)p-B)^{3q}\right] + h^{3q}\right).
\end{aligned}$$

With Theorem 4 by Skorski (2025) we get

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{h+B} - \frac{1}{np}\right)^q\right] \\
&\lesssim (np)^{-2q}\left(\mathcal{O}\left(((n-h)p)^{q/2}\right) + h^q\right) + (np)^{-3q}\left(\mathcal{O}\left(((n-h)p)^q\right) + h^{2q}\right) \\
&\quad + (np)^{-3q}\left(\mathcal{O}\left(((n-h)p)^{3q/2}\right) + h^{3q}\right) \\
&= \mathcal{O}\left((np)^{-q3/2}\right),
\end{aligned}$$

which completes the proof. □

Chapter 3

Random forests

In this chapter, we introduce random forests, the central method of our work, in regression models. First, we provide a clear definition of regression trees, which were previously sketched in the introduction. This will enable us to provide the essential definition of a random forest. We will briefly discuss the effect of the number of trees in the random forest. In Section 3.1, the classical random forest, which employs the CART-split criterion by Breiman et al. (1984), is introduced. The connection of random forests to U-statistics, which is highly relevant for our work, is explained in Section 3.2. In the subsequent Section 3.3, we will introduce centered purely random forests, the random forest version we will consider throughout. In particular, we cover two specific versions, the well known uniform centered purely random forest and the newly proposed Ehrenfest centered purely random forest, and explain their characteristics. Section 3.4 deals with kernel random forests introduced by Scornet (2016b) and concludes the chapter.

For an explanation of regression trees and an illustration in Figure 1.1, we refer to the introduction of the thesis. Here, we give the formal definition of randomized regression trees in the regression model from (2.1). We use the random variable θ, which is independent of the training sample $\mathcal{D}_n$, to capture the subsampling and the randomness in the construction of the partition. This is rather generic, and the form of θ depends on the type of random partition that is used. Note that there is no standard domain space for θ. In particular, it is not a real number and is used for several random selections. Let $I_\theta \subset [n]$ denote the index set of the subsample selected by θ. The size of the subsamples, which is a tuning parameter of the random forest, is denoted by r_n. Further, let $A_n(x_0, \theta, \mathcal{D}_n)$ denote the cell in the final partition of the tree that contains x_0. The term

$$\sum_{j \in I_\theta} \mathbb{I}\{X_j \in A_n(x_0, \theta, \mathcal{D}_n)\}$$

is equal to the number of observations with index in I_θ that fall into this cell. Similar to the introduction, the tree estimator is defined as

$$\tilde{m}_n(x_0, \theta, \mathcal{D}_n) := \sum_{j \in I_\theta} Y_j \frac{\mathbb{I}\{X_j \in A_n(x_0, \theta, \mathcal{D}_n)\}}{\sum_{l \in I_\theta} \mathbb{I}\{X_l \in A_n(x_0, \theta, \mathcal{D}_n)\}}, \tag{3.1}$$

which is in line with our intuition for the construction: the estimator calculates the average of all Y_i values for which the X_i falls within the same cell of the feature space partition

as the argument x_0. The tree estimator is structurally similar to kernel estimators, and the indicator function is similar to a randomized box kernel. If one were to use a single regression tree outside of a random forest, one would use the entire training sample.

We use N to denote the number of trees in a random forest, which is another tuning parameter. Let $\theta_1, \ldots, \theta_N$ be i.i.d. copies of θ. We define the random forest estimator, which aggregates N trees by averaging, as

$$\begin{aligned}\hat{m}_{N,n}(x_0, (\theta_i)_{i=1}^N, \mathcal{D}_n) &:= \frac{1}{N}\sum_{i=1}^N \tilde{m}_n(x_0, \theta_i, \mathcal{D}_n) \\ &= \frac{1}{N}\sum_{i=1}^N \sum_{j \in I_{\theta_i}} Y_j \frac{\mathbb{I}\{X_j \in A_n(x_0, \theta_i, \mathcal{D}_n)\}}{\sum_{l \in I_{\theta_i}} \mathbb{I}\{X_l \in A_n(x_0, \theta_i, \mathcal{D}_n)\}}. \end{aligned} \tag{3.2}$$

This is a rather general form of the random forest estimator. Its variants depend on the choice of the distribution of θ, which controls the subsampling and the tree construction. Let $\mathbb{E}_\theta$ denote the expectation only with respect to θ. Connected to θ and the number of trees we define

$$\hat{m}_n(x_0, \mathcal{D}_n) := \mathbb{E}_\theta[\tilde{m}_n(x_0, \theta, \mathcal{D}_n)]. \tag{3.3}$$

By the law of large numbers we get the connection to the number of trees which is

$$\begin{aligned}\lim_{N\to\infty} \hat{m}_{N,n}(x_0, (\theta_i)_{i=1}^N, \mathcal{D}_n) &= \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^N \tilde{m}_n(x_0, \theta_i, \mathcal{D}_n) \\ &= \mathbb{E}_\theta[\tilde{m}_n(x_0, \theta, \mathcal{D}_n)] = \hat{m}_n(x_0, \mathcal{D}_n).\end{aligned}$$

It holds that

$$\begin{aligned}&\operatorname{Var}\left(\hat{m}_{N,n}(x_0, (\theta_i)_{i=1}^N, \mathcal{D}_n)\right) \\ &\quad= \operatorname{Var}\left(\mathbb{E}\left[\hat{m}_{N,n}(x_0, (\theta_i)_{i=1}^N, \mathcal{D}_n) \mid \mathcal{D}_n\right]\right) + \mathbb{E}\left[\operatorname{Var}\left(\hat{m}_{N,n}(x_0, (\theta_i)_{i=1}^N, \mathcal{D}_n) \mid \mathcal{D}_n\right)\right] \\ &\quad= \operatorname{Var}\left(\hat{m}_n(x_0, \mathcal{D}_n)\right) + \mathbb{E}\left[\operatorname{Var}\left(\frac{1}{N}\sum_{i=1}^N \tilde{m}_n(x_0, \theta_i, \mathcal{D}_n) \,\middle|\, \mathcal{D}_n\right)\right] \\ &\quad= \operatorname{Var}\left(\hat{m}_n(x_0, \mathcal{D}_n)\right) + \frac{1}{N}\mathbb{E}\left[\operatorname{Var}\left(\tilde{m}_n(x_0, \theta, \mathcal{D}_n) \mid \mathcal{D}_n\right)\right].\end{aligned}$$

The expectation of the estimator does not change with increasing N. This suggests that one would want to choose N as large as possible to minimize the variance. The larger N is, the closer the random forest is to the "infinite" random forest from (3.3). In practice, the returns of increasing N may be diminishing, such that a relatively small N can be sufficient. In the following sections, we will cover different types of random forests that correspond to different choices for the distribution of θ.

3.1 The original random forest and the CART-split criterion

In the introduction, we previously referenced the classical random forest variant that employs the CART-split criterion by Breiman et al. (1984) to determine the selection of

the cell/hyperrectangle splitting in the construction of a single tree. For completeness, we will describe this variant in this section because it is widely used and is usually meant when the forest variant is not further specified. To explain the criterion that is the main feature of this random forest, we assume that we are constructing a tree on the entire data set $\mathcal{D}_n$. Further consider a generic cell/hyperrectangle A and denote the number of independent variables X_i falling into A by $\#A$. A split or cut in A is a pair (j, z), where j is some integer in $\{1, \dots, p\}$ corresponding to an axis, and z is the position of the cut along the j-th coordinate, within the boundaries of A. We use $\mathcal{C}_A$ to denote the set of all possible cuts in A. We use the notation $X_i = (X_i^{(1)}, \dots, X_i^{(p)})$ for the individual features of one X_i. For any $(j, z) \in \mathcal{C}_A$ the CART-split criterion is given by

$$L_n(j,z) = \frac{1}{\#A} \sum_{i=1}^{n} (Y_i - \bar{Y}_A)^2 \mathbb{I}\{X_i \in A\} \\ - \frac{1}{\#A} \sum_{i=1}^{n} (Y_i - \bar{Y}_{A_{j,<z}} \mathbb{I}\{X_i^{(j)} < z\} - \bar{Y}_{A_{j,\geq z}} \mathbb{I}\{X_i^{(j)} \geq z\})^2 \mathbb{I}\{X_i \in A\},$$

where $A_{j,<z} = \{x \in A : x^{(j)} < z\}$, $A_{j,\geq z} = \{x \in A : x^{(j)} \geq z\}$, and $\bar{Y}_A$ (resp. $\bar{Y}_{A_{j,<z}}, \bar{Y}_{A_{j,\geq z}}$) is the average of the Y_i such that X_i belongs to A (resp. $A_{j,<z}, A_{j,\geq z}$). If no point is in a cell we define the average to be 0. For each cell A, the best cut (j_n^*, z_n^*) is selected by maximizing $L_n(j, z)$ over some set $\mathcal{M}_d \subset \{1, \dots, p\}$ and $\mathcal{C}_A$. In particular, we set

$$(j_n^*, z_n^*) \in \underset{j \in \mathcal{M}_d, (j,z) \in \mathcal{C}_A}{\arg\max} L_n(j, z).$$

For two $X_{i_1}^{(j)} < X_{i_2}^{(j)}$ that are neighbors, i.e. there is no observation X_{i_3} with $X_{i_1}^{(j)} < X_{i_3}^{(j)} < X_{i_2}^{(j)}$, all $z \in (X_{i_1}^{(j)}, X_{i_2}^{(j)})$ lead to the same value of $L_n(j, z)$. To remove these ties, the best splitting is always performed centered between $X_{i_1}^{(j)}$ and $X_{i_2}^{(j)}$. The intuition behind this criterion is to maximize the difference between the variance (of the response Y) in the original cell and the variance in the two cells after the cut. In other words, we want to minimize the variance in the two resulting cells after the cut. This makes sense since our goal is to use the average of the observations in one final cell of the tree as the estimator for all x in that cell. The complete random forest algorithm with some additional tuning parameters is summarized in Algorithm 3.1. The already introduced tuning parameter r_n determines the number of data points that are used to construct a single tree. The tuning parameter d_R is the number of feature space dimensions that are eligible to be cut by the CART criterion. The tuning parameter s_L determines the maximal number of data points in each leaf of the tree. Using these data points and the CART criterion, we construct a tree such that the leaves form a partition of $[0, 1]^p$.

We briefly discuss the role of θ in this particular type of random forest. One aspect to consider is the choice of whether the samples for the individual trees are generated with or without replacement. It should be noted that while this is not an explicit tuning parameter, it is determined by the rather abstract distribution of θ. If the subsampling is done with replacement, the algorithm runs in a bootstrap mode. In such cases, r_n is usually equal to n. If the subsampling is done without replacement, it is common to select $r_n < n$. The selection of d_R also has an impact on the "randomness" of the forest.

Algorithm 3.1 Random Forest

Input: Training set $\mathcal{D}_n$, number of trees N, $r_n \in [n]$, $d_R \in [p]$, $s_L \in [r_n]$ and $x_0 \in [0,1]^p$.

for $j = 1, \ldots, N$ **do**
  Select a subsample of size r_n with (or without) replacement, uniformly from $\mathcal{D}_n$.
  For the rest of this iteration only use the selected subsample.
  Set $\mathcal{P} = ([0,1]^p)$, the list containing the cell associated with the root of the tree.
  Set $\mathcal{P}_{final} = \emptyset$ an empty list.
  while $\mathcal{P} \neq \emptyset$ **do**
    Set A equal to the first entry of $\mathcal{P}$.
    if A contains less than s_L points or if all $X_i \in A$ are equal **then**
      Remove the cell A from the list $\mathcal{P}$.
      $\mathcal{P}_{final} \leftarrow (\mathcal{P}_{final}, A)$.
    else
      Select uniformly, without replacement, a subset $\mathcal{M}_d \subset [p]$ with $|\mathcal{M}_d| = d_R$.
      Select (j_n^*, z_n^*), the best split in A, by maximizing $L_n(j,z)$ for all $j \in \mathcal{M}_d$.
      Split the cell A at (j_n^*, z_n^*) and denote the resulting cells by A_L and A_R.
      Remove the cell A from the list $\mathcal{P}$.
      $\mathcal{P} \leftarrow (\mathcal{P}, A_L, A_R)$.
    end if
  end while
  Select $A_n(x_0, \theta, \mathcal{D}_n) \in \mathcal{P}_{final}$ containing x_0.
  Compute $\tilde{m}_n(x_0; \theta_j, \mathcal{D}_n)$ at x_0 equal to the average of the $Y_i \in A_n(x_0, \theta, \mathcal{D}_n)$.
end for
Compute the random forest estimate $\hat{m}_{N,n}(x_0; \theta_1, \ldots, \theta_N, \mathcal{D}_n)$ at the query point x_0.

Output: Prediction of the random forest at x_0.

Another critical tuning parameter is s_L, which denotes the size of the leaves in the trees. If $s_L = 1$, the trees are called fully grown. Intuitively, single fully grown trees appear to lack consistency due to their tendency towards overfitting.

Now, we discuss some possible choices for the tuning parameters to develop an intuition for their effects. Suppose we choose $r_n = n$ and do the subsampling without replacement, this simply means that each tree is constructed on the full data set $\mathcal{D}_n$. If we additionally choose $d_R = p$, there is no randomness left in the construction of each tree. This implies that every tree is exactly the same, and therefore the forest is also equal to the trees. This further implies that there is no advantage in using a forest over a tree in this case. Intuitively, it is clear that the forest benefits from diversity among the trees. Although this example is a very special case, it shows some effects of different tuning parameter selections.

One way to ensure diversity within the trees is to use a subsampling procedure that guarantees different subsamples such that the trees also differ. Using the bootstrap subsampling with $r_n = n$ in expectation, about one third of the data set is not included in each bootstrap subsample. Using subsampling without replacement, the diversity depends on the choice of r_n. The choice of d_R can also lead to diversity. A small value leads

to a greater chance of different cuts of the same cell via the CART-split criterion. The parameter θ captures both of these aspects.

Example 3.1. We describe a specific tree construction with the CART criterion and characterize the corresponding distribution of θ. To simplify the example, consider a tree where the number of splits is fixed and deterministic. Further, the subsamples are selected with replacement and contain $r_n = n$ observations. For $i = 1, \dots, n$ let U_i be i.i.d. random variables uniformly distributed on $\{1, \dots, n\}$. These random variables determine the subsample. We build all branches of the tree to the exact depth k. Thus, we omit the tuning parameter s_L. When building the tree to exact depth k, we need a total of $2^k - 1$ splits in total. Let S_i for $i \in \{1, \dots 2^k - 1\}$ be i.i.d. uniformly distributed on $\{M \subset \{1, \dots, p\}, |M| = d_R\}$. The S_i determine the subsets of eligible coordinates for the CART-splits. We can denote $\theta = ((U_j)_{j=1}^n, (S_i)_{i=1}^{2^k-1})$.

In the standard random forest, the tree depth is not deterministic because it depends on the previous splits in the tree. In the case of subsampling without replacement, the distributions of the U_i change accordingly.

3.2 Random forests as U-statistics

Before introducing the forest variants that we will mainly consider in our work, we explain the U-statistics perspective on random forests that we will use for these variants. There are multiple articles that interpret slightly adjusted random forests as generalized incomplete U-statistics, see e.g. Mentch and Hooker (2016) or Wager and Athey (2018). This section is loosely based on the work by Peng et al. (2022). The sample that we consider for the U-statistic version of a random forest is the training sample from the regression model (2.1).

For now, let us consider a fixed point x_0 at which we want to estimate the regression function m. To interpret a random forest as a U-statistic, the random forest needs to use subsampling without replacement and a subsample size r_n that is less than n. In particular, it does not use classical bootstrap subsampling. The randomized regression trees will be the kernel of the U-statistic applied to different subsets.

The trees from (3.1) and the kernel of a generalized U-statistic are both randomized in addition to their dependence on the training sample. The randomization of the regression tree done by θ also includes the subsampling. In Example 3.1 we denoted θ as a pair of random variables/vectors, where the first entry describes the subsampling and the second entry describes the randomization in the partition construction. We can do this similarly in the general case. The partition randomization part is denoted by a random variable ω. These are the random variables ω_I from (2.11) that randomize the kernel of the U-statistic.

We will slightly change the way the subsampling is done in the random forest. To get a generalized U-statistic as in the definition (2.11), we use Bernoulli random variables ρ_I for $I \in B_{r_n,n}$ with $\mathbb{P}(\rho_I = 1) = N/\binom{n}{r_n}$ to select the subsamples to be used in the random forest. This is a small difference from the original notion of a random forest. Instead of drawing an independent subsample for each tree, we use the ρ_I to randomize on which of the subsamples we will grow a tree. With the original method, the same subsample may occur more than once, but for typical choices of the tuning parameters this is rather

unlikely. Therefore, it should not be a drawback that using the same sample twice is not possible with the U-statistic. Another difference is that the number of trees for the U-statistic version is random and binomially distributed.

Given a subsample I, we denote the kernel of the random forest U-statistic by

$$h_n^{(T)}(x_0, (X_i, Y_i)_{i\in I}, \omega_I) := \sum_{j\in I} \frac{Y_j \mathbb{I}\{X_j \in A_n(x_0, \omega_I, (X_i, Y_i)_{i\in I})\}}{\sum_{l\in I} \mathbb{I}\{X_l \in A_n(x_0, \omega_I, (X_i, Y_i)_{i\in I})\}}.$$

This kernel is a single regression tree and $h_n^{(T)}$ is symmetric in the (X_i, Y_i) since we are summing over all elements in I. The kernel depends on the fixed $x_0 \in [0,1]^p$, but we will sometimes drop the x_0 in the notation of $h_n^{(T)}$. The random forest is

$$\begin{aligned} U^{(\mathrm{RF})}_{n,r_n,N,\omega}(x_0) &:= \frac{1}{\hat{N}} \sum_{I\in B_{r_n,n}} \rho_I h_n^{(T)}(x_0, (X_i, Y_i)_{i\in I}, \omega_I) \\ &= \frac{1}{\hat{N}} \sum_{I\in B_{r_n,n}} \rho_I \sum_{j\in I} \frac{Y_j \mathbb{I}\{X_j \in A_n(x_0, \omega_I, (X_i, Y_i)_{i\in I})\}}{\sum_{l\in I} \mathbb{I}\{X_l \in A_n(x_0, \omega_I, (X_i, Y_i)_{i\in I})\}} \end{aligned} \tag{3.4}$$

with $\hat{N}$ from (2.10). Using (3.4), a first approach is to apply Theorem 2.13 to get asymptotic normality of the random forest.

Note that in general this estimator will be biased for $m(x_0)$. For the parameter ϑ_n from the definition of a U-statistic, we have

$$\vartheta_n(x_0) = \mathbb{E}[h_n^{(T)}(x_0, (X_i, Y_i)_{i=1}^{r_n}, \omega)].$$

This expectation is a function in x_0 that should approximate m. The quality of this approximation depends on the tuning parameters of the random forest.

Despite the similarities between the random forest from the previous section and this generalized U-statistic they are different objects. The generalized U-statistic version is a random forest where two equal subsamples are not allowed and the number of trees is random and binomially distributed.

Based on this perspective on random forests, Theorem 2.13 allows us to prove a pointwise central limit theorem for a specific random forest. However, the result is rather abstract and, in particular, does not include the bias. Further, we need to know $\zeta^n_{1,\omega}$ and $\zeta^n_{r_n}$ to apply the theorem to a specific case. In general, these problems can be complex, which is why we will start to tackle them in the case of a simpler version of random forest, which will be introduced in the next section.

3.3 Centered purely random forests

The version of random forests we want to consider are centered purely random forests (CPRF). Purely random forests are random forests where the partitions of the trees do not depend on the training sample. Centered random forests means that each split in the tree construction is placed in the middle of the splitted cell.

For the tree construction, we perform $k \in \mathbb{N}$ iterations of splitting, and in each iteration every existing cell is split. This means that each final cell is the result of k splits, where

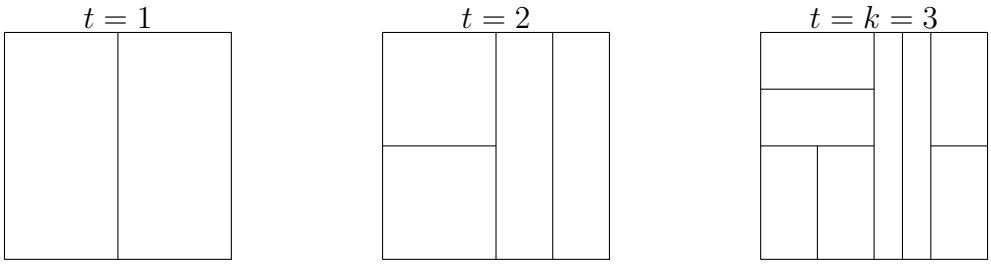

Figure 3.1: Centered hierarchical partition of a regression tree for $k = 3$ in three steps.

k is a tuning that is usually chosen dependent on n. For each split, a feature is randomly drawn from $\{1, \dots, p\}$ and the cell is split in the middle orthogonal to that feature. This implies that the volume of each cell is halved in each step. Thus, the procedure results in cells that all have a volume of 2^{-k}. An example of a realization of the hierarchical partition for $p = 2$ and $k = 3$ is illustrated in Figure 3.1.

For now we allow different distributions of the feature selection in the partition building. A possible choice is to draw the feature uniformly because we do not have any information to favor some features over others. Later, in Section 3.3.3, we will see a different way to draw the features that favors splits along the longer sides of a cell.

A similar type of random forest than the one we consider was first introduced by Breiman (2004). In this report the mean squared error of a random forest model is analyzed. The regression model therein has strong and weak features. The split probabilities for the strong and weak features are parameters of the algorithm. It is assumed that the splits of strong features are centered. Biau (2012) shows consistency in this model and proves bounds on the bias and variance terms. Both of these articles consider the infinite random forest estimator from equation (3.3). Klusowski (2021) improves the mean squared prediction error over the bound by Biau (2012). Under the assumption that the probabilities of selecting the feature to be cut are constant for all steps in the tree construction, he further proves that centered random forests do not achieve the minimax optimal rate of $n^{-\frac{2}{p+2}}$, see e.g. Tsybakov (2009), for the mean squared error in a regression model with Lipschitz continuous function m. He also discusses choosing these probabilities data dependent on the relevance of the features. This is similar to honest random forests, which we will consider as an extension in Chapter 7. Moreover, Klusowski (2021) shows that a uniform distribution over the set of relevant features achieves the best possible rate for these random forests.

Later in this section, we consider two types of CRPFs based on two different randomizations of the feature selection for the splits. First, we recall the well-studied CPRF, which selects features uniformly. Second, we propose the Ehrenfest CPRF, whose nonuniform split selection allows for minimax rate optimal estimation of the regression function. To achieve this, we will need to drop the assumption that the probabilities of selecting a feature are constant throughout the tree construction. Before discussing the two variants, we will cover some general characteristics of CPRFs.

The important difference and simplification of (centered) purely random forests compared to classical ones is that the partitions do not depend on the training sample $\mathcal{D}_n$. Therefore, we can omit $\mathcal{D}_n$ in the notation of the cells. In particular, we use $A_k(x_0, \omega)$

to denote the cell that contains x_0 and is the result of the splitting random variable ω. Indirectly it still depends on n but the index k emphasizes that k is the tuning for the cell size. Before we give the definition of a CPRF, we note that the definition consists of two parts, both of which are essential. The first part of the definition deals only with the partition construction, and the second part defines the estimator based on this partition.

Definition 3.2 (Centered purely random forest (CPRF))**.** 1. For $k \in \mathbb{N}$ let $t \in [k]$ and $j \in \{1, \dots, 2^{t-1}\}$. For $p \in \mathbb{N}$ the random variables $D_{t,j} \in [p]$ create a centered hierarchical partition with depth k of the feature space $[0,1]^p$. Starting with $[0,1]^p$, the partition is constructed in k steps, where in each step each cell in the partition is split into two new cells. For each splitting, a specific coordinate is selected, and the splitting is performed orthogonally at the midpoint of the interval corresponding to the coordinate. In the t-th step the $(D_{t,j})_{j=1}^{2^{t-1}}$ decide which coordinates are split in the 2^{t-1} existing cells. Let ω be a random variable aggregating the information in the $D_{t,j}$, then we denote the final partition as

$$\{A_k^{(i)}(\omega) \mid i \in \{1, \dots, 2^k\}\}.$$

For $x_0 \in [0,1]^p$ let $A_k(x_0, \omega)$ denote the unique element of the partition with $x_0 \in A_k^{(i)}(\omega)$. For $l \in [p]$, $t \in [k]$ let $S_{l,t}(x_0, \omega)$ denote the number of splits orthogonal to the l-th coordinate that were used in the first t steps in the construction of $A_k(x_0, \omega)$. They satisfy $\sum_{l=1}^{p} S_{l,t}(x_0, \omega) = t$. For $t = k$ we abbreviate $S_{l,k}(x_0, \omega) = S_l(x_0, \omega)$ and assume that $S_l(x_1, \omega) \stackrel{d}{=} S_l(x_2, \omega)$ for all $l \in [p]$ and $x_1, x_2 \in [0,1]^p$. We call a centered purely random forest symmetric if $S_l(x_0, \omega) \stackrel{d}{=} S_1(x_0, \omega)$ for all $l \in [p]$.

2. Let $r_n \in \mathbb{N}$ with $r_n < n$, for $I \in B_{r_n,n}$ let ω_I be i.i.d. copies of ω. The centered purely random forest regression estimator (in U-statistic form, see Definition 2.9 and (3.4)) on the training sample $\mathcal{D}_n = (X_i, Y_i)_{i=1}^n$ is

$$U_{n,r_n,N,\omega}^{(\mathrm{RF})}(x_0) = \frac{1}{\hat{N}} \sum_{I \in B_{r_n,n}} \rho_I h_n^{(T)}(x_0, (X_i, Y_i)_{i \in I}, \omega_I), \tag{3.5}$$

with kernel

$$h_n^{(T)}(x_0, (X_i, Y_i)_{i \in I}, \omega_I) = \sum_{j \in I} \frac{Y_j \mathbb{I}\{X_j \in A_k(x_0, \omega_I)\}}{\sum_{i \in I} \mathbb{I}\{X_i \in A_k(x_0, \omega_I)\}}.$$

We define $h_n^{(T)}(x_0, (X_i, Y_i)_{i \in I}, \omega_I) = 0$ if

$$\sum_{i \in I} \mathbb{I}\{X_i \in A_k(x_0, \omega_I)\} = 0.$$

Remark 3.3. Throughout, we consider centered purely random forests whose $(D_{t,j})_{j=1}^{2^{t-1}}$ are conditionally independent given $\{D_{s,j} \mid s < t, 1 \le j \le 2^{s-1}\}$ for all t. This means that given a partition at step t, the subsequent partitioning of its cells is independent of each other. From here on we will usually use ω instead of the $D_{t,j}$ to shorten the notation.

Remark 3.4. The centered splitting implies that the volume of the cells is halved in each step. In particular, the volume of all cells in the final partition is equal to 2^{-k}. The average length of the cells along a single coordinate is $2^{-k/p}$. Thus, this tuning parameter has a similar effect as the bandwidth for kernel regression estimators.

Remark 3.5. The exact form of $A_k(x_0, \omega)$ is determined by $S(x_0, \omega) = (S_l(x_0, \omega))_{l=1}^p$ and the order of the splits does not impact the form of the cell. For $x \in [0,1]^p$ we denote $x = (x^{(1)}, \dots, x^{(p)})$. The projection of the cell $A_k(x_0, \omega)$ on the l-th coordinate is

$$A_k^{(l)}(x_0, \omega) := \{\xi \in [0,1] : \exists x \in A_k(x_0, \omega), x^{(l)} = \xi\}. \tag{3.6}$$

Definition 3.2 implies that this interval has length $2^{-S_l(x_0,\omega)}$ and that its endpoints are in $2^{-S_l(x_0,\omega)}\mathbb{N}_0$. Therefore it holds that

$$A_k^{(l)}(x_0, \omega) = 2^{-S_l(x_0,\omega)}\big(\lfloor x_0^{(l)} 2^{S_l(x_0,\omega)} \rfloor, \lfloor x_0^{(l)} 2^{S_l(x_0,\omega)} \rfloor + 1\big]. \tag{3.7}$$

This does not depend on the order of the splits, and thus $A_k(x_0, \omega)$, the Cartesian product of the $A_k^{(l)}(x_0, \omega)$, does not depend on the order of the splits. The choice that the cells are open at the left endpoint and closed at the right endpoint is arbitrary.

Remark 3.6. A CPRF is not necessarily symmetric, but in the absence of any assumptions about the features, it is a natural assumption to treat them equally. When the regression model includes assumptions about the features, for example, having strong and weak features as considered by Breiman (2004) and Biau (2012), symmetry is not appropriate. From now on, the random forests studied in this thesis will all be symmetric.

The notation $U_{n,r_n,N,\omega}^{(\mathrm{RF})}$ for the CPRF does not differ from the one in (3.4). We can keep the shorter notation for convenience because we will only consider the CPRF from here on. Let us denote

$$U_{n,r_n,N,\omega}^{(m)}(x_0) := \frac{1}{\hat{N}} \sum_{I \in B_{r_n,n}} \rho_I \sum_{j \in I} m(X_j) \frac{\mathbb{I}\{X_j \in A_k(x_0, \omega_I)\}}{\sum_{i \in I} \mathbb{I}\{X_i \in A_k(x_0, \omega_I)\}} \tag{3.8}$$

and

$$U_{n,r_n,N,\omega}^{(\varepsilon)}(x_0) := \frac{1}{\hat{N}} \sum_{I \in B_{r_n,n}} \rho_I \sum_{j \in I} \varepsilon_j \frac{\mathbb{I}\{X_j \in A_k(x_0, \omega_I)\}}{\sum_{i \in I} \mathbb{I}\{X_i \in A_k(x_0, \omega_I)\}}, \tag{3.9}$$

such that we have the decomposition

$$U_{n,r_n,N,\omega}^{(\mathrm{RF})}(x_0) = U_{n,r_n,N,\omega}^{(m)}(x_0) + U_{n,r_n,N,\omega}^{(\varepsilon)}(x_0).$$

We note that $U_{n,r_n,N,\omega}^{(m)}$ and $U_{n,r_n,N,\omega}^{(\varepsilon)}$ are both generalized incomplete U-statistics. For any set $I \subset [n]$, with $|I| = r_n$ and $j \in I$ let us further denote

$$W_{j,k}(x_0, \omega, I) := \frac{\mathbb{I}\{X_j \in A_k(x_0, \omega)\}}{\sum_{i \in I} \mathbb{I}\{X_i \in A_k(x_0, \omega)\}}.$$

To simplify the notation we use

$$W_{j,k}(x_0, I) := W_{j,k}(x_0, \omega_I, I) \qquad \text{and} \qquad W_{j,k}(x_0, \omega) := W_{j,k}(x_0, \omega, [r_n]). \tag{3.10}$$

The kernels of the two U-statistics in (3.8) and (3.9) are

$$h_n^{(m)}(x_0, (X_j)_{j=1}^{r_n}, \omega) = \sum_{j=1}^{r_n} m(X_j) \frac{\mathbb{I}\{X_j \in A_k(x_0, \omega)\}}{\sum_{i=1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0, \omega)\}} = \sum_{j=1}^{r_n} m(X_j) W_{j,k}(x_0, \omega)$$

and

$$h_n^{(\varepsilon)}(x_0, (X_j, \varepsilon_j)_{j=1}^{r_n}, \omega) = \sum_{j=1}^{r_n} \varepsilon_j \frac{\mathbb{I}\{X_j \in A_k(x_0, \omega)\}}{\sum_{i=1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0, \omega)\}} = \sum_{j=1}^{r_n} \varepsilon_j W_{j,k}(x_0, \omega). \tag{3.11}$$

The kernel $h_n^{(\varepsilon)}$ and thus also $U_{n,r_n,N,\omega}^{(\varepsilon)}$ have expectation zero due to the independence of the ε_j from X_j and ω. We note that the $W_{j,k}(x_0, \omega, I)$ are the weights of the observations in the kernel applied to the subsample I. If there is at least one observation in $A_k(x_0, \omega)$ the sum of the weights is equal to one. Otherwise Definition 3.2 yields that their sum is zero.

Similar to (2.4) we observe that $U_{n,r_n,N,\omega}^{(m)}$ is the expectation of $U_{n,r_n,N,\omega}^{(\mathrm{RF})}(x_0)$ conditioned on the observations X_j, the ω_I and the ρ_I. We have

$$\begin{aligned}
&\mathbb{E}\left[U_{n,r_n,N,\omega}^{(\mathrm{RF})}(x_0) \middle| (X_j)_{j=1}^n, (\omega_I)_{I \in B_{r_n,n}}, (\rho_I)_{I \in B_{r_n,n}}\right] \\
&= \frac{1}{\hat{N}} \sum_{I \in B_{r_n,n}} \rho_I \sum_{j \in I} \mathbb{E}[Y_j \mid X_j] W_{j,k}(x_0, I) \\
&= \frac{1}{\hat{N}} \sum_{I \in B_{r_n,n}} \rho_I \sum_{j \in I} m(X_j) W_{j,k}(x_0, I) \\
&= U_{n,r_n,N,\omega}^{(m)}(x_0).
\end{aligned} \tag{3.12}$$

In general, this only holds for purely random forests because otherwise the cells depend on the errors through the Y_i. In particular, we do not necessarily have

$$\mathbb{E}\left[\varepsilon_1 \frac{\mathbb{I}\{X_1 \in A_n(x_0, \omega, (X_i, Y_i)_{i=1}^{r_n})\}}{\sum_{l=1}^{r_n} \mathbb{I}\{X_l \in A_n(x_0, \omega, (X_i, Y_i)_{i=1}^{r_n})\}} \,\middle|\, (X_j)_{j=1}^{r_n}, \omega\right] = 0.$$

Using the terms from (3.8) and (3.9), we can decompose the error of the estimator into two parts

$$U_{n,r_n,N,\omega}^{(\mathrm{RF})}(x_0) - m(x_0) = U_{n,r_n,N,\omega}^{(\varepsilon)}(x_0) + U_{n,r_n,N,\omega}^{(m)}(x_0) - m(x_0). \tag{3.13}$$

Just like for the histogram, we call $U_{n,r_n,N,\omega}^{(m)} - m$ the approximation error. The identity in (3.12) illustrates that this is the part of the error that is due to the estimator's general ability to approximate m. We call $U_{n,r_n,N,\omega}^{(\varepsilon)}$ the stochastic error, but note that there are alternative names in the literature, see for instance Biau (2012), who uses the term estimation error.

3.3.1 Characteristics of centered purely random forests

In this section, we introduce three key characteristics of centered purely random forests that will be important for the results in the following chapters. Furthermore, we discuss the interactions between these characteristics.

3.3.1.1 The diameter of the cells

As we have already seen for the histogram estimator, the diameter of a cell in the partition is important for the ability of the estimator to approximate the regression function. We start with some observations that hold for all distributions of the splits. Remark 3.5 explains that the length of the cell $A_k(x_0, \omega)$ along the l-th coordinate is equal to $2^{-S_l(x_0,\omega)}$. For $A_k(x_0, \omega)$ we obtain

$$\mathfrak{d}(A_k(x_0, \omega)) = \Big(\sum_{l=1}^{p} 2^{-2S_l(x_0,\omega)} \Big)^{1/2}.$$

We are interested in bounds for the diameter and its moments. If $q \leq 2$, it holds that $\|\xi\|_{2/q} \leq \|\xi\|_1$ for any $\xi \in \mathbb{R}^p$. This implies

$$\mathfrak{d}(A_k(x_0, \omega))^q = \left(\sum_{l=1}^{p} 2^{-2S_l(x_0,\omega)} \right)^{q/2} \leq \sum_{l=1}^{p} 2^{-qS_l(x_0,\omega)}.$$

Taking the expectation we obtain

$$\mathbb{E}\left[\mathfrak{d}(A_k(x_0, \omega))^q\right] \leq \mathbb{E}\left[\sum_{l=1}^{p} 2^{-qS_l(x_0,\omega)} \right] = p\mathbb{E}\left[2^{-qS_1(x_0,\omega)}\right]. \tag{3.14}$$

The last equality holds due to the symmetry that implies $S_l(x_0, \omega) \stackrel{d}{=} S_1(x_0, \omega)$ for all $l \in [p]$. Without additional assumptions on the distribution of $S_l(x_0, \omega)$, Jensen's inequality yields a lower bound via

$$\mathbb{E}\left[2^{-qS_1(x_0,\omega)}\right] \geq 2^{-q\mathbb{E}[S_1(x_0,\omega)]} = 2^{-qk/p}. \tag{3.15}$$

Equality in this lower bound holds in the deterministic case $S_l(x_0, \omega) = k/p$.

For confidence intervals and bands, we need the stochastic error to be the dominant term in the error decomposition. We will see that the expected absolute approximation error depends on the expected diameter, similar to (2.6) for the histogram estimator. Therefore, it is desirable to have a small diameter close to the optimal rate to get a good bound for the approximation error.

3.3.1.2 Cell intersections

Another important characteristic of a CPRF is the expected volume of $A_k(x_0, \omega_1) \cap A_k(x_0, \omega_2)$ for i.i.d. copies ω_1 and ω_2 of ω. For any $A \subset \mathbb{R}^p$ we denote the volume of A by

$$\mathbb{V}(A) := \int 1_A(x)dx. \tag{3.16}$$

Further let us denote the expected volume of the intersection

$$\mathcal{V}_{\cap,k} := \mathbb{E}\left[\mathbb{V}(A_k(x_0, \omega_1) \cap A_k(x_0, \omega_2))\right]. \tag{3.17}$$

The intersection is not empty because it contains x_0. Without loss of generality let $S_l(x_0, \omega_1) \geq S_l(x_0, \omega_2)$. In this case, Remark 3.5 implies for $A_k^{(l)}$ from (3.6) that

$$A_k^{(l)}(x_0, \omega_1) \cap A_k^{(l)}(x_0, \omega_2) = A_k^{(l)}(x_0, \omega_1), \tag{3.18}$$

and in particular, the diameter of this intersection is the minimum of the diameters. For the volume of the intersection this yields

$$\begin{aligned}
\mathbb{V}(A_k(x_0, \omega_1) \cap A_k(x_0, \omega_2)) &= \int_{A_k(x_0,\omega_1) \cap A_k(x_0,\omega_2)} dx \\
&= \prod_{l=1}^{p} \int_{A_k^{(l)}(x_0,\omega_1) \cap A_k^{(l)}(x_0,\omega_2)} dx^{(l)} \\
&= \prod_{l=1}^{p} \min\{2^{-S_l(x_0,\omega_1)}, 2^{-S_l(x_0,\omega_2)}\} \\
&= \prod_{l=1}^{p} 2^{-\max\{S_l(x_0,\omega_1), S_l(x_0,\omega_2)\}} \\
&= 2^{-\sum_{l=1}^{p} \max\{S_l(x_0,\omega_1), S_l(x_0,\omega_2)\}}.
\end{aligned} \tag{3.19}$$

Hence, we can write $\mathcal{V}_{\cap,k}$ as the expectation of (3.19). Note that the expectation of this does not depend on x_0, because Definition 3.2 yields that the distribution of the splits $S_l(x_0, \omega_1)$ and $S_l(x_0, \omega_2)$ does not depend on x_0. The fact that $\sum_{l=1}^{p} S_l(x_0, \omega) = k$ for all ω implies that

$$\begin{aligned}
2^{-2k} &= 2^{-\sum_{l=1}^{p} (S_l(x_0,\omega_1) + S_l(x_0,\omega_2))} \\
&\leq 2^{-\sum_{l=1}^{p} \max\{S_l(x_0,\omega_1), S_l(x_0,\omega_2)\}} \\
&\leq 2^{-\sum_{l=1}^{p} S_l(x_0,\omega_1)} \\
&= 2^{-k}.
\end{aligned}$$

Applying the expectation to the above and using (3.19) we obtain

$$2^{-2k} \leq \mathcal{V}_{\cap,k} \leq 2^{-k}. \tag{3.20}$$

We note that

$$\sum_{l=1}^{p} \max\{S_l(x_0, \omega_1) S_l(x_0, \omega_2)\} = k + \frac{1}{2} \sum_{l=1}^{p} |S_l(x_0, \omega_1) - S_l(x_0, \omega_2)|.$$

Thus, we get

$$\mathcal{V}_{\cap,k} = \mathbb{E}\left[2^{-\sum_{i=1}^{p} \max\{S_l(x_0,\omega_1) S_l(x_0,\omega_2)\}}\right] = 2^{-k} \mathbb{E}\left[2^{-\frac{1}{2}\sum_{l=1}^{p} |S_l(x_0,\omega_1) - S_l(x_0,\omega_2)|}\right]. \tag{3.21}$$

This can be helpful to prove a better lower bound for $\mathcal{V}_{\cap,k}$ if we know more about the distribution of the splits.

3.3.1.3 The finest partition of the feature space

For a fixed distribution of the random variable ω we call a subset $A \subset [0,1]^p$ of the feature space undividable, if for all realizations w of ω there exists an $x \in [0,1]^p$ with $A \subset A_k(x,w)$. This means that A is a subset of some cell in any partition. We call a set A maximum undividable if there is no other undividable set $\tilde{A} \neq A$ with $A \subset \tilde{A}$. Let $\mathcal{S}_u$ denote the set of all maximum undividable sets. Then $\mathcal{S}_u$ is a partition of the feature space. This is implied by the following two observations. First, two maximum undividable sets are always disjoint by definition. Second, every element x in the feature space must be contained in an undividable subset, since $\{x\}$ is undividable. For different distributions of ω, the size of the maximum undividable sets can vary, which directly affects the number of elements in the partition $\mathcal{S}_u$. The partition $\mathcal{S}_u$ implies an equivalence relation on $[0,1]^p$. If x_1 and x_2 are in the same maximum undividable set, we have

$$A_k(x_1,\omega) = A_k(x_2,\omega)$$

almost surely. We note that this also implies that every realization of $A_k(x,\omega)$ can be represented as a disjoint union of maximum undividable sets.

Let us denote $N_f(k) := |\mathcal{S}_u|$ and let $\mathcal{X}_k \subset [0,1]^p$ with $|\mathcal{X}_k| = N_f(k)$ be a set containing exactly one element in each of the maximum undividable sets. The above explanation shows that $N_f(k)$ and $\mathcal{X}_k$ can depend on the distribution of ω. We do not capture this in their notation for convenience, so we need to keep this dependence in mind. We are interested in $N_f(k)$ for the distributions of ω that we consider.

For $\mu_l \in \{0, 2^{-k}, \ldots, 1-2^{-k+1}, 1-2^{-k}\}$, let us consider the hypercubes of the form

$$\prod_{l=1}^{p} \left[\mu_l, \mu_l + 2^{-k}\right]. \tag{3.22}$$

These hypercubes form a sufficiently fine partition such that each cube cannot be split by a centered tree of depth k. This is the case because there can be at most k splits in one direction. The number of cubes is 2^{kp}, which implies that $N_f(k) \leq 2^{kp}$ for any CPRF. If we consider a CPRF where all splits are always drawn uniformly from $[p]$, the probability of splitting one feature k times is greater than zero. Therefore, we have $N_f(k) = 2^{kp}$ for this type of CPRF.

3.3.1.4 The interaction of the characteristics

The three characteristics of CPRFs that we introduced above affect each other because they all depend on the same object, which is the distribution of the splits.

We already mentioned that the expected diameter is preferred to be small, because it directly affects the approximation error. We will see that the variance of the leading term of the stochastic error U-statistic scales linearly with $2^{2k}\mathcal{V}_{\cap,k}$. To construct confidence intervals and bands we will need that the stochastic error dominates the approximation error. For this dominance it is necessary that the variance of the stochastic error and thus $\mathcal{V}_{\cap,k}$ is large enough. Further, $\mathcal{V}_{\cap,k}$ is a measure for the similarity of the trees in a CPRF. If the trees are more similar at some point x_0, the intersections of the cells around

x_0 will be larger on average. One would expect a random forest to benefit from having more diverse trees. This is reflected by the effect of $\mathcal{V}_{\cap,k}$ on the variance.

For the confidence band results we will need uniform bounds on the approximation error and different remainder terms. We will see that the supremum over $x_0 \in [0,1]^p$ will in fact be a maximum over $x_0 \in \mathcal{X}_k$. This will allow us to use union bounds in which the sum has $N_f(k)$ terms. To handle the union bounds, we will exploit that a finite number of moments of ε exist. In general, the number of existing moments needs to be larger if $N_f(k)$ is larger. Thus, a smaller $N_f(k)$ leads to weaker assumptions on the errors. To achieve that $N_f(k) < 2^{kp}$, one needs that some of 2^{kp} cubes from (3.22) are never split. In other words, one needs to restrict the diversity of the partitions, which will also lead to a larger $\mathcal{V}_{\cap,k}$.

In summary, we face a trade-off when selecting the distribution of the splits. On the one hand, distributions with more diverse partitions lead to a smaller variance of the random forest, which is generally desirable, and will lead to smaller confidence sets. On the other hand, less diverse partitions lead to a smaller expected diameter, and thus a smaller approximation error bound, because their cells are closer to cubes. A smaller approximation error is helpful for the dominance of the stochastic error and for the construction of confidence sets. A larger $\mathcal{V}_{\cap,k}$ will also help this dominance of the stochastic error. The final advantage of less diverse partitions is a smaller $N_f(k)$, which allows weaker assumptions on the error moments when proving the confidence band results.

Earlier we mentioned that the expected diameter is an essential part of the bound for the approximation error. However, we need to note that the effect of the partition diversity on the approximation error is not fully captured by this bound. We know that any CPRF estimator is constant on the undividable cells. Consider an approximation of a continuous function by a function that is piecewise constant on the cells of some grid. As the grid is refined, the class of piecewise constant functions on the finer grid is a superset of the class of constant functions on the rougher grid. In the larger function class, we can always find an approximation that is at least as good as the best approximation from the smaller class. The estimator for m produced by an algorithm with a larger $N_f(k)$ is an element of a larger function class than an algorithm with a smaller $N_f(k)$. In other words, an algorithm with a small number of undividable sets is limited by fewer elements in the image of its estimated regression function. This effect opposes the effect of the smaller expected diameter. However, it is not clear whether this effect actually results in a smaller approximation error. The magnitude of both effects can depend on the shape of the specific regression function. In general, it is not clear how to improve the bound for the approximation error uniformly in a Hölder class for the regression functions. Still, it is important to be aware of both effects, for example when interpreting the results of simulations.

The two variants of CPRFs, which we will present in the next two sections, will have a different diversity of their partitions and thus their characteristics will differ. The Ehrenfest CPRF will restrict the diversity of the cells through the construction algorithm. Therefore, it will have a smaller approximation error bound, and in Chapter 4 we will see that it can reach the minimax optimal rate for the mean squared error. The uniform CPRF will have more diverse partitions. In Chapter 5 we will see that one can still prove confidence bands for this RF version. However, the assumptions on n, r_n and k for the

confidence band result will be less flexible, because the approximation error bound will be larger and $\mathcal{V}_{\cap,k}$ will be smaller. Further, one needs more existing moments of ε for the confidence band result to hold. A small advantage is that their confidence band radius is slightly smaller, a nuance that remains ambiguous within the theory. However, this distinction but will be observed in the simulations in Chapter 6.

3.3.2 The uniform centered purely random forest

The uniform CPRF is characterized and named by the fact that the probability of splitting each feature is $1/p$. Purely random forest types similar to the uniform CPRF were already considered in the literature, for instance by Breiman (2004) or Biau (2012). When constructing the tree, all splits in all steps are drawn independently of each other. More precisely the $D_{t,j}$ from Definition 3.2 are i.i.d. and uniformly distributed on $[p]$. Let $x_0 \in [0,1]^p$ be fixed, without loss of generality the cell $A_k(x_0,\omega)$ is constructed by $(D_{t,1})_{t=1}^k$. It then holds that $S_l(x_0,\omega) = \sum_{t=1}^k \mathbb{I}\{D_{t,1} = l\}$ and thus $S_l(x_0,\omega) \sim \mathrm{Bin}(k, 1/p)$.

In a multivariate regression model with Lipschitz continuous regression function Klusowski (2021) proves an upper bound of the rate of the mean squared prediction error of this random forest. The rate does not match the minimax optimal rate for nonparametric regression and Klusowski (2021) further proves that the rate cannot be improved.

Let us consider the characteristics from Section 3.3.1 for the uniform CPRF. In Section 3.3.1.3 we already argued that $N_f(k) = 2^{kp}$ for the uniform CPRF. The binomial distribution of $S_l(x_0,\omega)$ can be utilized to derive a bound on the diameter by employing the moment-generating function of the Binomial distribution. For $q \leq 2$ (3.14) yields

$$\begin{aligned}\mathbb{E}\left[\mathfrak{d}(A_k(x_0,\omega))^q\right] &\leq p\mathbb{E}\left[2^{-qS_1(x_0,\omega)}\right] \\ &= p\mathbb{E}\left[\exp\left(\log(2^{-q})S_1(x_0,\omega)\right)\right] = p\left(\frac{p-1+2^{-q}}{p}\right)^k. \end{aligned} \tag{3.23}$$

The lower bound in (3.15) implies that the minimal rate for any distribution of $S_1(x_0,\omega)$ is $2^{-kq/p}$. Here we have a worse rate because

$$\frac{p-1+2^{-q}}{p} \geq 2^{-q/p}. \tag{3.24}$$

For $p = 1$ equality holds in (3.24), which is obvious because in this case the splitting is deterministic. If $p \geq 2$, (3.24) can be proved by plugging $2^{-q} \leq 1$ into $g(\xi) := p-1+\xi-p\xi^p$ and using that $g(\xi) \geq 0$ for all $\xi \in [0,1]$. This holds because g is concave on $[0,1]$, $g(0) = p-1 > 0$ and $g(1) = 0$.

We proceed with $\mathcal{V}_{\cap,k}$. To get the mean squared error rate Klusowski (2021) proves an upper bound for the approximation error of centered random forests. One major step in proving this is to bound a term analogous to $\mathcal{V}_{\cap,k}$. This bound relies on Klusowski (2021, Lemma S.1.) from the supplementary material of the article. We will use this lemma to bound

$$\mathbb{E}\left[2^{-\frac{1}{2}\sum_{l=1}^p |S_l(x_0,\omega_1)-S_l(x_0,\omega_2)|}\right].$$

We note that $(S_l(x_0,\omega_1))_{l=1}^p$ is multinomial distributed with p trials and probabilities all equal to $1/p$. The same holds for its independent copy $(S_l(x_0,\omega_2))_{l=1}^p$. The lemma yields

that

$$\frac{p^p}{(47)^p k^{p-1}} \leq \mathbb{E}\left[2^{-\frac{1}{2}\sum_{l=1}^{p}|S_l(x_0,\omega_1)-S_l(x_0,\omega_2)|}\right] \leq \frac{8^p p^{p/2}}{\sqrt{k^{p-1}}}.$$

With (3.21), this implies

$$2^{-k}\frac{p^p}{(47)^p k^{p-1}} \leq \mathcal{V}_{\cap,k} \leq 2^{-k}\frac{8^p p^{p/2}}{\sqrt{k^{p-1}}},$$

and for the rate in k we get

$$2^{-k}k^{-(p-1)} \lesssim \mathcal{V}_{\cap,k} \lesssim 2^{-k}k^{-(p-1)/2}. \tag{3.25}$$

3.3.3 The Ehrenfest centered purely random forest

We will introduce the following model to get a random forest that has the optimal rate for the approximation error. To achieve this, the diameter needs to be of the same rate as the diameter of a hypercube. This is in a sense closer to a histogram than the RF with uniform distribution for feature selection and the trees are less diverse. This will also imply that $\mathcal{V}_{\cap,k}$ is larger due to the reduced diversion in the partitions. As mentioned earlier, the result by Klusowski (2021) tells us that we cannot improve the error rate if the feature selection probabilities are equal in each step. Therefore, we will drop this assumption. Instead, we will use a Markov model and the probabilities for the feature selection will depend on all previously selected features. A similar idea is used by Mondrian trees as introduced by Mourtada et al. (2020). For Mondrian trees, the probability of selecting a feature to split is proportional to the length of the cell in that feature. The Ehrenfest trees will also lead to higher selection probabilities for features in which the considered cell is longer.

3.3.3.1 The two-dimensional case

For clarity, we first explain the two-dimensional case where we use the Ehrenfest model, see for instance Bhattacharya and Waymire (2022, Chapter 11, Example 3). In this model, we consider two numbered containers, each initially containing B identical balls, and in each step a uniformly chosen ball from the total of $2B$ balls is moved from its current container to the other. The model has the Markov property, because its transition probabilities depend only on the current state.

We link this model to the partition construction on $[0,1]^2$. For simplicity, we consider only the splits for the cell containing a fixed point x_0. We start with an Ehrenfest model in its initial balanced state and use the first k steps to determine the splits for $A_k(x_0,\omega)$. In each step, we split feature l if the selected ball leaves container l. For the next step this decreases the probability of splitting the same feature again and increases the probability of splitting the other feature. By favoring the feature that has been split less often, this procedure results in a more balanced number of splits than the uniform selection of a feature in each step. In practice, one needs to copy the current state of the Ehrenfest model after each split. The two models will then evolve independently for the two resulting cells. An illustration of this process can be found in Figure 3.2.

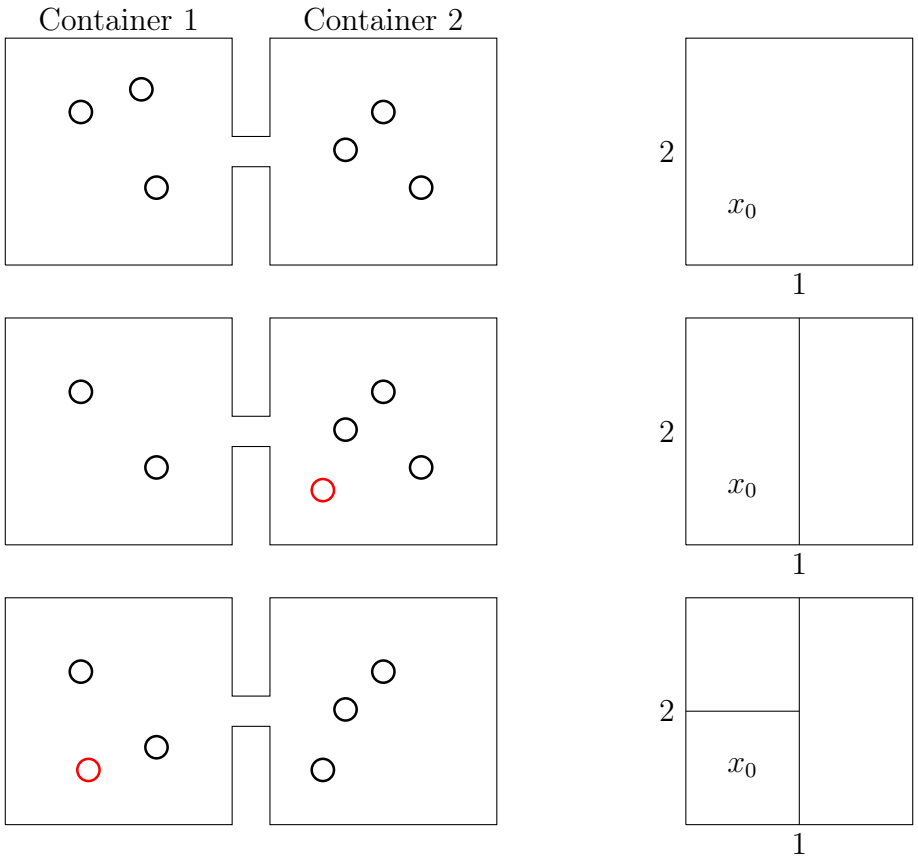

Figure 3.2: First three states of a two-dimensional Ehrenfest model with $B = 3$ and corresponding cell construction. The red ball moved in the previous step.

For now we drop the dependence of $S_{l,t}$ on x_0 and ω in the notation for convenience. Let $B_{l,t}$ denote the number of balls in container l after t construction steps. The state of the Ehrenfest model after t steps is described by $(B_{1,t}, B_{2,t})$. For $D_{t,j}$ from Definition 3.2 we drop the second index, because we only consider the construction of the cell around x_0. Thus, $D_t \in \{1, 2\}$ is the random variable that determines which coordinate is split in the t-th step. It holds that $S_{l,t} = \sum_{s=1}^{t} \mathbb{I}\{D_s = l\}$. For $l \in \{1, 2\}$, the probability of selecting a ball in container l and thus splitting feature l is

$$\mathbb{P}(D_{t+1} = l \mid B_{1,t}, B_{2,t}) = \frac{B_{l,t}}{2B}.$$

We note that it would be sufficient to condition on $B_{1,t}$ because $B_{1,t} + B_{2,t} = 2B$. In the case $p = 2$ it holds that $t = S_{1,t} + S_{2,t}$. Further, the relation between the number of balls and the number of splits is

$$2(S_{1,t} - S_{2,t}) = B_{2,t} - B_{1,t}. \tag{3.26}$$

We obtain

$$2S_{1,t} - t = S_{1,t} - S_{2,t} = \frac{B_{2,t} - B_{1,t}}{2} = \frac{2B - 2B_{1,t}}{2} = B - B_{1,t}, \tag{3.27}$$

and analogously $2S_{2,t} - t = B - B_{2,t}$. This implies

$$\mathbb{P}(D_{t+1} = l \mid B_{1,t}, B_{2,t}) = \frac{B_{l,t}}{2B} = \frac{B + t - 2S_{l,t}}{2B} = \frac{1}{2} + \frac{1}{2B}(t - 2S_{l,t}).$$

The equation describes how the probability of splitting a feature l decreases in the current number of its splits in relation to t. The number of balls B is a tuning parameter that weights the influence of this effect. For $B \to \infty$ the probability goes to $1/2$. Using (3.27) we get for the squared diameter after the final step $t = k$ that

$$\begin{aligned}
\mathfrak{d}(A_k(x_0, \omega))^2 &= 2^{-2S_{1,k}} + 2^{-2S_{2,k}} \\
&= 2^{-2S_{1,k}} + 2^{-2(k - S_{1,k})} \\
&= 2^{-k} \left(2^{-(2S_{1,k} - k)} + 2^{2S_{1,k} - k}\right) \\
&= 2^{-k} \left(2^{B_{1,k} - B} + 2^{B - B_{1,k}}\right) \\
&\leq 2^{-k} 2^{B+1},
\end{aligned} \tag{3.28}$$

which is $\mathcal{O}(2^{-k})$ if the tuning parameter B does not depend on k.

3.3.3.2 The p-dimensional case

We generalize the two-dimensional model to dimension p. The model has p numbered containers that each contain B identical balls in the initial state. Still, in each step, a uniformly chosen ball from the total of pB balls is moved from its current container to another. We will describe the target container selection later. We keep the connection to the Ehrenfest model by splitting feature $l \in [p]$ if a ball in container l is selected. The current state of the model after t steps is described by $(B_{l,t})_{l=1}^p$. We now have $D_t \in [p]$, and for $l \in [p]$ the probability of selecting a ball in container l and thus splitting feature l is

$$\mathbb{P}(D_{t+1} = l \mid (B_{l,t})_{l=1}^p) = \frac{B_{l,t}}{pB}. \tag{3.29}$$

This implies for $l \in [p]$ that

$$\mathbb{P}(S_{l,t+1} - S_{l,t} = j \mid B_{l,t}) = \begin{cases} \frac{B_{l,t}}{pB}, & j = 1, \\ 1 - \frac{B_{l,t}}{pB}, & j = 0, \\ 0, & \text{otherwise.} \end{cases}$$

This describes the selection of a ball, but unlike the case of $p = 2$, the ball has more than one container to move to. It remains to specify the random selection of this target container to fully describe the transition of the p-dimensional Ehrenfest model to the next state.

An important feature of the case $p = 2$ is the direct connection between $(S_{l,t})_{l=1}^p$ and $(B_{l,t})_{l=1}^p$ in (3.26). We lose this connection for $p > 2$ because the state of the model is not solely described by the $(S_{l,t})_{l=1}^p$ as they are missing the information about the target containers. A direct consequence of (3.26) was (3.27) which allowed the bound for the diameter in (3.28). In other words the reason for this bound is that the splitting probability for a feature is zero if it was split too many times relative to the other features.

We now explain our choice of the target probabilities which preserves this property in the p-dimensional case. We use the term "target" so that the probabilities are not confused with the transition probabilities of the entire model. These transition probabilities describe the Markovian behavior of the entire model state $(B_{l,t})_{l=1}^p$ for $t \in \mathbb{N}$ and are only implicitly defined by (3.29) and the target probabilities below.

We prohibit the addition of balls to a container if the corresponding feature has already been split too many times in relation to t/p. How many splits are too many is controlled by a tuning parameter Δ. Lemma 3.7 below shows that this will lead to a deterministic bound for $S_{l,t} - t/p$ similar to the equality in (3.27). The target probability of a ball leaving container i, for entering container $j \neq i$ is

$$\varrho((S_{l,t})_{l=1}^p, j, i) = \frac{\mathbb{I}\{pS_{j,t} < t + p\Delta\}}{\sum_{l \in \{1,\dots,p\}\setminus\{i\}} \mathbb{I}\{pS_{l,t} < t + p\Delta\}}. \tag{3.30}$$

The choice in (3.30) implies that we need to choose $\Delta > B/p$. It is necessary that $\varrho((S_{l,t})_{l=1}^p, j, i) > 0$ for at least one $j \neq i$. Consider the start of the splitting procedure, where each container has B balls. Assume that all the balls in containers 2 to p are chosen in consecutive order and move into container one. At step $t = (p-1)B$ we then have $S_{j,(p-1)B} = B$ for $j \in \{2, \dots, p\}$. If $B \geq p\Delta$ we have

$$pS_{j,t} = pS_{j,(p-1)B} = pB \geq (p-1)B + p\Delta = t + p\Delta$$

and hence $\varrho((S_{l,t})_{l=1}^p, j, i) = 0$ for all $j \neq i$. The model could not continue to operate in this case.

When we construct an Ehrenfest tree, we start with one p-variate Ehrenfest model in the first step. When a cell is split, we need two copies of the current state of the model for the two child cells. From there on, the copies of the model evolve independently.

The lemma below gives us an almost sure bound for $|S_{j,k} - \frac{k}{p}|$ that does not depend on k. This will directly lead to three corollaries with bounds for the three characteristics introduced in Section 3.3.1.

Lemma 3.7. *For the Ehrenfest tree on $[0,1]^p$ with target probabilities given in (3.30) it almost surely holds that*

$$S_{j,k} \leq \frac{k}{p} + \Delta + (p-1)B =: \frac{k}{p} + C^{(1)}_{\Delta,B}$$

$$\text{and} \qquad S_{j,k} \geq \frac{k}{p} - (p-1)\Delta - (p-1)^2 B =: \frac{k}{p} - C^{(2)}_{\Delta,B}.$$

Proof. Without loss of generality we consider the splits in direction 1. From (3.30) we get an upper bound for $S_{1,t}$. We consider the step in which the bound in (3.30) is exceeded. More precisely let t be such that $pS_{1,t} \geq t + p\Delta$ and $pS_{1,t-1} < t - 1 + p\Delta$. This still yields the upper bound

$$pS_{1,t} = p(S_{1,t-1} + 1) < t - 1 + p\Delta + p = t + p\Delta + (p-1).$$

To exceed the threshold in time step t one ball had to leave container 1. Thus there can be at most $pB - 1$ balls left in the container. For any s the fraction $S_{1,s}/s$ will be the largest if all the balls leave the container consecutively. We get

$$pS_{1,t+pB-1} = p(S_{1,t} + pB - 1) = pS_{1,t} + p^2B - p \leq t + p\Delta + (p-1) + p^2B - p = t + p\Delta + p^2B - 1.$$

For $s = t + pB - 1$ this is

$$pS_{1,s} \leq s + p\Delta - pB + p^2 B = s + p\Delta + (p^2 - p)B.$$

Since t and s were arbitrary and we considered the worst case we get for any t that

$$S_{1,t} \leq \frac{t}{p} + \Delta + (p-1)B$$

which proves the first claim. Using this we get for any t that

$$\begin{aligned} S_{1,t} &= t - \sum_{l=2}^{p} S_{l,t} \\ &\geq t - (p-1)\left(\frac{t}{p} + \Delta + (p-1)B\right) \\ &= \frac{t}{p} - (p-1)\Delta + (p-1)^2 B. \end{aligned}$$

□

Remark 3.8. The only thing that the proof of Lemma 3.7 uses about the probabilities $\varrho((S_{l,t})_{l=1}^p, j, i)$, is that they are zero if $pS_{j,t} \geq t + p\Delta$. Hence Lemma 3.7 holds for any adjusted probabilities

$$\varrho((S_{l,t})_{l=1}^p, j, i) = w((S_{l,t})_{l=1}^p, j, i)\mathbb{I}\{pS_{j,t} < t + p\Delta\}$$

with weights $w((S_{l,t})_{l=1}^p, j, i) > 0$ that satisfy

$$\sum_{j \in \{1,\dots,p\}\setminus\{i\}} w((S_{l,t})_{l=1}^p, j, i)\mathbb{I}\{pS_{j,t} < t + p\Delta\} = 1.$$

This allows for some flexibility in the distribution of the splits. How different choices of the $w((S_{l,t})_{l=1}^p, j, i)$ and also of Δ and B affect the performance remains to be analyzed.

The corollaries below use the above lemma to bound $\mathfrak{d}(A_k(x_0, \omega))$, $\mathcal{V}_{\cap,k}$ and $N_f(k)$.

Corollary 3.9. *For the Ehrenfest tree on $[0,1]^p$ with target probabilities given in (3.30) it almost surely holds that*

$$\mathfrak{d}(A_k(x_0, \omega)) \leq \sqrt{p} 2^{C^{(1)}_{\Delta,B}} 2^{-k/p} = \mathcal{O}(2^{-k/p}).$$

Proof. Lemma 3.7 directly implies an almost sure bound for the diameter. We obtain

$$\mathfrak{d}(A_k(x_0, \omega))^2 = \sum_{l=1}^{p} 2^{-2S_l(x_0,\omega)} = 2^{-2k/p} \sum_{l=1}^{p} 2^{2(k/p - S_l(x_0,\omega))} \leq 2^{-2k/p} p 2^{2C^{(1)}_{\Delta,B}}$$

and taking the square root this yields the claim. The resulting term is $\mathcal{O}(2^{-k/p})$ because $C^{(1)}_{\Delta,B}$ does not depend on k. □

Corollary 3.10. *For the Ehrenfest trees on $[0,1]^p$ with target probabilities given in (3.30) it holds that*

$$\mathcal{V}_{\cap,k} \geq 2^{-k} 2^{-p(C^{(1)}_{\Delta,B} + C^{(2)}_{\Delta,B})/2} \qquad \text{and} \qquad \mathcal{V}_{\cap,k} = \Theta(2^{-k}).$$

Proof. Equation (3.21) gives us

$$\mathcal{V}_{\cap,k} = \mathbb{E}\left[2^{-\sum_{i=1}^{p} \max\{S_l(x_0,\omega_1),S_l(x_0,\omega_2)\}}\right] = 2^{-k}\mathbb{E}\left[2^{-\frac{1}{2}\sum_{l=1}^{p} |S_l(x_0,\omega_1)-S_l(x_0,\omega_2)|}\right].$$

Lemma 3.7 yields for any ω that

$$\frac{k}{p} - C^{(1)}_{\Delta,B} \leq S_l(x_0,\omega) \leq \frac{k}{p} + C^{(2)}_{\Delta,B}.$$

Hence

$$|S_l(x_0,\omega_1) - S_l(x_0,\omega_2)| \leq C^{(1)}_{\Delta,B} + C^{(2)}_{\Delta,B}$$

which leads to

$$\mathcal{V}_{\cap,k} = 2^{-k}\mathbb{E}\left[2^{-\frac{1}{2}\sum_{l=1}^{p} |S_l(x_0,\omega_1)-S_l(x_0,\omega_2)|}\right] \geq 2^{-k}2^{-p(C^{(1)}_{\Delta,B}+C^{(2)}_{\Delta,B})/2}.$$

This implies $2^{-k} = \mathcal{O}(\mathcal{V}_{\cap,k})$ because the constants do not depend on k and with the bound $\mathcal{V}_{\cap,k} \leq 2^{-k}$ from (3.20) it holds that $\mathcal{V}_{\cap,k} = \Theta(2^{-k})$. □

Corollary 3.11. *For the Ehrenfest tree on $[0,1]^p$ with target probabilities given in (3.30) it holds that $N_f(k) = \mathcal{O}(2^k)$.*

Proof. Lemma 3.7 implies

$$S_{j,k} \leq \frac{k}{p} + C^{(2)}_{\Delta,B}$$

for all l and k. This implies that the size of the undividable sets is larger than or equal to

$$2^{-p(\frac{k}{p}+C^{(2)}_{\Delta,B})} = 2^{-(k+pC^{(2)}_{\Delta,B})}.$$

This implies for the number of the cells that

$$N_f(k) \leq 2^{(k+pC^{(2)}_{\Delta,B})} = \mathcal{O}(2^k).$$ □

The above results will be used in the proofs of Chapter 4 and Chapter 5.

3.4 Kernel random forests

Kernel random forests (KeRF) introduced by Scornet (2016b) are regression estimators similar to random forests but they utilize a different structure more similar to kernel regression estimators such as the Nadaraya-Watson estimator (Nadaraya (1964), Watson (1964)). Like random forests, they are based on an ensemble of tree-partitions of the feature space. The primary distinction between these two estimators arises from the method by which the observations are weighted within the estimation. KeRF can be employed for the same variants of partition creation as random forests, including both data dependent and purely random methods. The KeRF estimator corresponding to the random forest in (3.2) is

$$\hat{m}^{K}_{N,n}(x_0,(\theta_j)_{j=1}^{N},\mathcal{D}_n) := \frac{\sum_{i=1}^{N}\sum_{j\in I_{\theta_i}} Y_j \mathbb{I}\{X_j \in A_n(x_0,\theta_i,\mathcal{D}_n)\}}{\sum_{i=1}^{N}\sum_{j\in I_{\theta_i}} \mathbb{I}\{X_j \in A_n(x_0,\theta_i,\mathcal{D}_n)\}}.$$

Since we only focus on centered purely random forests in a U-statistic form we will do the same for the KeRF. For the sake of simplicity, we will limit its consideration to the complete version, which incorporates all trees. We denote

$$\begin{aligned} U^{(\mathrm{KRF})}_{n,r_n,\omega}(x_0) &:= \frac{\sum_{I\in B_{r_n,n}} \sum_{j\in I} Y_j \mathbb{I}\{X_j \in A_k(x_0,\omega_I)\}}{\sum_{I\in B_{r_n,n}} \sum_{j\in I} \mathbb{I}\{X_j \in A_k(x_0,\omega_I)\}} \\ &= \frac{\sum_{j=1}^n Y_j \sum_{I\in B_{r_n,n}: j\in I} \mathbb{I}\{X_j \in A_k(x_0,\omega_I)\}}{\sum_{j=1}^n \sum_{I\in B_{r_n,n}: j\in I} \mathbb{I}\{X_j \in A_k(x_0,\omega_I)\}}. \end{aligned} \tag{3.31}$$

It is important to note that this estimator is not a U-statistic, although its numerator and denominator are both U-statistics. The second representation implies that the estimator is a weighted average of the Y_j with weights

$$\frac{\sum_{I\in B_{r_n,n}: j\in I} \mathbb{I}\{X_j \in A_k(x_0,\omega_I)\}}{\sum_{I\in B_{r_n,n}} \sum_{i\in I} \mathbb{I}\{X_i \in A_k(x_0,\omega_I)\}} = \frac{\sum_{I\in B_{r_n,n}: j\in I} \mathbb{I}\{X_j \in A_k(x_0,\omega_I)\}}{\sum_{i=1}^n \sum_{I\in B_{r_n,n}: i\in I} \mathbb{I}\{X_i \in A_k(x_0,\omega_I)\}}. \tag{3.32}$$

The analogous weight in the standard random forest is

$$\frac{r_n}{n} \frac{1}{\binom{n-1}{r_n-1}} \sum_{I\in B_{r_n,n}: j\in I} \frac{\mathbb{I}\{X_j \in A_k(x_0,\omega_I)\}}{\sum_{i\in I} \mathbb{I}\{X_i \in A_k(x_0,\omega_I)\}}, \tag{3.33}$$

since

$$U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0) = \sum_{j=1}^n Y_j \frac{r_n}{n} \frac{1}{\binom{n-1}{r_n-1}} \sum_{I\in B_{r_n,n}: j\in I} \frac{\mathbb{I}\{X_j \in A_k(x_0,\omega_I)\}}{\sum_{i\in I} \mathbb{I}\{X_i \in A_k(x_0,\omega_I)\}}.$$

Their weights are the main difference between the KeRF and the standard random forest. In the standard RF the trees are all weighted equally in the estimation of $m(x_0)$. Thus, the weights of the Y_j in (3.33) itself are an equally weighted average in the trees.

The KeRF is not an average of regression tree estimators. Nonetheless it depends on the same collection of partitions. The weight of Y_j from (3.32) is not weighted equally in the partitions based on the ω_I. Instead, the partition based on ω_I is weighted more heavily if its cell around x_0 contains relatively many observations.

Thus, an advantage of the KeRF weights is that the estimator is less sensitive to cells with few observations, and that cells containing no observations do not affect the estimator. However, this structure can also lead to disadvantages in the estimation. Consider a fixed $x_0 \in [0,1]^p$ and suppose that the X_j are not uniformly distributed. The fact that the estimator weights partitions more heavily if they contain many observations means that on average cells around x_0 that contain regions of relatively higher density are relatively more important in the estimation. If the density is larger in the regions of the feature space that are more distant to x_0, the weight of cells with larger diameters is increased. For a general m, this can lead to a worse approximation error.

A natural question is how much the RF and KeRF estimators differ. Scornet (2016b, Proposition 3) provides a general upper bound on

$$\Big| \frac{\hat{m}_{N,n}(x_0, (\theta_j)_{j=1}^N, \mathcal{D}_n)}{\hat{m}^K_{N,n}(x_0, (\theta_j)_{j=1}^N, \mathcal{D}_n)} - 1 \Big|, \tag{3.34}$$

and analyzes the upper bound for different partition schemes. For centered random forests with partition depth k, the proposition implies that (3.34) converges to zero for the right choice of k, for instance $k = \Theta((\log_2 n)/3)$, if the distribution of X is uniform. The bound can be extended to the case where the density f_X of X satisfies $0 < c_X \leq f_X \leq C_X$, but in this case the upper bound does not converge to zero. This means that the uniform assumption is necessary to prove asymptotic equivalence with this bound, but it may still be possible to prove it with a different technique. In general, this suits the considerations from above about the possible differences in the weights for a non constant density. KeRF will reappear in Chapter 5, where we will prove confidence bands for the estimator from (3.31).

While the focus of Chapter 4 is on CPRFs, the proof techniques should allow for CLTs for KeRFs. One would have to apply Theorem 2.12 to the numerator of the estimator and handle the denominator with Slutsky's theorem. The U-statistic in the numerator has an even simpler structure than the standard random forest U-statistic.

Chapter 4

A central limit theorem for centered purely random forests

In this chapter we will apply the results for generalized incomplete U-statistics, especially Theorem 2.13, to the U-statistic version of the CPRF from Definition 3.2. Before stating the results for the asymptotic normal distribution, we include a bound for the mean squared error of a CPRF and compare the rates of the two variants from Section 3.3. Following the results, we explain the proof strategy in Section 4.3 and give the detailed proofs of all results in this chapter in Section 4.4.

From here on we assume that the training sample $\mathcal{D}_n = (X_i, Y_i)_{i=1}^n$ is from the regression model (2.1) under a few additional assumptions. The first assumption is that the density f_X of X satisfies

$$c_X \leq f_X(x) \leq C_X \tag{4.1}$$

for all $x \in [0,1]^p$ with constants $0 < c_X < C_X < \infty$. The second assumption is that m is Hölder continuous of order $\alpha \in (0,1]$ with Hölder constant C_H. Further, we require the second moments of ε to exist and denote $\sigma^2 := \mathrm{Var}(\varepsilon)$.

4.1 The mean squared error

For the mean squared prediction error of the complete U-statistic version of the CPRF we get the following proposition.

Proposition 4.1. *Consider a CPRF from Definition 3.2 and suppose* $0 < c_X \leq f_X \leq C_X$, $\mathbb{E}[\varepsilon_1^2] < \infty$, $r_n < n$ *and* m *is Hölder continuous of order* $\alpha \in (0,1]$. *It then holds that*

$$\mathbb{E}\left[\left(U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0) - m(x_0)\right)^2\right] = \mathcal{O}\left(\mathbb{E}\left[\mathfrak{d}(A_k(x_0,\omega))^\alpha\right]^2\right) + \mathcal{O}((1 - c_X 2^{-k})^{r_n}) + \mathcal{O}\Big(\frac{2^{2k}\mathcal{V}_{\cap,k}}{n}\Big) + \mathcal{O}\Big(\frac{2^k r_n}{n^2}\Big).$$

The first term in the bound corresponds to the approximation error. We already mentioned that it is controlled by the expected diameter of the cells. The second term is necessary to bound the part of the approximation error that is caused by empty cells that contain no observations. The final two terms are associated with the stochastic

error. To elaborate, the first term corresponds to the first order contributions from the Hoeffding-decomposition of the stochastic error, while the subsequent term corresponds to the remaining components of the Hoeffding-decomposition.

We want to compare the MSE with that of a single regression tree. For an appropriate comparison we consider a regression tree estimator for n observations. For a single regression tree, a purely random partition is less practical than for a random forest. Nonetheless, we consider a centered purely random tree since this is in line with the random forest estimator. Similar to (3.1) we consider

$$\tilde{m}_n(x_0, \omega, \mathcal{D}_n) = \sum_{j=1}^{n} Y_j \frac{\mathbb{I}\{X_j \in A_k(x_0, \omega)\}}{\sum_{i=1}^{n} \mathbb{I}\{X_i \in A_k(x_0, \omega)\}}.$$

For the regression tree, the first term in the bound of the MSE in Proposition 4.1 would be $\mathbb{E}\left[\mathfrak{d}(A_k(x_0, \omega))^{2\alpha}\right]$, which is larger due to Jensen's inequality. The underlying cause of this disparity is that the average diameter of numerous trees is close to the expected diameter. This illustrates why a random forest has a smaller approximation error than a regression tree. The assumption $r_n < n$ is necessary to ensure that the estimator uses more than one subsample. In the second term, the exponent r_n would be replaced by n.

Lemma 2.15 yields that the stochastic error of the regression tree fulfills

$$\begin{aligned}
&\operatorname{Var}\left(\frac{\sum_{j=1}^{n} \varepsilon_j \mathbb{I}\{X_j \in A_k(x_0, \omega)\}}{\sum_{i=1}^{n} \mathbb{I}\{X_i \in A_k(x_0, \omega)\}}\right) \\
&\quad = n\sigma^2 \mathbb{E}\left[\left(\frac{\mathbb{I}\{X_1 \in A_k(x_0, \omega)\}}{1 + \sum_{i=2}^{n} \mathbb{I}\{X_i \in A_k(x_0, \omega)\}}\right)^2\right] \\
&\quad \leq C_X n\sigma^2 2^{-k} \mathbb{E}\left[\left(\frac{1}{1 + \sum_{i=2}^{n} \mathbb{I}\{X_i \in A_k(x_0, \omega)\}}\right)^2\right] \\
&\quad \leq C_X n\sigma^2 2^{-k} 2c_X^{-2} \frac{2^{2k}}{n^2} \\
&\quad = \mathcal{O}\left(\frac{2^k}{n}\right).
\end{aligned}$$

Thus, the rate of the stochastic error of the random forest is smaller, if $2^k \mathcal{V}_{\cap,k} = o(1)$ and $r_n/n = o(1)$. The reason for the smaller stochastic error is, that the random forest utilizes more observations for the estimation at a fixed x_0. On average, the tree uses $n\mathbb{P}(X_1 \in A_k(x_0, \omega))$ observations for the estimation of $m(x_0)$, which is the expected number of observations in $A_k(x_0, \omega)$. The random forest uses more observations, because all observations with $X_i \in \bigcup_{I \in B_{r_n,n}} A_k(x_0, \omega_I)$ affect the estimator at x_0. On average this are more than $n\mathbb{P}(X_1 \in A_k(x_0, \omega))$ and the effect is captured by the term $2^k \mathcal{V}_{\cap,k} \leq 1$. The second term of the stochastic error corresponding to the higher order terms in the Hoeffding-decomposition lacks a similar interpretation.

In total, the MSE rate of a CPRF is better than that of a comparable regression tree. The gain in the approximation error depends on the behavior of the diameter. For the stochastic error, we can conclude that the gain is at most polynomial in k for both CPRF

types from Section 3.3. This is due to the fact that $2^k \mathcal{V}_{\cap,k} \lesssim k^{-(p-1)/2}$, as demonstrated in (3.25). Overall, the ensemble of trees in the random forest has two beneficial effects on the mean squared error. It allows more observations to be used to estimate $m(x_0)$ than in a single regression tree, and it also reduces the variance of the average diameter of the cells around x_0.

The corollary below shows the MSE rates for the Ehrenfest and uniform CPRF. In particular, the appropriate choice of k leads to the minimax optimal rate, see e.g. Tsybakov (2009), for Ehrenfest CPRF. With the (approximately) best choice for k, the uniform CPRF does not achieve this rate.

Corollary 4.2. *Suppose $0 < c_X \leq f_X \leq C_X$, $\mathbb{E}[\varepsilon_1^2] < \infty$, $r_n < n$ and that m is Hölder continuous of order $\alpha \in (0, 1]$. For the Ehrenfest CPRF it holds that*

$$\mathbb{E}\big[(U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0) - m(x_0))^2\big] = \mathcal{O}(2^{-2\alpha k/p}) + \mathcal{O}\big(\frac{2^k}{n}\big) + \mathcal{O}((1 - c_X 2^{-k})^{r_n}).$$

If $k = \lfloor \log_2(c_k n^{\frac{1}{(1+2\alpha/p)}}) \rfloor$ for a constant c_k and $r_n = \mathcal{O}(n)$ satisfies $n^{\frac{1}{(1+2\alpha/p)}} \log n = o(r_n)$, it holds that

$$\mathbb{E}\big[(U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0) - m(x_0))^2\big] = \mathcal{O}(n^{-\frac{2\alpha}{2\alpha+p}}).$$

For $b(p, \alpha) = (p - 1 + 2^{-\alpha})/p$ the uniform CPRF satisfies

$$\begin{aligned}\mathbb{E}\left[\big(U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0) - m(x_0)\big)^2\right] &= \mathcal{O}\left(b(p,\alpha)^{2k}\right) + \mathcal{O}((1 - c_X 2^{-k})^{r_n}) \\ &\quad + \mathcal{O}\big(\frac{2^k k^{-(p-1)/2}}{n}\big) + \mathcal{O}\big(\frac{2^k r_n}{n^2}\big).\end{aligned}$$

For a constant c_k and $\nu = \frac{2\log_2 b(p,\alpha)}{2\log_2 b(p,\alpha)-1}$ choose $k = \lfloor \log_2(c_k(n(\log_2 n)^{(p-1)/2})^{1-\nu}) \rfloor$. If $r_n = \Theta(nk^{-(p-1)/2})$, it holds that

$$\mathbb{E}\left[\big(U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0) - m(x_0)\big)^2\right] = \mathcal{O}\left(\big(n(\log_2 n)^{(p-1)/2}\big)^{-\nu}\right).$$

We note that $b(p, \alpha) < 1$ implies that $\nu > 0$. The bound $b(p, \alpha) \geq 2^{-\alpha/p}$ from (3.24) further implies

$$\nu = \frac{2\log_2 b(p,\alpha)}{2\log_2 b(p,\alpha) - 1} \leq \frac{-2\alpha/p}{-2\alpha/p - 1} = \frac{2\alpha}{2\alpha + p}, \tag{4.2}$$

because $g(\xi) = \xi/(\xi - 1)$ is monotonically decreasing for $\xi \in (-\infty, 1)$. Therefore, in the case of $p > 1$, the rate of the uniform CPRF is worse than the rate of the Ehrenfest CPRF. In particular, it is not minimax optimal. If m is Lipschitz continuous the rate of the uniform CPRF matches that of Klusowski (2021, Theorem 2), which cannot be improved.

Mourtada et al. (2020) proved that Mondrian trees and forests, which are purely random, reach the minimax optimal rate over the class of α-Hölder functions for $\alpha \in (0, 1]$. To the best of our knowledge, their result is the first to do so for any purely random forest. Although the convergence rate is the same for single trees and forests, they discuss that forests perform better in practice. They attribute this disparity to the smaller

approximation error of forests compared to single trees. They further prove that Mondrian forests benefit from additional smoothness of the regression function, while single Mondrian trees do not, which explains the disparity in some cases. The approximation error rate for Ehrenfest trees and forests is also the same due to the almost sure bound on the diameter. However, it seems possible that a similar result for smoother regression functions may hold for the Ehrenfest CPRF. A notable difference from our work is that Mourtada et al. (2020) do not incorporate subsampling for the trees, but instead use the entire training sample for the estimation performed by each tree.

To reach its optimal rate, the Ehrenfest CPRF only needs that $n^{\frac{1}{(1+2\alpha/p)}} \log n = o(r_n)$, which ensures $2^k \log n = o(r_n)$ but allows $r_n = \Theta(n)$. In order to optimize the rate of the uniform CPRF, it is necessary that r_n/n converges to zero at a logarithmic rate. For $r_n = \Theta(n)$, the rate of the term corresponding to the higher order terms from the Hoeffding-decomposition would be larger. This finding indicates that the construction of confidence intervals for the uniform CPRF based on these specific bounds is not possible when $r_n = \Theta(n)$. A smaller choice of r_n is eligible, provided that the term corresponding to the nonempty cells is sufficiently small.

4.2 The central limit theorem

In this section we will finally state the central limit theorem, but first we need several definitions. Let us denote

$$p_{x_0}(\omega) := \mathbb{P}(X_1 \in A_k(x_0, \omega) \mid \omega) \qquad \text{and} \tag{4.3}$$

$$p_{x_0} := \mathbb{E}[p_{x_0}(\omega)] = \mathbb{P}(X_1 \in A_k(x_0, \omega)). \tag{4.4}$$

In general, these terms are functions in x_0 as soon as the density is not constant. However, they are strongly connected with the density and

$$p_{x_0}(\omega) = \mathbb{P}(X_1 \in A_k(x_0, \omega) \mid \omega) = \int_{A_k(x_0,\omega)} f_X(x)dx. \tag{4.5}$$

By construction of the centered trees, it holds that

$$\int_{A_k(x_0,\omega)} dx = 2^{-k}$$

and thus (4.1) implies

$$c_X 2^{-k} \leq p_{x_0}(\omega) \leq C_X 2^{-k} \qquad \text{and} \tag{4.6}$$

$$c_X 2^{-k} \leq p_{x_0} \leq C_X 2^{-k}. \tag{4.7}$$

Let us define

$$K_k(x_0, x) = \mathbb{E}\left[\mathbb{I}\{x \in A_k(x_0, \omega)\} p_{x_0}(\omega)^{-1}\right], \tag{4.8}$$

which can be seen as a kernel function. We denote its second moment by

$$\Psi_k(x_0) := \mathbb{E}\big[K_k(x_0, X_1)^2\big]. \tag{4.9}$$

Remark 4.3. The term

$$\mathbb{P}(x \in A_k(x_0, \omega))$$

is related to the above kernel. If X_1 is uniformly distributed, it is equal to $K_k(x_0, x)2^{-k}$ and otherwise it is $\Theta(K_k(x_0, x)2^{-k})$. Scornet (2016b) call this function the connection function of the infinite random forest. They calculate the function explicitly for a centered RF with uniform distribution of the split directions. The formula can be found in Proposition 5 of their article, which also includes a graphical representation.

Theorem 4.4. *Consider a CPRF from Definition 3.2 and suppose* $0 < c_X \leq f_X \leq C_X$, $\mathbb{E}[\varepsilon_1^6] < \infty$ *and* m *is Hölder continuous of order* $\alpha \in (0, 1]$. *Under the conditions*

$$\frac{r_n}{n2^k \mathcal{V}_{\cap,k}} \to 0, \tag{4.10}$$

$$\frac{n}{N2^{2k} \mathcal{V}_{\cap,k}} \to 0, \tag{4.11}$$

$$\frac{2^k}{r_n} \log\left(n2^{-2k} \mathcal{V}_{\cap,k}^{-1}\right) \to 0, \tag{4.12}$$

$$\mathbb{E}\left[\mathfrak{d}(A_k(x_0, \omega))^\alpha\right]^2 \frac{n}{2^{2k} \mathcal{V}_{\cap,k}} \to 0, \tag{4.13}$$

it holds that

$$\sqrt{\frac{n}{\Psi_k(x_0)}} \left(U^{(\mathrm{RF})}_{n,r_n,N,\omega}(x_0) - m(x_0)\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2). \tag{4.14}$$

Assumption (4.11) ensures that the number of trees is sufficiently large to reduce the variance to that of the complete U-statistic case. As previously mentioned in Remark 3.4, the cell size in a CPRF is 2^{-k}, and $2^{-k/p}$ is comparable to the bandwidth for kernel regression estimators. The condition in (4.12) is necessary to prevent cells in the tree partitions that contain no observation. The last assumption in (4.13) is an undersmoothing assumption. An assumption of this type is standard in the theory and necessary for the dominance of the stochastic error over the approximation error. It implies that the bandwidth parameter $2^{-k/p}$ needs to be slightly smaller than the choice that optimizes the means squared error. Condition (4.10) ensures that the first order terms in the Hoeffding-decomposition of the stochastic error are asymptotically leading. It is derived from the condition $\zeta^n_{r_n}/(n\zeta^n_{1,\omega}) \to 0$ from Theorem 2.13.

Remark 4.5. According to Peng et al. (2022) the condition $\zeta^n_{r_n}/(n\zeta^n_{1,\omega}) \to 0$ from Theorem 2.13 can be weakened to

$$\frac{r_n}{n}\left(\frac{\zeta^n_{r_n}}{r_n \zeta^n_{1,\omega}} - 1\right) \to 0.$$

This suggests that the assumption $r_n/n \to 0$ might be circumvented. In order to achieve this objective, it is necessary that $\zeta^n_{r_n}(r_n\zeta^n_{1,\omega})^{-1} \to 1$. For the Ehrenfest CPRF the auxiliary results in this chapter allow to prove that $\zeta^n_{r_n}(r_n\zeta^n_{1,\omega})^{-1} = \mathcal{O}(1)$. Nevertheless the precision of our results is insufficient to prove the convergence to one due to the lack of sharpness in several bounds. With $\mathcal{V}_{\cap,k} \leq 2^{-k}$ (4.10) implies that

$$\frac{r_n}{n} \leq \frac{r_n}{n2^k \mathcal{V}_{\cap,k}} \to 0$$

is a necessary condition.

If we want to apply Theorem 4.4, for example to get confidence intervals we need to calculate the variance term $\Psi_k(x_0)$. Equation (4.20) below yields that $\Psi_k(x_0) = \Theta(2^{2k}\mathcal{V}_{\cap,k})$ for every x_0. However, it is not always possible to get a closed form solution of this term, even if the distribution of X is known. It is determined by the distribution of ω and X. If the distribution of X is known, we can get an arbitrarily close Monte Carlo approximation of the term by simulating enough copies of X and ω. If the distribution of X is not known one could either estimate the density of X first, or one could use bootstrap samples to estimate $\Psi_k(x_0)$.

For the Ehrenfest CPRF and the uniform CPRF we get the corollaries below. Their proofs are postponed to Section 4.4.1, following the proof of Theorem 4.4.

Corollary 4.6. *Consider the Ehrenfest CPRF and suppose* $0 < c_X \leq f_X \leq C_X$, $\mathbb{E}[\varepsilon_1^6] < \infty$ *and* m *is Hölder continuous of order* $\alpha \in (0, 1]$. *Under the assumptions*

$$\begin{aligned} \frac{r_n}{n} &\to 0, \\ \frac{n}{N2^k} &\to 0, \\ \frac{2^k}{r_n} \log\left(n2^{-k}\right) &\to 0, \\ \text{and} \qquad 2^{-2\alpha k/p}\frac{n}{2^k} &\to 0, \end{aligned}$$

the convergence in distribution in (4.14) holds. The tuning parameters $r_n = \lfloor n(\log n)^{-1} \rfloor$ *and* $k = \lfloor \log_2((\log n) n^{\frac{p}{p+2\alpha}}) \rfloor$ *fulfill the assumptions above and imply that the asymptotic variance in (4.14) satisfies*

$$\Psi_k(x_0)/n = \Theta\left(n^{-\frac{2\alpha}{2\alpha+p}}(\log n)\right).$$

It is noteworthy that the above choice for 2^k exceeds the value presented in Corollary 4.2 by a logarithmic term. This is a required consequence of the undersmoothing assumption (4.13). The asymptotic variance rate is larger than the optimal MSE rate by a factor of $\log n$ for the same reason. For the optimal MSE rate, it was possible to choose $r_n = \Theta(n)$, which is not eligible here due to the first assumption regarding the terms in the Hoeffding-decomposition. The corollary below gives us the asymptotic distribution of the uniform CPRF.

Corollary 4.7. *Consider the uniform CPRF and suppose* $0 < c_X \leq f_X \leq C_X$, $\mathbb{E}[\varepsilon_1^6] < \infty$ *and* m *is Hölder continuous of order* $\alpha \in (0, 1]$. *Under the assumptions*

$$\begin{aligned} \frac{r_n k^{(p-1)}}{n} &\to 0, \\ \frac{nk^{(p-1)}}{N2^k} &\to 0, \\ \frac{2^k}{r_n} \log\left(n2^{-k}k^{(p-1)}\right) &\to 0, \\ \text{and} \qquad \left(\frac{p-1+2^{-\alpha}}{p}\right)^{2k} \frac{nk^{(p-1)}}{2^k} &\to 0, \end{aligned}$$

the convergence in distribution in (4.14) holds. For ν as in Corollary 4.2 the tuning parameters $k = \lfloor \log_2(c_k(n(\log_2 n)^p)^{1-\nu}) \rfloor$ and $r_n = \lfloor n(\log_2 n)^{-p} \rfloor$ fulfill the assumptions above and imply that the asymptotic variance in (4.14) satisfies

$$\Psi_k(x_0)/n = \mathcal{O}\left(n^{-\nu}(\log_2 n)^{p(1/2-\nu)+1/2}\right).$$

Once more, k is larger than in Corollary 4.2. We recall that $0 < \nu \leq \frac{2\alpha}{2\alpha+p}$, see (4.2), which yields

$$p(1/2-\nu)+1/2 \geq p\left(1/2 - \frac{2\alpha}{2\alpha+p}\right) + 1/2 \geq p\left(1/2 - \frac{2}{2+p}\right) + 1/2 > 0$$

for any $p \in \mathbb{N}$. Therefore, the logarithmic term in the asymptotic variance rate has a positive exponent. This implies that the rate is larger than the MSE rate in Corollary 4.2. Further, r_n is smaller by $(\log_2 n)^{-(p+1)/2}$ due to the first assumption and the gap between the lower and upper bound for $\mathcal{V}_{\cap,k}$ in (3.25).

The upper bound for ν illustrates that the asymptotic variance rate for the uniform CPRF is larger than the one for the Ehrenfest CPRF in Corollary 4.6. The primary factor contributing to this discrepancy is the greater approximation error associated with the fourth assumption. As is the case with the MSE, the assumptions on r_n are more restrictive for the uniform CPRF. This is caused by the disparity in the rate of $\mathcal{V}_{\cap,k}$ in comparison to the Ehrenfest case.

4.3 Proof strategy

To prove the central limit theorem, we use the error decomposition of the U-statistic estimator in (3.13) and treat the two error terms separately. We will prove that the approximation error converges to zero in probability at a sufficiently fast rate. Similar bounds are also used when proving consistency of a random forest, see for example Klusowski (2021, Theorem 1).

For the stochastic error we will apply Theorem 2.13. To use the result we need to calculate and bound the terms

$$\zeta^n_{1,\omega}(x_0) = \operatorname{Var}\left(\mathbb{E}[h^{(\varepsilon)}_n(Z_1, \ldots, Z_{r_n}, \omega) \mid Z_1]\right) \tag{4.15}$$

$$\text{and} \qquad \zeta^n_{r_n}(x_0) = \operatorname{Var}\left(h^{(\varepsilon)}_n(x_0, Z_1, \ldots, Z_{r_n}, \omega)\right). \tag{4.16}$$

We note that $\zeta^n_{1,\omega}(x_0)$ is the variance of $h^{(\varepsilon)}_{n,1}$ from the Hájek projection, see (2.23). From here on these terms will correspond to the U-statistic $U^{(\varepsilon)}_{n,r_n,N,\omega}$ even though this is not denoted through their notation. We point out that they depend on the fixed point x_0 at which we consider the estimator. The application of Theorem 2.13 leads to the first statement in the theorem below. The second statement follows by the connection of $\zeta^n_{1,\omega}$ and $\zeta^n_{r_n}$ to Ψ_k and $\mathcal{V}_{\cap,k}$ if N is large enough.

Theorem 4.8. *For a centered purely random forest suppose that*

$$\frac{r_n}{n2^k\mathcal{V}_{\cap,k}} \to 0, \tag{4.17}$$

further assume that $\mathbb{E}[\varepsilon_1^6] < \infty$ *and that* $2^k \leq cr_n$ *for some sufficiently small* $c > 0$*. For the incomplete U-statistic* $U^{(\varepsilon)}_{n,r_n,N,\omega}(x_0)$ *with* $\zeta^n_{1,\omega}(x_0)$ *from (4.15) and* $\zeta^n_{r_n}(x_0)$ *from (4.16) it holds that*

$$\frac{U^{(\varepsilon)}_{n,r_n,N,\omega}(x_0)}{\sqrt{r_n^2\zeta^n_{1,\omega}(x_0)/n + \zeta^n_{r_n}(x_0)/N}} \xrightarrow{d} \mathcal{N}(0,1).$$

If the additional conditions $2^k = o(r_n)$ *and*

$$\frac{n}{Nr_n 2^k \mathcal{V}_{\cap,k}} \to 0 \tag{4.18}$$

are met the convergence is reduced to

$$\frac{U^{(\varepsilon)}_{n,r_n,N,\omega}(x_0)}{\sqrt{\Psi_k(x_0)/n}} \xrightarrow{d} \mathcal{N}(0,\sigma^2).$$

In the calculations and bounds for $\zeta^n_{1,\omega}$ and $\zeta^n_{r_n}$, the terms Ψ_k and $\mathcal{V}_{\cap,k}$ will appear. We briefly explain the connection of the latter two. From (3.17) we have

$$\mathcal{V}_{\cap,k} = \mathbb{E}\left[\int_{A_k(x_0,\omega_1)\cap A_k(x_0,\omega_2)} dx\right],$$

which is the expected volume of $A_k(x_0,\omega_1) \cap A_k(x_0,\omega_2)$. It holds that

$$\begin{aligned}\mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega_1)\cap A_k(x_0,\omega_2)\}\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega_1)\cap A_k(x_0,\omega_2)\} \mid \omega_1,\omega_2\right]\right] \\ &= \mathbb{E}\left[\int_{A_k(x_0,\omega_1)\cap A_k(x_0,\omega_2)} f_X(x)dx\right].\end{aligned}$$

Using the assumptions on the density from (4.1) we have

$$c_X\mathcal{V}_{\cap,k} \leq \mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega_1)\cap A_k(x_0,\omega_2)\}\right] \leq C_X\mathcal{V}_{\cap,k}. \tag{4.19}$$

The bound on $p_{x_0}(\omega)$ in (4.6) yields

$$\begin{aligned}\Psi_k(x_0) &= \mathbb{E}\left[K_k(x_0,X_1)^2\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega_1)\}\frac{1}{p_{x_0}(\omega_1)} \mid X_1\right]\mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega_2)\}\frac{1}{p_{x_0}(\omega_2)} \mid X_1\right]\right] \\ &\leq c_X^{-2}2^{2k}\mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega_1)\} \mid X_1\right]\mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega_2)\} \mid X_1\right]\right] \\ &= c_X^{-2}2^{2k}\mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega_1)\cap A_k(x_0,\omega_2)\} \mid X_1\right]\right] \\ &= c_X^{-2}2^{2k}\mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega_1)\cap A_k(x_0,\omega_2)\}\right],\end{aligned}$$

and similarly

$$\Psi_k(x_0) \geq C_X^{-2}2^{2k}\mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega_1)\cap A_k(x_0,\omega_2)\}\right].$$

Together with (4.19) this yields

$$\frac{c_X}{C_X^2}2^{2k}\mathcal{V}_{\cap,k} \leq \Psi_k(x_0) \leq \frac{C_X}{c_X^2}2^{2k}\mathcal{V}_{\cap,k} \tag{4.20}$$

uniformly in $x_0 \in [0,1]^p$. That means for all x_0 that $\Psi_k(x_0) = \Theta(2^{2k}\mathcal{V}_{\cap,k})$. For the Ehrenfest CPRF we know that this has the exact rate 2^{-k}. With these observations we can prove bounds for $\zeta^n_{1,\omega}(x_0)$ and $\zeta^n_{r_n}(x_0)$, which will allow the application of Theorem 2.13.

4.4 Proofs

This section contains all the proofs in the chapter. First, the proof of Theorem 4.4 and its corollaries are presented in Section 4.4.1. Next, the proof of Proposition 4.1 and Corollary 4.2 are provided in Section 4.4.2. Subsequently, we prove Theorem 4.8 in Section 4.4.3. The proofs of Lemma 4.10 and Lemma 4.11 used there can be found in Section 4.4.5.

The proofs of Theorem 4.4 and Proposition 4.1 require the following proposition regarding the approximation error, the proof of which is postponed to Section 4.4.4.

Proposition 4.9. *Suppose* $\rho_I \sim Ber(N\binom{n}{r_n}^{-1})$, $0 < c_X \leq f_X \leq C_X$ *and* m *is Hölder continuous of order* α *with constant* C_H. *For the absolute approximation error it holds that*

$$\begin{aligned}&\mathbb{E}\left[\frac{\hat{N}}{N}\left|U^{(m)}_{n,r_n,N,\omega}(x_0) - m(x_0)\right|\right]\\ &\quad\leq \frac{2}{\sqrt{N}}\|m\|_\infty + C_H\mathbb{E}\left[\mathfrak{d}(A_k(x_0,\omega))^\alpha\right] + |m(x_0)|(1-c_X2^{-k})^{r_n}.\end{aligned}$$

and for a constant C *independent of all parameters and the distributions of all considered random variables the squared approximation error fulfills*

$$\begin{aligned}&\mathbb{E}\left[\frac{\hat{N}^2}{N^2}(U^{(m)}_{n,r_n,N,\omega}(x_0) - m(x_0))^2\right]\\ &\quad\leq C\left(\frac{1}{N}\|m\|^2_\infty + C_H^2\mathbb{E}\left[\mathfrak{d}(A_k(x_0,\omega))^\alpha\right]^2 + C_H^2\frac{p^\alpha}{\binom{n}{r_n}} + |m(x_0)|^2(1-c_X2^{-k})^{r_n}\right).\end{aligned}$$

In the complete U-statistic case it holds that

$$\mathbb{E}\left[\left|U^{(m)}_{n,r_n,\omega}(x_0) - m(x_0)\right|\right] \leq C_H\mathbb{E}\left[\mathfrak{d}(A_k(x_0,\omega))^\alpha\right] + |m(x_0)|(1-c_X2^{-k})^{r_n}$$

and

$$\begin{aligned}&\mathbb{E}\left[(U^{(m)}_{n,r_n,\omega}(x_0) - m(x_0))^2\right]\\ &\quad\leq \left(C_H^2\mathbb{E}\left[\mathfrak{d}(A_k(x_0,\omega))^\alpha\right]^2 + C_H^2\frac{p^\alpha}{\binom{n}{r_n}} + |m(x_0)|^2(1-c_X2^{-k})^{r_n}\right).\end{aligned}$$

4.4.1 Proof of Theorem 4.4, Corollary 4.6 and Corollary 4.7

In this section we gather the proof of Theorem 4.4 and its corollaries.

Proof of Theorem 4.4. We use that

$$U^{(\mathrm{RF})}_{n,r_n,N,\omega}(x_0) - m(x_0) = U^{(\varepsilon)}_{n,r_n,N,\omega}(x_0) + U^{(m)}_{n,r_n,N,\omega}(x_0) - m(x_0).$$

Condition (4.11) and (4.12) imply

$$\frac{n}{Nr_n2^k\mathcal{V}_{\cap,k}} = \frac{n}{N2^{2k}\mathcal{V}_{\cap,k}}\frac{2^k}{r_n} \to 0.$$

Together with (4.10) the second statement in Theorem 4.8 yields

$$\frac{U^{(\varepsilon)}_{n,r_n,N,\omega}(x_0)}{\sqrt{\Psi_k(x_0)/n}} \xrightarrow{d} \mathcal{N}(0,\sigma^2).$$

With (4.20) we have

$$(\Psi_k(x_0)/n)^{-1/2} = \mathcal{O}\big(n^{1/2}2^{-k}\mathcal{V}_{\cap,k}^{-1/2}\big)$$

Condition (4.12) implies

$$n2^{-2k}\mathcal{V}_{\cap,k}^{-1}(1-c_X2^{-k})^{2r_n} \to 0.$$

Thus, Proposition 4.9 yields

$$\begin{aligned}
&\mathbb{E}\left[\frac{\hat{N}}{N}|U^{(m)}_{n,r_n,N,\omega}(x_0) - m(x_0)|\right](\Psi_k(x_0)/n)^{-1/2} \\
&= \mathcal{O}\big(n^{1/2}2^{-k}\mathcal{V}_{\cap,k}^{-1/2}N^{-1/2}\big) + \mathcal{O}\big(\mathbb{E}\left[\mathfrak{d}(A_k(x_0,\omega))^{\alpha}\right]n^{1/2}2^{-k}\mathcal{V}_{\cap,k}^{-1/2}\big) \\
&\qquad + \mathcal{O}\big(n^{1/2}2^{-k}\mathcal{V}_{\cap,k}^{-1/2}(1-c_X2^{-k})^{r_n}\big) \\
&\to 0
\end{aligned}$$

due to (4.11) and (4.13). The assumption $N \to \infty$ implies $\hat{N}/N \xrightarrow{\mathbb{P}} 1$ and with Slutsky's theorem we obtain

$$\frac{U^{(m)}_{n,r_n,N,\omega}(x_0) - m(x_0)}{\sqrt{\Psi_k(x_0)/n}} \xrightarrow{\mathbb{P}} 0$$

which implies the claim. □

Proof of Corollary 4.6. The corollary follows directly from Theorem 4.4 by plugging in $\mathfrak{d}(A_k(x_0,\omega)) = \mathcal{O}(2^{-k/p})$ and $\mathcal{V}_{\cap,k} = \Theta(2^{-k})$ which are the results of Corollary 3.9 and Corollary 3.10. For the second statement let $k = \lfloor \log_2((\log n)n^{\frac{1}{(1+2\alpha/p)}})\rfloor$ and $r_n = \lfloor n(\log n)^{-1}\rfloor$. The undersmoothing assumption holds because

$$n2^{-k(1+2\alpha/p)} \lesssim n\left((\log n)n^{\frac{1}{(1+2\alpha/p)}}\right)^{-(1+2\alpha/p)} \lesssim (\log n)^{-1} \to 0.$$

Further the choice for r_n yields

$$\begin{aligned}
\frac{2^k}{r_n}\log\left(n2^{-k}\right) &\lesssim (\log n)^2 n^{\frac{1}{(1+2\alpha/p)}-1}\log\left(n2^{-k}\right) \\
&= (\log n)^2\log\left(n2^{-k}\right)n^{-\frac{2\alpha}{p+2\alpha}} \to 0
\end{aligned}$$

and $r_n = o(n)$. Lastly (4.20) yields

$$\Psi_k(x_0)/n = \Theta(2^{2k}\mathcal{V}_{\cap,k}/n) = \Theta(2^k/n) = \Theta((\log n)n^{-\frac{2\alpha}{2\alpha+p}}).$$ □

Proof of Corollary 4.7. The corollary is directly deduced from (3.23) and (3.25), as these yield

$$\mathbb{E}[\mathfrak{d}(A_k(x_0,\omega))^\alpha] \leq p\left(\frac{p-1+2^{-\alpha}}{p}\right)^k$$

and $\mathcal{V}_{\cap,k} \gtrsim 2^{-k}k^{-(p-1)}$. For $b(p,\alpha) = (p-1+2^{-\alpha})/p < 1$ and $\nu = \frac{2\log_2 b(p,\alpha)}{2\log_2 b(p,\alpha)-1}$ as in Corollary 4.2 choose

$$\begin{aligned} k &= \log_2(c_k(n(\log_2 n)^p)^{1-\nu}) \\ &= (1-\nu)\log_2(n) + \log_2(c_k(\log_2 n)^{p(1-\nu)}) \\ &= \mathcal{O}(\log_2 n) \end{aligned} \tag{4.21}$$

and $r_n = \Theta(n(\log_2 n)^{-p})$. This yields

$$\frac{r_n k^{(p-1)}}{n} \lesssim (\log_2 n)^{-1} \to 0$$

as well as

$$\begin{aligned} \frac{2^k}{r_n}\log\left(n2^{-k}k^{(p-1)}\right) &\lesssim \frac{(n(\log_2 n)^p)^{1-\nu}}{n(\log_2 n)^{-p}}\log\left(nc_k^{-1}(n(\log_2 n)^p)^{\nu-1}(\log_2 n)^{p-1}\right) \\ &= n^{-\nu}(\log_2 n)^{p(2-\nu)}\log\left(n^\nu c_k^{-1}(\log_2 n)^{p\nu-1}\right) \to 0. \end{aligned}$$

We obtain

$$\begin{aligned} b(p,\alpha)^{2k} &= (2^k)^{2\log_2 b(p,\alpha)} \\ &= \left(c_k^{1/(1-\nu)} n(\log_2 n)^p\right)^{(1-\nu)2\log_2 b(p,\alpha)} \\ &\lesssim \left(n\log_2^p n\right)^{-\nu}. \end{aligned}$$

which yields

$$\begin{aligned} b(p,\alpha)^{2k}\frac{nk^{(p-1)}}{2^k} &\lesssim \left(n\log_2^p n\right)^{-\nu}\frac{n\left(\log_2(c_k(n(\log_2 n)^p)^{1-\nu})\right)^{p-1}}{(n(\log_2 n)^p)^{1-\nu}} \\ &= (\log_2 n)^{-p}\left((1-\nu)\log_2 n + \log_2(c_k(\log_2 n)^{(1-\nu)p})\right)^{p-1} \\ &\lesssim (\log_2 n)^{-p}\left((1-\nu)\log_2 n\right)^{p-1} \\ &\lesssim (\log_2 n)^{-1} \to 0. \end{aligned}$$

Lastly (4.21), (4.20) and (3.25) yield

$$\begin{aligned} \Psi_k(x_0)/n &= \Theta(2^{2k}\mathcal{V}_{\cap,k}/n) \\ &= \mathcal{O}\left(2^k k^{-(p-1)/2}/n\right). \\ &= \mathcal{O}\left((n(\log_2 n)^p)^{1-\nu}(\log_2 n)^{-(p-1)/2}/n\right) \\ &= \mathcal{O}\left(n^{-\nu}(\log_2 n)^{p(1-\nu)-(p-1)/2}\right) \\ &= \mathcal{O}\left(n^{-\nu}(\log_2 n)^{p(1/2-\nu)+1/2}\right). \end{aligned}$$

□

4.4.2 Proof of Proposition 4.1 and Corollary 4.2

Proof of Proposition 4.1. We use that

$$\mathbb{E}\left[\left(U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0) - m(x_0)\right)^2\right] = \mathbb{E}\left[\left(U^{(m)}_{n,r_n,\omega}(x_0) - m(x_0)\right)^2\right] + \mathbb{E}\left[U^{(\varepsilon)}_{n,r_n,\omega}(x_0)^2\right].$$

In the complete U-statistic case Proposition 4.9 yields

$$\begin{aligned}&\mathbb{E}\left[\left(U^{(m)}_{n,r_n,\omega}(x_0) - m(x_0)\right)^2\right]\\ &\quad= \mathcal{O}\left(\mathbb{E}\left[\mathfrak{d}(A_k(x_0,\omega))^\alpha\right]^2\right) + \mathcal{O}((1-c_X 2^{-k})^{r_n}) + \mathcal{O}\left(\binom{n}{r_n}^{-1}\right). \qquad (4.22)\end{aligned}$$

For the second term (2.27) implies with Lemma 4.11, Lemma 4.10 and (4.20) that

$$\begin{aligned}\mathbb{E}\left[\left(U^{(\varepsilon)}_{n,r_n,\omega}(x_0)\right)^2\right] &= \mathrm{Var}\left(U^{(\varepsilon)}_{n,r_n,\omega}(x_0)\right)\\ &\le \frac{r_n^2}{n}\zeta^n_{1,\omega}(x_0) + \frac{r_n^2}{n^2}\zeta^n_{r_n}(x_0)\\ &= \mathcal{O}\left(\frac{1}{n}\Psi_k(x_0)\right) + \mathcal{O}\left(\frac{2^k r_n}{n^2}\right)\\ &= \mathcal{O}\left(\frac{2^{2k}\mathcal{V}_{\cap,k}}{n}\right) + \mathcal{O}\left(\frac{2^k r_n}{n^2}\right). \qquad (4.23)\end{aligned}$$

Noting that $r_n < n$ implies that

$$\frac{1}{\binom{n}{r_n}} = \mathcal{O}\left(\frac{1}{n}\right) = \mathcal{O}\left(\frac{2^{2k}\mathcal{V}_{\cap,k}}{n}\right),$$

(4.22) and (4.23) yield the claim. □

Proof of Corollary 4.2 . For the whole proof we can omit the term with the binomial coefficient from Proposition 4.1, because $\binom{n}{r_n} \ge n$ if $r_n < n$. We use Proposition 4.1, Corollary 3.9 and Corollary 3.10 to obtain

$$\begin{aligned}\mathbb{E}\left[\left(U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0) - m(x_0)\right)^2\right] &= \mathcal{O}(\mathbb{E}\left[\mathfrak{d}(A_k(x_0,\omega_I))^\alpha\right]^2) + \mathcal{O}((1-c_X 2^{-k})^{r_n})\\ &\quad + \mathcal{O}\big(\frac{2^{2k}\mathcal{V}_{\cap,k}}{n}\big) + \mathcal{O}\big(\frac{2^k r_n}{n^2}\big)\\ &= \mathcal{O}(2^{-2\alpha k/p}) + \mathcal{O}((1-c_X 2^{-k})^{r_n}) + \mathcal{O}\big(\frac{2^k}{n}\big)\end{aligned}$$

because $r_n/n \le 1$. If we choose $k = \lfloor \log_2(c_k n^{\frac{1}{(1+2\alpha/p)}})\rfloor$ which implies $2^k = \Theta(n^{\frac{1}{(1+2\alpha/p)}})$ we get

$$\begin{aligned}\mathbb{E}\left[\left(U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0) - m(x_0)\right)^2\right] &= \mathcal{O}(2^{-2\alpha k/p}) + \mathcal{O}\left(2^k/n\right) + \mathcal{O}((1-c_X 2^{-k})^{r_n})\\ &= \mathcal{O}(n^{-\frac{2\alpha/p}{(1+2\alpha/p)}}) + \mathcal{O}\big(n^{\frac{1}{(1+2\alpha/p)}-1}\big) + \mathcal{O}((1-c_X n^{-\frac{1}{(1+2\alpha/p)}})^{r_n})\end{aligned}$$

$$= \mathcal{O}(n^{-\frac{2\alpha}{p+2\alpha}})$$

because

$$(1 - c_X n^{-\frac{1}{(1+2\alpha/p)}})^{r_n} n^{\frac{2\alpha}{p+2\alpha}} = \mathcal{O}\left(\exp(-c_X r_n n^{-\frac{1}{(1+2\alpha/p)}} + \tfrac{2\alpha}{p+2\alpha}\log n)\right) = o(1)$$

due to the assumption on r_n. For the uniform CPRF, Proposition 4.1, (3.23) and (3.25) yield

$$\begin{aligned}\mathbb{E}\left[\left(U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0) - m(x_0)\right)^2\right] &= \mathcal{O}\left(\left(\frac{p-1+2^{-\alpha}}{p}\right)^{2k}\right) + \mathcal{O}((1-c_X 2^{-k})^{r_n}) \\ &\quad + \mathcal{O}\left(\frac{2^k k^{-(p-1)/2}}{n}\right) + \mathcal{O}\left(\frac{2^k r_n}{n^2}\right). \end{aligned} \tag{4.24}$$

Using $r_n = \Theta(nk^{-(p-1)/2})$ the last two terms have the same rate. The choice of k implies $k = \mathcal{O}(\log_2 n)$. Using this, the choice of r_n and $2^k = c_k(n(\log_2 n)^{(p-1)/2})^{1-\nu}$ it holds that

$$\begin{aligned}(1 - c_X 2^{-k})^{r_n} &= \mathcal{O}(\exp(-c_X r_n 2^{-k})) \\ &= \mathcal{O}\left(\exp\left(-c_X n k^{-(p-1)/2} c_k^{-1}(n(\log_2 n)^{(p-1)/2})^{\nu-1}\right)\right) \\ &= \mathcal{O}\left(\exp\left(-c_X c_k^{-1} n^{\nu}(\log_2 n)^{-(p-1)/2}(\log_2 n)^{(p-1)(\nu-1)/2}\right)\right) \\ &= \mathcal{O}\left(\exp\left(-c_X c_k^{-1} n^{\nu}(\log_2 n)^{(\nu-2)(p-1)/2}\right)\right) \\ &= \mathcal{O}\left(\left(n(\log_2 n)^{(p-1)/2}\right)^{-\nu}\right)\end{aligned}$$

because $\nu > 0$. Thus, it remains to handle the first and third term in (4.24). Recalling the definition of $b(p,\alpha)$ and ν we get

$$\begin{aligned}b(p,\alpha)^{2k} &= 2^{2k\log_2 b(p,\alpha)} \\ &= (2^k)^{2\log_2 b(p,\alpha)} \\ &= \left(n(\log_2 n)^{(p-1)/2}\right)^{(1-\nu)2\log_2 b(p,\alpha)} \\ &= \left(n(\log_2 n)^{(p-1)/2}\right)^{-\nu}.\end{aligned}$$

For the last term we obtain

$$\begin{aligned}\frac{2^k k^{-(p-1)/2}}{n} &= n^{-1}(n(\log_2 n)^{(p-1)/2})^{1-\nu}\left\{\log_2\left(c_k(n(\log_2 n)^{(p-1)/2})^{1-\nu}\right)\right\}^{-(p-1)/2} \\ &= n^{-\nu}\left((\log_2 n)^{(p-1)/2}\right)^{1-\nu}\left\{\log_2\left(c_k(n(\log_2 n)^{(p-1)/2})^{1-\nu}\right)\right\}^{-(p-1)/2} \\ &= n^{-\nu}\left((\log_2 n)^{(p-1)/2}\right)^{1-\nu}\left\{\log_2(n^{1-\nu}) + \log_2\left(c_k(\log_2 n)^{(p-1)(1-\nu)/2}\right)\right\}^{-(p-1)/2} \\ &\lesssim n^{-\nu}\left((\log_2 n)^{(p-1)/2}\right)^{1-\nu}\left\{\log_2(n^{1-\nu})\right\}^{-(p-1)/2} \\ &\lesssim n^{-\nu}\left((\log_2 n)^{(p-1)/2}\right)^{1-\nu}\left\{\log_2^{p-1}(n)\right\}^{-1/2} \\ &= \left(n(\log_2 n)^{(p-1)/2}\right)^{-\nu}\end{aligned}$$

which completes the proof. □

4.4.3 Proof of Theorem 4.8

The proof of Theorem 4.8 relies on the following two lemmas, which deal with the functions $\zeta_{r_n}^n(x_0)$ and $\zeta_{1,\omega}^n(x_0)$. Their proofs are postponed to Section 4.4.5.

Lemma 4.10. *For the $\zeta_{r_n}^n$ that corresponds to the incomplete U-statistic $U_{n,r_n,N,\omega}^{(\varepsilon)}(x_0)$ it holds that*

$$\zeta_{r_n}^n(x_0) = \sigma^2 \mathbb{E}\left[\frac{\mathbb{I}\{\sum_{i=1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0,\omega)\} > 0\}}{\sum_{i=1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0,\omega)\}}\right] \leq \sigma^2 \frac{2^{k+1}}{(r_n+1)c_X}.$$

Lemma 4.11. *We define*

$$\lambda(h,\xi) := 1 - (1-\xi)^h. \tag{4.25}$$

For the $\zeta_{1,\omega}^n(x_0)$ from (2.17) that corresponds to the incomplete U-statistic $U_{n,r_n,N,\omega}^{(\varepsilon)}(x_0)$ it holds that

$$\begin{aligned}
h_{n,1}^{(\varepsilon)}(Z_1) &= \frac{\varepsilon_1}{r_n}\mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega)\} p_{x_0}(\omega)^{-1} \lambda(r_n, p_{x_0}(\omega)) | X_1\right], \qquad (4.26)\\
\zeta_{1,\omega}^n(x_0) &= \frac{\sigma^2}{r_n^2}\mathbb{E}\left[\frac{\mathbb{I}\{X_1 \in A_k(x_0,\omega_1) \cap A_k(x_0,\omega_2)\}}{p_{x_0}(\omega_1)p_{x_0}(\omega_2)} \lambda(r_n, p_{x_0}(\omega_1))\lambda(r_n, p_{x_0}(\omega_2))\right],\\
\zeta_{1,\omega}^n(x_0) &\geq \frac{\sigma^2}{r_n^2}\Psi_k(x_0)\lambda(r_n, c_X 2^{-k})^2,\\
\zeta_{1,\omega}^n(x_0) &\leq \frac{\sigma^2}{r_n^2}\Psi_k(x_0).
\end{aligned}$$

Remark 4.12. If X is uniformly distributed we have $p_{x_0}(\omega) = 2^{-k}$ and therefore

$$\zeta_{1,\omega}^n(x_0) = \sigma^2 \mathcal{V}_{\cap,k} \left(\frac{2^k}{r_n}\right)^2 \left(1 - (1-2^{-k})^{r_n}\right)^2.$$

Proof of Theorem 4.8. We want to apply Theorem 2.13 and therefore need to check its assumptions. We prove (2.24), (2.25) and the condition on the moments of $h_n^{(\varepsilon)}$ in this order. Using the results of Lemma 4.10 and Lemma 4.11 we get

$$\begin{aligned}
\frac{\zeta_{r_n}^n(x_0)}{n\zeta_{1,\omega}^n(x_0)} &\leq \frac{1}{n}\frac{2^{k+1}}{(r_n+1)c_X} r_n^2 \Psi_k^{-1}(x_0)\lambda(r_n, c_X 2^{-k})^{-2}\\
&\leq \frac{2}{c_X}\frac{2^k r_n}{n}\frac{C_X^2}{c_X} 2^{-2k}\mathcal{V}_{\cap,k}^{-1}\lambda(r_n, c_X 2^{-k})^{-2}\\
&= \frac{2C_X^2}{c_X^2}\frac{r_n}{n 2^k \mathcal{V}_{\cap,k}}\lambda(r_n, c_X 2^{-k})^{-2}
\end{aligned}$$

because of (4.20). Using the assumption that $2^k \leq c r_n$ we get that

$$\begin{aligned}
\lambda(r_n, c_X 2^{-k})^{-2} &= \left(1 - (1 - c_X 2^{-k})^{r_n}\right)^{-2}\\
&\leq \left(1 - (1 - \frac{c_X}{c r_n})^{r_n}\right)^{-2} \to (1 - \exp(-c_X/c))^{-2}
\end{aligned}$$

which implies $\lambda(r_n, c_X 2^{-k})^{-2} = \mathcal{O}(1)$. Therefore (2.24) holds with (4.17). We proceed with (2.25) and first note that (4.26) from Lemma 4.11, (4.6) and $\lambda(r_n, p_{x_0}(\omega)) \leq 1$ yield

$$\begin{aligned} h_{n,1}^{(\varepsilon)}(Z_1) &= \frac{\varepsilon_1}{r_n} \mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega)\} p_{x_0}(\omega)^{-1} \lambda(r_n, p_{x_0}(\omega)) | X_1\right] \\ &\leq \varepsilon_1 \frac{2^k}{c_X r_n} \mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega)\} | X_1\right]. \end{aligned}$$

For $\delta = 1$ we obtain with the lower bound for $\zeta_{1,\omega}^n$ from Lemma 4.11, $\lambda(r_n, c_X 2^{-k})^{-1} = \mathcal{O}(1)$ from above and (4.20) that

$$\begin{aligned} \frac{\mathbb{E}\left[|h_{n,1}^{(\varepsilon)}(Z_1)|^3\right]^2}{n(\zeta_{1,\omega}^n)^3} &\leq \mathbb{E}\left[|\varepsilon_1|^3\right]^2 \left(\frac{2^k}{c_X r_n}\right)^6 \frac{\mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega)\} | X_1\right]^3\right]^2}{n\left(\frac{\sigma^2}{r_n^2} \Psi_k(x_0) \lambda(r_n, c_X 2^{-k})^2\right)^3} \\ &\lesssim 2^{6k} \frac{\mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega)\} | X_1\right]^2\right]^2}{n\left(2^{2k} \mathcal{V}_{\cap,k}\right)^3} \\ &= \frac{1}{n \mathcal{V}_{\cap,k}} \\ &= \mathcal{O}\left(\frac{r_n}{n 2^k \mathcal{V}_{\cap,k}}\right) \to 0 \end{aligned}$$

due to (4.17).

We note that $\vartheta_n = \mathbb{E}[h_n^{(\varepsilon)}] = 0$, hence the last condition we need to check is

$$\frac{\mathbb{E}\left[|h_n^{(\varepsilon)}(x_0, Z_1, \ldots, Z_{r_n}, \omega)|^{2l}\right]}{\mathbb{E}\left[|h_n^{(\varepsilon)}(x_0, Z_1, \ldots, Z_{r_n}, \omega)|^{l}\right]^2} \leq C$$

for $l = 2, 3$ and a constant C. We recall the definition of $h_n^{(\varepsilon)}$ from (3.11) and the abbreviated notation of the weights from (3.10) with

$$h_n^{(\varepsilon)}(x_0, Z_1, \ldots, Z_{r_n}, \omega) = \sum_{j=1}^{r_n} \varepsilon_j W_{j,k}(x_0, \omega).$$

We start with $l = 2$. We use that $\mathbb{E}[\varepsilon_1] = 0$ and $c_X 2^{-k} \leq p_{x_0}(\omega) \leq C_X 2^{-k}$, and with the usual arguments involving the conditional expectation we obtain

$$\begin{aligned} &\mathbb{E}\left[h_n^{(\varepsilon)}(x_0, Z_1, \ldots, Z_{r_n}, \omega)^4\right] \\ &= \mathbb{E}\left[\left(\sum_{j=1}^{r_n} \varepsilon_j W_{j,k}(x_0,\omega)\right)^4\right] \\ &= \mathbb{E}\left[\sum_{j=1}^{r_n} \varepsilon_j^4 W_{j,k}(x_0,\omega)^4\right] + 3\mathbb{E}\left[\sum_{j=1}^{r_n} \sum_{i=1, i\neq j}^{r_n} \varepsilon_j^2 \varepsilon_i^2 W_{j,k}(x_0,\omega)^2 W_{i,k}(x_0,\omega)^2\right] \end{aligned}$$

$$
\begin{aligned}
&= r_n \mathbb{E}\left[\varepsilon_1^4\right] \mathbb{E}\left[\frac{\mathbb{I}\{X_1 \in A_k(x_0,\omega)\}}{\left(1+\sum_{i=2}^{r_n}\mathbb{I}\{X_i \in A_k(x_0,\omega)\}\right)^4}\right] \\
&\quad + 3r_n(r_n-1)\sigma^4 \mathbb{E}\left[\frac{\mathbb{I}\{X_1, X_2 \in A_k(x_0,\omega)\}}{\left(2+\sum_{i=3}^{r_n}\mathbb{I}\{X_i \in A_k(x_0,\omega)\}\right)^4}\right] \\
&\leq r_n 2^{-k} C_X \mathbb{E}\left[\varepsilon_1^4\right] \mathbb{E}\left[\left(1+\sum_{i=2}^{r_n}\mathbb{I}\{X_i \in A_k(x_0,\omega)\}\right)^{-4}\right] \\
&\quad + 3r_n(r_n-1)2^{-2k} C_X^2 \sigma^4 \mathbb{E}\left[\left(2+\sum_{i=3}^{r_n}\mathbb{I}\{X_i \in A_k(x_0,\omega)\}\right)^{-4}\right] \\
&\leq r_n 2^{-k} C_X \left(\mathbb{E}\left[\varepsilon_1^4\right] + 3r_n 2^{-k} C_X \sigma^4\right) \mathbb{E}\left[\left(1+\sum_{i=2}^{r_n}\mathbb{I}\{X_i \in A_k(x_0,\omega)\}\right)^{-4}\right] \\
&\leq r_n^2 2^{-2k} C_X \left(\mathbb{E}\left[\varepsilon_1^4\right] c + 3C_X\sigma^4\right) 2^{4k} c_X^{-4} \frac{24}{r_n^4} \\
&= r_n^{-2} 2^{2k} \frac{24 C_X}{c_X^4}\left(\mathbb{E}\left[\varepsilon_1^4\right] c + 3C_X\sigma^4\right) \\
&= \mathcal{O}(r_n^{-2}2^{2k}) \qquad (4.27)
\end{aligned}
$$

by using Lemma 2.15. Similarly we obtain

$$
\begin{aligned}
&\mathbb{E}\left[h_n^{(\varepsilon)}(x_0, Z_1, \ldots, Z_{r_n}, \omega)^2\right] \\
&\quad = \mathbb{E}\left[\left(\sum_{j=1}^{r_n}\varepsilon_j W_{j,k}(x_0,\omega)\right)^2\right] \\
&\quad = \mathbb{E}\left[\sum_{j=1}^{r_n}\varepsilon_j^2 W_{j,k}^2(x_0,\omega)\right] \\
&\quad = \sigma^2 r_n \mathbb{E}\left[\frac{\mathbb{I}\{X_1 \in A_k(x_0,\omega)\}}{\left(1+\sum_{i=2}^{r_n}\mathbb{I}\{X_i \in A_k(x_0,\omega)\}\right)^2}\right] \\
&\quad \geq \sigma^2 r_n c_X 2^{-k} \mathbb{E}\left[\frac{1}{\left(1+\sum_{i=2}^{r_n}\mathbb{I}\{X_i \in A_k(x_0,\omega)\}\right)^2}\right] \\
&\quad \geq \sigma^2 r_n c_X 2^{-k} 2^{2k} C_X^{-2} \frac{2}{r_n(r_n+1)}\left(1-(r_n+1)C_X 2^{-k}(1-c_X 2^{-k})^{r_n} - (1-c_X 2^{-k})^{r_n+1}\right) \\
&\quad = \sigma^2 \frac{2c_X}{C_X^2}\frac{2^k}{(r_n+1)}\left(1-(r_n+1)C_X 2^{-k}(1-c_X 2^{-k})^{r_n} - (1-c_X 2^{-k})^{r_n+1}\right). \qquad (4.28)
\end{aligned}
$$

If $2^k/r_n$ is sufficiently small the term in the brackets is larger or equal than a constant. With (4.27) this leads to

$$
\frac{\mathbb{E}\left[h_n^{(\varepsilon)}(x_0, Z_1, \ldots, Z_{r_n}, \omega)^4\right]}{\mathbb{E}\left[h_n^{(\varepsilon)}(x_0, Z_1, \ldots, Z_{r_n}, \omega)^2\right]^2} = \mathcal{O}\Big(r_n^{-2}2^{2k}\frac{(r_n+1)^2}{2^{2k}}\Big) = \mathcal{O}(1).
$$

We continue with $l = 3$. With (2.7) and analogue arguments to the first case we get

$$\begin{aligned}
&\mathbb{E}\left[h_n^{(\varepsilon)}(x_0, Z_1, \dots, Z_{r_n}, \omega)^6\right] \\
&= \mathbb{E}\left[\left(\sum_{j=1}^{r_n} \varepsilon_j W_{j,k}(x_0, \omega)\right)^6\right] \\
&\leq C_{6,2}\mathbb{E}\left[\left(\sum_{j=1}^{r_n} \varepsilon_j^2 W_{j,k}^2(x_0, \omega)\right)^3\right] \\
&= C_{6,2}\mathbb{E}\left[\sum_{j=1}^{r_n}\sum_{i=1}^{r_n}\sum_{l=1}^{r_n} \varepsilon_j^2\varepsilon_i^2\varepsilon_l^2 W_{j,k}(x_0, \omega)^2 W_{i,k}(x_0, \omega)^2 W_{l,k}(x_0, \omega)^2\right] \\
&= C_{6,2}\sum_{j=1}^{r_n}\sum_{i=1}^{r_n}\sum_{l=1}^{r_n} \mathbb{E}\left[\varepsilon_j^2\varepsilon_i^2\varepsilon_l^2 W_{j,k}(x_0, \omega)^2 W_{i,k}(x_0, \omega)^2 W_{l,k}(x_0, \omega)^2\right] \\
&= C_{6,2}\sum_{j=1}^{r_n} \mathbb{E}\left[\varepsilon_j^6 W_{j,k}(x_0, \omega)^6\right] \\
&\quad + C_{6,2}3\sum_{j=1}^{r_n}\sum_{i=1, i\neq j}^{r_n} \mathbb{E}\left[\varepsilon_j^4\varepsilon_i^2 W_{j,k}(x_0, \omega)^4 W_{i,k}(x_0, \omega)^2\right] \\
&\quad + C_{6,2}\sum_{j=1}^{r_n}\sum_{i=1, i\neq j}^{r_n}\sum_{l=1, l\notin\{j,i\}}^{r_n} \mathbb{E}\left[\varepsilon_j^2\varepsilon_i^2\varepsilon_l^2 W_{j,k}(x_0, \omega)^2 W_{i,k}(x_0, \omega)^2 W_{l,k}(x_0, \omega)^2\right] \\
&= C_{6,2} r_n \mathbb{E}\left[\varepsilon_1^6\right] \mathbb{E}\left[W_{1,k}(x_0, \omega)^6\right] + 3r_n(r_n - 1)\mathbb{E}\left[\varepsilon_1^4\varepsilon_2^2\right] \mathbb{E}\left[W_{1,k}(x_0, \omega)^4 W_{2,k}(x_0, \omega)^2\right] \\
&\quad + C_{6,2} r_n(r_n - 1)(r_n - 2)\sigma^6 \mathbb{E}\left[W_{1,k}(x_0, \omega)^2 W_{2,k}(x_0, \omega)^2 W_{3,k}(x_0, \omega)^2\right] \\
&\leq C_{6,2} r_n \mathbb{E}\left[\varepsilon_1^6\right] \mathbb{E}\left[W_{1,k}(x_0, \omega)^6\right] + 3r_n(r_n - 1)\mathbb{E}\left[\varepsilon_1^4\varepsilon_2^2\right] \mathbb{E}\left[W_{1,k}(x_0, \omega)^4 W_{2,k}(x_0, \omega)^2\right] \\
&\quad + C_{6,2} r_n(r_n - 1)(r_n - 2)\sigma^6 \mathbb{E}\left[W_{1,k}(x_0, \omega)^2 W_{2,k}(x_0, \omega)^2 W_{3,k}(x_0, \omega)^2\right] \\
&= \mathcal{O}(r_n^3 2^{-3k})\mathbb{E}\left[\left(1 + \sum_{i=2}^{r_n} \mathbb{I}\{X_i \in A_k(x_0, \omega)\}\right)^{-6}\right] \\
&= \mathcal{O}(r_n^{-3} 2^{3k})
\end{aligned}$$

due to Lemma 2.15. With (4.28) we get

$$\mathbb{E}\left[\left|\sum_{j=1}^{r_n} \varepsilon_j W_{j,k}(x_0, \omega)\right|^3\right]^{-2} \leq \mathbb{E}\left[\left(\sum_{j=1}^{r_n} \varepsilon_j W_{j,k}(x_0, \omega)\right)^2\right]^{-3} = \mathcal{O}\big(r_n^3 2^{-3k}\big).$$

Together this yields

$$\frac{\mathbb{E}\left[|h_n^{(\varepsilon)}(x_0, Z_1, \dots, Z_{r_n}, \omega)|^6\right]}{\mathbb{E}\left[|h_n^{(\varepsilon)}(x_0, Z_1, \dots, Z_{r_n}, \omega)|^3\right]^2} = \mathcal{O}(1).$$

Since all its conditions are met Theorem 2.13 now yields the first claim

$$\frac{U_{n,r_n,N,\omega}^{(\varepsilon)}(x_0)}{\sqrt{r_n^2 \zeta_{1,\omega}^n(x_0)/n + \zeta_{r_n}^n(x_0)/N}} \xrightarrow{d} \mathcal{N}(0, 1).$$

For the second statement in Theorem 4.8 it remains to prove that

$$\frac{r_n^2\zeta_{1,\omega}^n(x_0)/n + \zeta_{r_n}^n(x_0)/N}{\Psi_k(x_0)/n} \to \sigma^2.$$

With Lemma 4.10, (4.18) and (4.20) we obtain

$$\frac{n\zeta_{r_n}^n(x_0)}{N\Psi_k(x_0)} \le \sigma^2 \frac{n2^{k+1}}{(r_n+1)c_X} \frac{C_X^2}{c_X N 2^{2k}\mathcal{V}_{\cap,k}} = \mathcal{O}\Big(\frac{n}{Nr_n 2^k \mathcal{V}_{\cap,k}}\Big) \to 0.$$

Lemma 4.11 yields

$$\begin{aligned} &\frac{r_n^2\zeta_{1,\omega}^n(x_0)}{\Psi_k(x_0)} \le \sigma^2 \\ \text{and} \qquad &\frac{r_n^2\zeta_{1,\omega}^n(x_0)}{\Psi_k(x_0)} \ge \sigma^2\big(1-(1-c_X 2^{-k})^{r_n}\big)^2 \to \sigma^2 \end{aligned}$$

because $2^k = o(r_n)$. Together this implies $r_n^2\zeta_{1,\omega}^n(x_0)\Psi_k^{-1}(x_0) \to \sigma^2$ and therefore the second claim of the theorem. □

4.4.4 Proof of Proposition 4.9

In order to show this bound we use the beneficial decomposition

$$\begin{aligned} &\frac{\hat{N}}{N}(U_{n,r_n,N,\omega}^{(\mathrm{RF})}(x_0) - m(x_0)) \\ &\quad= \frac{1}{N}\sum_{I\in B_{r_n,n}} \rho_I \sum_{j\in I} m(X_j)W_{j,k}(x_0,I) - \frac{\hat{N}}{N}m(x_0) \\ &\quad= \frac{1}{N}\sum_{I\in B_{r_n,n}} \rho_I \sum_{j\in I} m(X_j)W_{j,k}(x_0,I) - \frac{1}{N}\sum_{I\in B_{r_n,n}} \rho_I m(x_0) \\ &\quad= \frac{1}{N}\sum_{I\in B_{r_n,n}} \rho_I\big(\sum_{j\in I} m(X_j)W_{j,k}(x_0,I) - m(x_0)\big) \\ &\quad= \sum_{I\in B_{r_n,n}} \left(\rho_I/N - \binom{n}{r_n}^{-1}\right)\left(\sum_{j\in I} m(X_j)W_{j,k}(x_0,I) - m(x_0)\right) \\ &\qquad\quad + \binom{n}{r_n}^{-1} \sum_{I\in B_{r_n,n}} \left(\sum_{j\in I} m(X_j)W_{j,k}(x_0,I) - m(x_0)\right) \\ &\quad= \sum_{I\in B_{r_n,n}} \left(\rho_I/N - \binom{n}{r_n}^{-1}\right)\left(\sum_{j\in I} m(X_j)W_{j,k}(x_0,I) - m(x_0)\right) \\ &\qquad\quad + \binom{n}{r_n}^{-1} \sum_{I\in B_{r_n,n}} \sum_{j\in I} (m(X_j)-m(x_0))W_{j,k}(x_0,I) \\ &\qquad\quad + m(x_0)\binom{n}{r_n}^{-1} \sum_{I\in B_{r_n,n}} \left(\sum_{j\in I} W_{j,k}(x_0,I) - 1\right). \end{aligned}$$

We note that Definition 3.2 implies

$$\sum_{j\in I} W_{j,k}(x_0, I) = \mathbb{I}\{\exists j \in I : X_j \in A_k(x_0, \omega_I)\} \leq 1. \tag{4.29}$$

Due to the fact that the ρ_I are independent of all other random variables and $\mathbb{E}[\rho_I/N - \binom{n}{r_n}^{-1}] = 0$ it holds that

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{I\in B_{r_n,n}}\left(\rho_I/N - \binom{n}{r_n}^{-1}\right)\left(\sum_{j\in I} m(X_j)W_{j,k}(x_0,I) - m(x_0)\right)\right)^2\right] \\
&= \mathbb{E}\left[\sum_{I\in B_{r_n,n}}\left(\rho_I/N - \binom{n}{r_n}^{-1}\right)^2\left(\sum_{j\in I} m(X_j)W_{j,k}(x_0,I) - m(x_0)\right)^2\right] \\
&= \binom{n}{r_n}\mathbb{E}\left[\left(\frac{\rho_{[r_n]}}{N} - \binom{n}{r_n}^{-1}\right)^2\right]\mathbb{E}\left[\left(\sum_{j\in I} m(X_j)W_{j,k}(x_0,I) - m(x_0)\right)^2\right] \\
&\leq \frac{\binom{n}{r_n}}{N^2}\mathrm{Var}(\rho_I)\mathbb{E}\left[\left(\|m\|_\infty \sum_{j\in I} W_{j,k}(x_0,I) + |m(x_0)|\right)^2\right] \\
&\leq \frac{\binom{n}{r_n}}{N^2}\frac{N}{\binom{n}{r_n}}\left(1 - N/\binom{n}{r_n}\right)4\|m\|_\infty^2 \\
&\leq \frac{1}{N}4\|m\|_\infty^2.
\end{aligned}$$

For the second term we get

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}}\sum_{j\in I}\big(m(X_j) - m(x_0)\big)W_{j,k}(x_0,I)\right|\right] \\
&\leq C_H\mathbb{E}\left[\frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}}\mathfrak{d}(A_k(x_0,\omega_I))^\alpha\sum_{j\in I}W_{j,k}(x_0,I)\right] \\
&\leq C_H\mathbb{E}\left[\frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}}\mathfrak{d}(A_k(x_0,\omega_I))^\alpha\right] \\
&= C_H\mathbb{E}\left[\mathfrak{d}(A_k(x_0,\omega))^\alpha\right].
\end{aligned}$$

Noting that the diameter of the p-dimensional hypercube is $\sqrt{p}$, we further obtain

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}}\sum_{j\in I}(m(X_j) - m(x_0))W_{j,k}(x_0,I)\right)^2\right] \\
&\leq C_H^2\mathbb{E}\left[\left(\frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}}\mathfrak{d}(A_k(x_0,\omega_I))^\alpha\sum_{j\in I}W_{j,k}(x_0,I)\right)^2\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq C_H^2 \mathbb{E}\left[\left(\frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \mathfrak{d}(A_k(x_0,\omega_I))^\alpha\right)^2\right] \\
&= C_H^2 \mathrm{Var}\left(\frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \mathfrak{d}(A_k(x_0,\omega_I))^\alpha\right) + C_H^2 \mathbb{E}\left[\mathfrak{d}(A_k(x_0,\omega))^\alpha\right]^2 \\
&= C_H^2 \frac{1}{\binom{n}{r_n}} \mathrm{Var}\left(\mathfrak{d}(A_k(x_0,\omega))^\alpha\right) + C_H^2 \mathbb{E}\left[\mathfrak{d}(A_k(x_0,\omega))^\alpha\right]^2 \\
&\leq C_H^2 \frac{1}{\binom{n}{r_n}} \mathbb{E}\left[\mathfrak{d}(A_k(x_0,\omega))^{2\alpha}\right] + C_H^2 \mathbb{E}\left[\mathfrak{d}(A_k(x_0,\omega))^\alpha\right]^2 \\
&\leq C_H^2 \frac{p^\alpha}{\binom{n}{r_n}} + C_H^2 \mathbb{E}\left[\mathfrak{d}(A_k(x_0,\omega))^\alpha\right]^2 .
\end{aligned}
$$

For the third term we get with (4.29)

$$
\begin{aligned}
&\mathbb{E}\left[|m(x_0)\frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \big(\sum_{j \in I} W_{j,k}(x_0,I) - 1\big)|\right] \\
&\quad = \mathbb{E}\left[|m(x_0)\frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \big(\mathbb{I}\{\exists j \in I : X_j \in A_k(x_0,\omega_I)\} - 1\big)|\right] \\
&\quad = |m(x_0)|\mathbb{E}\left[\frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \mathbb{I}\{\nexists j \in I : X_j \in A_k(x_0,\omega_I)\}\right] \\
&\quad = |m(x_0)|\mathbb{E}\left[\mathbb{I}\{\nexists j \in \{1,\dots r_n\} : X_j \in A_k(x_0,\omega)\}\right] \\
&\quad = |m(x_0)|\mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\{\nexists j \in \{1,\dots r_n\} : X_j \in A_k(x_0,\omega)\} \mid \omega\right]\right] \\
&\quad = |m(x_0)|\mathbb{E}\left[(1 - p_{x_0}(\omega))^{r_n}\right] \\
&\quad \leq |m(x_0)|(1 - c_X 2^{-k})^{r_n}
\end{aligned}
$$

and similarly

$$
\mathbb{E}\left[\left(m(x_0)\frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \left(\sum_{j \in I} W_{j,k}(x_0,I) - 1\right)\right)^2\right] \leq |m(x_0)|^2(1 - c_X 2^{-k})^{r_n}.
$$

Combining the three parts yields the claim. □

4.4.5 Proof of Auxiliary Lemmas

Here we provide the remaining proofs of the auxiliary Lemmas 4.10 and 4.11 that handle $\zeta_{r_n}^n$ and $\zeta_{1,\omega}^n$.

Proof of Lemma 4.10. We start by using the definition and get

$$
\zeta_{r_n}^n(x_0) = \mathrm{Var}(h_n^{(\varepsilon)}(x_0, Z_1, \dots, Z_{r_n}, \omega))
$$

$$
\begin{aligned}
&= \operatorname{Var}\left(\sum_{j=1}^{r_n} \frac{\varepsilon_j \mathbb{I}\{X_j \in A_k(x_0,\omega)\}}{\sum_{i=1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0,\omega)\}}\right) \\
&= \mathbb{E}\left[\operatorname{Var}\left(\sum_{j=1}^{r_n} \frac{\varepsilon_j \mathbb{I}\{X_j \in A_k(x_0,\omega)\}}{\sum_{i=1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0,\omega)\}} \,\Big|\, \omega, X_1, \ldots, X_{r_n}\right)\right] \\
&\quad + \operatorname{Var}\left(\mathbb{E}\left[\sum_{j=1}^{r_n} \frac{\varepsilon_j \mathbb{I}\{X_j \in A_k(x_0,\omega)\}}{\sum_{i=1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0,\omega)\}} \,\Big|\, \omega, X_1, \ldots, X_{r_n}\right]\right) \\
&= \mathbb{E}\left[\sum_{j=1}^{r_n} \left(\frac{\mathbb{I}\{X_j \in A_k(x_0,\omega)\}}{\sum_{i=1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0,\omega)\}}\right)^2 \operatorname{Var}\left(\varepsilon_j \mid \omega, X_1, \ldots, X_{r_n}\right)\right] \\
&= \sigma^2 \mathbb{E}\left[\frac{\sum_{j=1}^{r_n} \mathbb{I}\{X_j \in A_k(x_0,\omega)\}}{(\sum_{i=1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0,\omega)\})^2}\right] \\
&= \sigma^2 \mathbb{E}\left[\frac{\mathbb{I}\{\sum_{i=1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0,\omega)\} > 0\}}{\sum_{i=1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0,\omega)\}}\right] \\
&= \sigma^2 \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{I}\{\sum_{i=1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0,\omega)\} > 0\}}{\sum_{i=1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0,\omega)\}} \,\Big|\, \omega\right]\right] \\
&\leq \sigma^2 \mathbb{E}\left[\frac{2}{(r_n+1)p_{x_0}(\omega)}\right] \\
&\leq \sigma^2 \frac{2^{k+1}}{(r_n+1)c_X}
\end{aligned}
$$

by using the second part of Lemma 2.14 which is Lemma 4.1 by Györfi et al. (2002) for the Binomial distribution with parameters r_n and $p_{x_0}(\omega)$. □

Proof of Lemma 4.11. With (2.17) we have

$$\zeta_{1,\omega}^n(x_0) = \operatorname{Var}\big(h_{n,1}^{(\varepsilon)}(Z_1)\big).$$

For the conditional expectation we obtain

$$
\begin{aligned}
h_{n,1}^{(\varepsilon)}(Z_1) &= \mathbb{E}\left[h_n^{(\varepsilon)}(Z_1, \ldots, Z_{r_n}, \omega) \mid Z_1\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{r_n} \frac{\varepsilon_j \mathbb{I}\{X_j \in A_k(x_0,\omega)\}}{\sum_{i=1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0,\omega)\}} \,\Big|\, Z_1\right] \\
&= \mathbb{E}\left[\sum_{j=2}^{r_n} \frac{\varepsilon_j \mathbb{I}\{X_j \in A_k(x_0,\omega)\}}{\sum_{i=1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0,\omega)\}} \,\Big|\, Z_1\right] + \varepsilon_1 \mathbb{E}\left[\frac{\mathbb{I}\{X_1 \in A_k(x_0,\omega)\}}{\sum_{i=1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0,\omega)\}} \,\Big|\, Z_1\right] \\
&= \varepsilon_1 \mathbb{E}\left[\frac{\mathbb{I}\{X_1 \in A_k(x_0,\omega)\}}{\sum_{i=1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0,\omega)\}} \,\Big|\, X_1\right]
\end{aligned}
$$

since ε_1 is independent of $\omega, X_1, \ldots, X_{r_n}$ as well as $\varepsilon_2, \ldots, \varepsilon_{r_n}$. With the first part of Lemma 2.14 we get for the conditional expectation that

$$\mathbb{E}\left[\frac{\mathbb{I}\{X_1 \in A_k(x_0,\omega)\}}{\sum_{i=1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0,\omega)\}} \,\Big|\, X_1\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\frac{\mathbb{I}\{X_1 \in A_k(x_0,\omega)\}}{1+\sum_{i=2}^{r_n}\mathbb{I}\{X_i \in A_k(x_0,\omega)\}} \;\Big|\; X_1\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{I}\{X_1 \in A_k(x_0,\omega)\}}{1+\sum_{i=2}^{r_n}\mathbb{I}\{X_i \in A_k(x_0,\omega)\}} \;\Big|\; X_1,\omega\right] \Big|\; X_1\right] \\
&= \mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega)\}\mathbb{E}\left[\frac{1}{1+\sum_{i=2}^{r_n}\mathbb{I}\{X_i \in A_k(x_0,\omega)\}} \;\Big|\; \omega\right] \Big|\; X_1\right] \\
&= \mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega)\}\frac{1}{r_n p_{x_0}(\omega)}\lambda(r_n, p_{x_0}(\omega)) \;\Big|\; X_1\right]
\end{aligned}
$$

and thus

$$
h_{n,1}^{(\varepsilon)}(Z_1) = \frac{\varepsilon_1}{r_n}\mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega)\}p_{x_0}(\omega)^{-1}\lambda(r_n, p_{x_0}(\omega)) \mid X_1\right],
$$

which yields the first claim. Since $h_{n,1}^{(\varepsilon)}$ is centered, we get

$$
\begin{aligned}
\zeta_{1,\omega}^n(x_0) &= \mathrm{Var}\left(h_{n,1}^{(\varepsilon)}(Z_1)\right) \\
&= \frac{\sigma^2}{r_n^2}\mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega)\}p_{x_0}(\omega)^{-1}\lambda(r_n, p_{x_0}(\omega)) \mid X_1\right]^2\right] \\
&= \frac{\sigma^2}{r_n^2}\mathbb{E}\left[\frac{\mathbb{I}\{X_1 \in A_k(x_0,\omega_1)\cap A_k(x_0,\omega_2)\}}{p_{x_0}(\omega_1)p_{x_0}(\omega_2)}\lambda(r_n, p_{x_0}(\omega_1))\lambda(r_n, p_{x_0}(\omega_2))\right],
\end{aligned}
$$

and

$$
\zeta_{1,\omega}^n(x_0) \le \frac{\sigma^2}{r_n^2}\mathbb{E}\left[\frac{\mathbb{I}\{X_1 \in A_k(x_0,\omega_1)\cap A_k(x_0,\omega_2)\}}{p_{x_0}(\omega_1)p_{x_0}(\omega_2)}\right] = \frac{\sigma^2}{r_n^2}\Psi_k(x_0)
$$

because $\lambda(h,\xi) \le 1$ and

$$
\begin{aligned}
\zeta_{1,\omega}^n(x_0) &\ge \frac{\sigma^2}{r_n^2}\mathbb{E}\left[\frac{\mathbb{I}\{X_1 \in A_k(x_0,\omega_1)\cap A_k(x_0,\omega_2)\}}{p_{x_0}(\omega_1)p_{x_0}(\omega_2)}\right]\lambda(r_n, c_X 2^{-k})^2 \\
&= \frac{\sigma^2}{r_n^2}\Psi_k(x_0)\lambda(r_n, c_X 2^{-k})^2
\end{aligned}
$$

because $p_{x_0}(\omega) \ge c_X 2^{-k}$. □

Chapter 5

Confidence bands for centered purely random forests

Theorem 4.4 from the previous chapter allows for the construction of pointwise confidence intervals for the regression function m at a fixed point x_0. This chapter is dedicated to confidence bands, which are a uniform generalization of confidence intervals. We will construct confidence bands based on the complete U-statistic version $U_{n,r_n,\omega}^{(\mathrm{RF})}$ of the CPRF from Definition 3.2. While the primary focus is on the complete version of the U-statistic, the extension of the results to the incomplete version $U_{n,r_n,N,\omega}^{(\mathrm{RF})}$ from (3.5) is feasible, see Corollary 5.5. The results leading to the main confidence band theorem further imply a uniform consistency result for centered purely random forests and slightly different versions of the pointwise results from the previous chapter.

These main results are gathered in Section 5.1. In Section 5.2, we will additionally state a confidence band result for the kernel random forest version of the CPRF from Section 3.4. The proof technique employed for the latter result differs slightly from the technique utilized for the random forest result, yet it exploits the same auxiliary results. As a byproduct of our proof technique, we will include a confidence band result for the histogram estimator in Section 5.3. This particular result is, to the best of our knowledge, a new addition to the literature. In Section 5.4, the distribution of the supremum of the Gaussian process is analyzed, a critical component in the construction of asymptotic confidence bands. The chapter concludes with an outline of the proof strategy in Section 5.5 and the collection of all proofs in Section 5.6.

As previously mentioned in the introduction, Theorem 2.4 by Chernozhukov et al. (2014b) is a fundamental component of our proof. It is noteworthy that there exists additional literature addressing the approximation of the supremum of an empirical process, for instance, more recently by Giessing (2023). For the proofs in our work, the result by Chernozhukov et al. (2014b) is sufficient.

In a connected article, Chernozhukov et al. (2014a) illustrate how the approximation by a Gaussian process can be employed to prove confidence bands for general nonparametric density estimators, even in the case where the supremum of the empirical process has no limit distribution. The results on asymptotic confidence bands presented in this chapter will also not be based on an explicit limit distribution.

In the introduction we already mentioned that a more classical proof method for confidence bands exists. This method utilizes a uniform approximation of the empirical process by a Gaussian process, instead of the approximation of the supremum that is sufficient for constructing confidence bands. This classical proof method will be covered later, in Section 7.2, alongside an explanation of its shortcomings in comparison to the proof of this chapter. In particular, it will be demonstrated that the uniform approximation of the entire empirical process with the current results in the literature has a larger error than the approximation of the single supremum.

5.1 Main results

In this section we will state the main result that allows us to construct confidence bands for the regression function. We still consider the regression setting from (2.1) with the assumption on the density from (4.1) and a Hölder continuous m. We use the definitions from (4.3), (4.4), (4.8) and (4.9). Let us define the function class

$$\mathcal{F}_k := \{f_{x_0,k}(x,s) \mid x_0 \in [0,1]^p\}, \tag{5.1}$$

where

$$f_{x_0,k}(x,s) := \sigma^{-1}\Psi_k^{-1/2}(x_0)sK_k(x_0,x), \tag{5.2}$$

with

$$\Psi_k(x_0) = \mathbb{E}[K_k(x_0,X_1)^2]$$

from (4.9) and K_k from (4.8). In Section 5.4 we will explain that $|\mathcal{F}_k| = N_f(k)$ for $N_f(k)$ from Section 3.3.1.3. Further let $\mathcal{H}(\alpha, C_H)$ denote the set of all α-Hölder functions with Hölder constant less or equal to C_H.

Theorem 5.1 (Asymptotic confidence band for CPRF)**.** *Consider any centered purely random forest with at most $N_f(k)$ undividable sets defined in Section 3.3.1.3. Let $\alpha \in (0,1]$ and for some $c < 1$ assume that*

$$r_n/n \leq c \qquad \text{and} \tag{5.3}$$

$$\mathbb{E}\left[\mathfrak{d}(A_k(x,\omega))^\alpha\right]^2 \frac{n(\log n)^2}{2^{2k}\mathcal{V}_{\cap,k}} \to 0. \tag{5.4}$$

Let $q_R \in 2\mathbb{N}$ and $4 \leq q_G \in \mathbb{N}$ be such that

$$N_f(k)^{2/q_R}\frac{(\log n)^2}{r_n\mathcal{V}_{\cap,k}} \to 0, \tag{5.5}$$

$$n^{2/q_G}\frac{(\log n)^5}{n\mathcal{V}_{\cap,k}} \to 0, \tag{5.6}$$

$$\mathbb{E}\left[|\varepsilon_1|^{\max\{q_R,q_G\}}\right] < \infty. \tag{5.7}$$

Let $\mathbf{S}_k$ be a sequence of random variables with $\mathbf{S}_k \overset{d}{=} \sup_{x_0\in[0,1]^p}|B_k f_{x_0,k}|$, where B_k is a sequence of centered Gaussian processes indexed by $\mathcal{F}_k$ and with covariance function

$$\mathrm{Cov}(B_k(f_{x_1,k}), B_k(f_{x_2,k})) = \Psi_k^{-1/2}(x_1)\Psi_k^{-1/2}(x_2)\mathbb{E}\left[K_k(x_1,X_1)K_k(x_2,X_1)\right]. \tag{5.8}$$

Let $\hat{\sigma}$ be an estimator of σ with

$$\mathbb{P}\left(|\hat{\sigma}^2 - \sigma^2| > (\log n)^{-2}\right) \to 0.$$

For $c_k(\beta) = F_{\mathbf{S}_k}^{-1}(1-\beta) := \inf\{\xi \in \mathbb{R} : \mathbb{P}(\mathbf{S}_k \leq \xi) \geq 1-\beta\}$ denote

$$\mathcal{C}_n(x) = \left[U_{n,r_n,\omega}^{(\mathrm{RF})}(x) - \hat{\sigma} c_k(\beta)\sqrt{\frac{\Psi_k(x)}{n}}, U_{n,r_n,\omega}^{(\mathrm{RF})}(x) + \hat{\sigma} c_k(\beta)\sqrt{\frac{\Psi_k(x)}{n}}\right]. \tag{5.9}$$

For $C_H > 0$ it holds that

$$\liminf_{n\to\infty} \inf_{m \in \mathcal{H}(\alpha, C_H)} \mathbb{P}(m(x) \in \mathcal{C}_n(x),\ \forall x \in [0,1]^p) \geq 1-\beta.$$

The result provides confidence bands for m, but like the confidence intervals in Chapter 4, they depend on the function Ψ_k. We recall that we can approximate Ψ_k by Monte Carlo simulation, if the distribution of X is known, otherwise one could use a bootstrap to estimate the function. The distribution of ω is more or less known from the partitioning algorithm and can therefore be used in this approximation. The fact that Ψ_k is not constant means that the confidence bands do not have a constant diameter. The main effect we will see is a smaller than average radius of the confidence band in regions of the feature space where f_X is relatively large. This can be seen from the definitions of $K_k(x_0, x)$ and $\Psi_k(x_0)$ and the observation that p_{x_0} is strongly related to f_X in the region around x_0. Note that the result holds uniformly for m in the class $\mathcal{H}(\alpha, C_H)$. This satisfies the definition of asymptotically honest confidence regions by Li (1989).

A noteworthy aspect of the result is that it does not utilize quantiles of a limit distribution. Instead, the quantiles $c_k(\beta)$ are those of $\mathbf{S}_k$, the supremum of the Gaussian process B_k, and thus depend on k. It is common in the literature for confidence bands to be based on a limit distribution, usually an extreme value distribution, see e.g., Bickel and Rosenblatt (1973). The structure of our result has two advantages. First, we do not need to know the limit distribution of $\mathbf{S}_k$. In particular, it is not even necessary that a limit distribution exists. The second advantage is that using a limit distribution adds another error term. The application of the Gaussian approximation in Theorem 2.4 will show that the distribution of $\mathbf{S}_k$ is close to the distribution of the leading term of the uniform stochastic error $\|U_{n,r_n,\omega}^{(\varepsilon)}\|_\infty$, see (3.9), when appropriately standardized. However, a potential limit distribution of $\mathbf{S}_k$ may be less close to this term, depending on the convergence rate of $\mathbf{S}_k$.

Before discussing the assumptions in the theorem, various effects on the radius of the confidence bands, and the estimation of σ, we state two corollaries, for the two types of CPRFs we introduced in Section 3.3.

Corollary 5.2 (Asymptotic confidence band for the Ehrenfest CPRF)**.** *For the Ehrenfest CPRF let $\alpha \in (0, 1]$ and assume that (5.3) holds and that*

$$\frac{n(\log n)^2}{2^{k(1+2\alpha/p)}} \to 0.$$

Further let $q_R \in 2\mathbb{N}$ and $4 \leq q_G \in \mathbb{N}$ and be such that

$$2^{k(1+2/q_R)}\frac{(\log n)^2}{r_n} \to 0,$$

$$n^{2/q_G}\frac{2^k(\log n)^5}{n} \to 0,$$
$$\mathbb{E}\left[|\varepsilon_1|^{\max\{q_R,q_G\}}\right] < \infty.$$

Let $\hat{\sigma}$, $c_k(\beta)$ and $\mathcal{C}_n$ be as is in Theorem 5.1 for ω distributed according to the Ehrenfest partition. For $C_H > 0$ it holds that

$$\liminf_{n\to\infty} \inf_{m\in\mathcal{H}(\alpha,C_H)} \mathbb{P}(m(x) \in \mathcal{C}_n(x),\, \forall x \in [0,1]^p) \geq 1-\beta.$$

The corollary for the Ehrenfest CPRF follows directly from Theorem 5.1 by noting that $\mathcal{V}_{\cap,k} = \Theta(2^{-k})$, $N_f(k) \lesssim 2^k$ and using $\mathfrak{d}(A_k(x,\omega))^\alpha \lesssim 2^{-\alpha k/p}$ almost surely. If q_R and q_G are sufficiently large, the assumptions are satisfied, for instance, by $k = \lfloor \log_2((\log n)^3 n^{\frac{p}{p+2\alpha}})\rfloor$ and $r_n = \lfloor cn \rfloor$, if $c \in (0,1)$. The choice of k is similar to that in Corollary 4.6, but $r_n = \Theta(n)$ is eligible here. This parameter choice implies that the rate of the radius is

$$\begin{aligned} c_k(\beta)\sqrt{2^k/n} &= \mathcal{O}\left(\sqrt{\log_2\left((\log n)^3 n^{\frac{p}{p+2\alpha}}\right)(\log n)^3 n^{-\frac{2\alpha}{p+2\alpha}}}\right) \\ &= \mathcal{O}\left(\sqrt{(\log n)^4 n^{-\frac{2\alpha}{p+2\alpha}}}\right), \end{aligned}$$

because $c_k(\beta) = \mathcal{O}(\sqrt{k})$, as mentioned later in (5.43). Compared to the asymptotic standard deviation in Corollary 4.6, this is larger in the logarithmic term. The corollary for the uniform CPRF below requires slightly stronger assumptions, similar to the pointwise case.

Corollary 5.3 (Asymptotic confidence band for the uniform CPRF)**.** *For the uniform CPRF let $\alpha \in (0,1]$ and assume that (5.3) holds and that*

$$\left(\frac{p-1+2^{-\alpha}}{p}\right)^{2k}\frac{n(\log n)^2 k^{(p-1)}}{2^k} \to 0.$$

Further let $q_R \in 2\mathbb{N}$ and $4 \leq q_G \in \mathbb{N}$ be such that

$$2^{2kp/q_R}\frac{2^k k^{(p-1)}(\log n)^2}{r_n} \to 0,$$
$$n^{2/q_G}\frac{2^k k^{(p-1)}(\log n)^5}{n} \to 0,$$
$$\mathbb{E}\left[|\varepsilon_1|^{\max\{q_R,q_G\}}\right] < \infty.$$

Let $\hat{\sigma}$, $c_k(\beta)$ and $\mathcal{C}_n$ be as is in Theorem 5.1 for ω distributed according to the uniform partition. For $C_H > 0$ it holds that

$$\liminf_{n\to\infty} \inf_{m\in\mathcal{H}(\alpha,C_H)} \mathbb{P}(m(x) \in \mathcal{C}_n(x),\, \forall x \in [0,1]^p) \geq 1-\beta.$$

Corollary 5.3 follows directly from Theorem 5.1 by noting that $N_f(k) \leq 2^{kp}$ and using the bounds in (3.23) and (3.25) which imply

$$\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha]^2 = \mathcal{O}\left(\left(\frac{p-1+2^{-\alpha}}{p}\right)^{2k}\right)$$

and $\mathcal{V}_{\cap,k} \gtrsim 2^{-k}k^{-(p-1)}$. If q_R and q_G are sufficiently large, the assumptions are again satisfied by a parameter choice similar to the pointwise CLT in Corollary 4.7. For $c \in (0,1)$, we choose $r_n = \lfloor cn \rfloor$ and $k = \lfloor \log_2((n(\log_2 n)^{p+2})^{1-\nu}) \rfloor$ for ν as in Corollary 4.2 with $0 < \nu \leq \frac{2\alpha}{2\alpha+p}$. We note that $r_n = \Theta(n)$ is eligible again. This parameter choice, $c_k(\beta) = \mathcal{O}(\sqrt{k})$ and (3.25) imply that the rate of the radius satisfies

$$\begin{aligned}
&c_k(\beta)\sqrt{\frac{\Psi_k(x)}{n}} \\
&\quad = \mathcal{O}\left(\sqrt{\frac{k2^{2k}\mathcal{V}_{\cap,k}}{n}}\right) \\
&\quad = \mathcal{O}\left(\sqrt{\frac{2^k k^{-(p-3)/2}}{n}}\right) \\
&\quad = \mathcal{O}\left(\sqrt{((\log_2 n)^{p+2})^{1-\nu}\log_2\left((n(\log_2 n)^{p+2})^{1-\nu}\right)^{-(p-3)/2}n^{-\nu}}\right) \\
&\quad = \mathcal{O}\left(\sqrt{(\log_2 n)^{-\nu(p+2)+(p+7)/2}n^{-\nu}}\right),
\end{aligned}$$

which is again larger as in the pointwise case in Corollary 4.7. We note that the bound for ν implies that the rate is worse than that of the Ehrenfest CPRF.

We continue with the discussion of the assumptions in Theorem 5.1. The two corollaries above allow us to consider the assumptions for the two specific CPRF types.

Assumption (5.3) This assumption is not a major restriction at all, and it is important to note that it allows $r_n = cn$, which is not possible with most asymptotic results from the literature. This is the maximum possible rate for r_n, since it cannot be greater than n. In the context of data dependent random forests, it is common to employ bootstrap subsamples of size $r_n = n$. The bootstrap subsampling implies that the number of distinct observations in the subsample is less than n. Nevertheless, the expectation of the number of distinct observations remains constant with respect to n. Therefore, the case $r_n = cn$ is highly relevant for random forests.

Contrary to our assumptions, in the literature it is sometimes assumed that r_n is not too large compared to n in order to prove asymptotic normality. See for example Mentch and Hooker (2016, Theorem 1). This is done to ensure that the first order terms, i.e. the Hájek projection, dominate in the Hoeffding-decomposition of the U-statistic. In Theorem 2.13, the assumption

$$\frac{r_n}{n}\frac{\zeta^n_{r_n}}{r_n\zeta^n_{1,\omega}} \to 0$$

captures this necessary dominance. In this chapter we use a different proof technique than in the pointwise case to show that the Hájek projection (or a similar term) is the leading term. Our more direct proof technique allows us to weaken the assumptions on the tuning parameter r_n with respect to n.

Assumption (5.4) This is an undersmoothing assumption similar to (4.13) in the pointwise case. The only distinction between the two is the presence of a logarithmic term. It

is important to note that this assumption involves a single x, which is sufficient to control the approximation error uniformly. In the proof of the bound for the approximation error in Section 5.6.10.2 we will exploit that

$$\frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \mathfrak{d}(A_k(x_0, \omega_I))^\alpha \approx \mathbb{E}[\mathfrak{d}(A_k(x_0, \omega))^\alpha]. \tag{5.10}$$

The term on the left hand side is an upper bound for a relevant part of the approximation error. We will utilize that its difference to the expectation on the right hand side of (5.10) is negligible. The tree construction of the CPRF implies that $\mathfrak{d}(A_k(x_1, \omega)) \stackrel{d}{=} \mathfrak{d}(A_k(x_2, \omega))$ for all $x_1, x_2 \in [0,1]^p$. Thus,

$$\mathbb{E}\left[\sup_{x_0 \in [0,1]^p} \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \mathfrak{d}(A_k(x_0, \omega_I))^\alpha\right] \approx \sup_{x_0 \in [0,1]^p} \mathbb{E}[\mathfrak{d}(A_k(x_0, \omega))^\alpha] = \mathbb{E}[\mathfrak{d}(A_k(x_0, \omega))^\alpha]$$

will be the relevant term for this part of the approximation error. This explains why the condition does not depend on the choice of x. In comparison, a single regression tree estimator lacks the averaging in I and thus, its approximation error bound involves

$$\mathbb{E}\left[\sup_{x_0 \in [0,1]^p} \mathfrak{d}(A_k(x_0, \omega))^\alpha\right].$$

This illustrates how the random forest profits from the averaging of many trees and why its uniform approximation error is smaller than that of a single regression tree with n observations.

If the diameter has the optimal rate $2^{-k/p}$ and $\mathcal{V}_{\cap,k} = \Theta(2^{-k})$, which is the case for the Ehrenfest forest, the assumption simplifies to

$$\mathbb{E}[\mathfrak{d}(A_k(x, \omega))^\alpha]^2 \frac{n(\log n)^2}{2^{2k}\mathcal{V}_{\cap,k}} \lesssim n(\log n)^2 2^{-k(1+2\alpha/p)} \to 0.$$

For the uniform CPRF the exponent of 2^{-k} is even smaller, it is effectively

$$1 + \log_2\left(\frac{p - 1 + 2^{-\alpha}}{p}\right)$$

instead of $(1 + 2\alpha/p)$ and additionally the logarithmic terms are larger.

This is the only assumption directly depending on p as p affects the diameter and thereby the bias. To illustrate this interaction, we focus on the Ehrenfest case. For a finite sample size n let us assume that for some p_1 we have that

$$\frac{n(\log n)^2}{2^{k_1(1+2\alpha/p_1)}} = \kappa$$

for some $\kappa > 0$. If we have a $p_2 > p_1$ and would want the fraction to be equal to the same κ we need to choose k_2 accordingly. We need that

$$2^{k_1(1+2\alpha/p_1)} = 2^{k_2(1+2\alpha/p_2)}$$

$$\Leftrightarrow k_1(1 + 2\alpha/p_1) = k_2(1 + 2\alpha/p_2)$$
$$\Leftrightarrow k_2 = k_1 \frac{p_1p_2 + 2p_2\alpha}{p_1p_2 + 2p_1\alpha}.$$

If we have an appropriate k for $p_1 = 2$ and we want to find a k that meets the assumption in the same way for $p_2 = 4$ we need to increase k by a factor

$$\frac{p_1p_2 + 2p_2\alpha}{p_1p_2 + 2p_1\alpha} = \frac{8 + 8\alpha}{8 + 4\alpha},$$

which is equal to $4/3$ for $\alpha = 1$.

Assumption (5.5) The reason for this assumption is the necessity to bound the uniform difference of $U^{(\varepsilon)}_{n,r_n,\omega}$ to its Hájek projection. In other words, it is required to ensure the dominance of the first order terms in the Hoeffding-decomposition. These first order terms can be handled by empirical process theory. Using a higher order approximation for the distribution of $U^{(\varepsilon)}_{n,r_n,\omega}$ might circumvent this assumption for confidence band construction, but it requires a more sophisticated proof technique, possibly involving U-processes.

Without the logarithmic terms, assumption (5.5) means

$$N_f(k)^{2/q_R} \frac{1}{r_n \mathcal{V}_{\cap,k}} \to 0. \tag{5.11}$$

Note that the assumption implicitly depends on p via $N_f(k)$. Since $N_f(k)$ increases with p, we need higher moments of the errors for greater p. This is similar to Chao et al. (2017, Theorem 2), where b_1 controls the tails of the errors and needs to be greater for greater p. If we can choose q_R large enough, (5.11) roughly means that $r_n\mathcal{V}_{\cap,k} \to \infty$ is required, which implies

$$2^k/r_n \leq \mathcal{V}^{-1}_{\cap,k} r_n^{-1} \to 0. \tag{5.12}$$

In particular, this means that a single tree is consistent. Consistency of a single tree trivially implies consistency of the random forest. It is not clear whether the considered forest is consistent with inconsistent trees. Nonetheless, in the case where both are consistent, the random forest has the faster convergence rate, as already discussed in Section 4.1.

We note that $2^k = \mathcal{O}(r_n)$ is necessary anyway, because otherwise the number of empty cells in a tree partition increases in n. This means that the only relevant case that contradicts (5.12) is $2^k = \Theta(r_n)$. Recalling the assumption $r_n \leq cn$, this allows for most reasonable choices for r_n.

Results in the case $2^k = \Theta(r_n)$ would be desirable for different types of random forests, since their construction often leads to a constant number of observations in each cell. Transferred to the purely random forest case, this corresponds to $2^k = \Theta(r_n)$. The discussion above indicates that for results in this regime, it may be necessary to use more than the first order terms in the Hoeffding-decomposition, since they are not necessarily dominant in this case. We have already explained why this is more difficult for asymptotic confidence bands, but for variance estimation of incomplete U-statistics, higher order terms have already been used by Xu et al. (2024), who also applied their results to random forests.

For purely random forests, we cannot control the number of observations in the cells directly, but only through the tuning parameters k and r_n. Therefore, an assumption like (5.5) is necessary. We will briefly explain how it ensures non-empty cells. Notice that $2^k \leq \mathcal{V}_{\cap,k}^{-1}$ and $2^k \leq N_f(k)$ directly implies

$$\frac{2^{k(1+2/q_R)}(\log n)^2}{r_n} \leq N_f(k)^{2/q_R} \frac{(\log n)^2}{r_n \mathcal{V}_{\cap,k}} \to 0.$$

Hence

$$\log(n(\log n)^2 2^{2k}) \lesssim 2^{2k/q_R}(\log n)^2 = o(r_n/2^k),$$

which ensures that

$$n(\log n)^2 \mathcal{V}_{\cap,k}^{-1}(1 - c_X 2^{-k})^{2r_n} \lesssim n(\log n)^2 2^{2k} \exp(-2c_X r_n/2^k) \to 0. \tag{5.13}$$

The term $(1 - c_X 2^{-k})^{2r_n}$ corresponds to the probability that a cell is empty, and the above convergence is necessary so that the union bound of these probabilities is negligible compared to the asymptotic variance.

Assumption (5.6) Omitting the logarithms, assumption (5.6) simplifies to

$$\mathcal{V}_{\cap,k} n^{1-2/q_G} \to \infty.$$

For both CPRFs, $\mathcal{V}_{\cap,k}$ has a rate 2^{-k}, at least up to terms polynomial in k. Note that $\mathcal{V}_{\cap,k} \leq 2^{-k}$ requires that n grows at a rate exceeding 2^k. We recall that $2^{-k/p}$ is comparable to a bandwidth for kernel estimators, which illustrates that (5.6) is similar to standard assumptions. The exact rate by which n must exceed 2^k depends mainly on the existing moments captured by q_G and further on the exact rate of $\mathcal{V}_{\cap,k}$ and the logarithmic terms from the assumption.

Assumption (5.7) The last assumption for the noise is fulfilled for all sub-Gaussian distributions and therefore includes multiple classes of distributions. It is related to the previous two assumptions. Depending on the existing moments of ε_1, these assumptions control which regimes of tuning parameters can be used, or vice versa, which moments have to exist in these regimes.

The radius Finally, we discuss the radius of the confidence bands. Besides Ψ_k, it is determined by n, σ, and $c_k(\beta)$. The dependence on σ is rather intuitive, as is the typical dependence on n, which is proportional to $n^{-1/2}$. The bounds in (4.20) yield $\Psi_k(x) = \Theta(2^{2k}\mathcal{V}_{\cap,k})$, which has rate 2^k for the Ehrenfest forest. Hence, for the Ehrenfest forest the radius has rate $c_k(\beta)\sqrt{2^k/n}$ for a fixed p. In (5.43) we will observe that $c_k(\beta) = \mathcal{O}(\sqrt{k})$. For the uniform CPRF, the rate of $\mathcal{V}_{\cap,k}$ is smaller by a term polynomial in k. Nonetheless, we cannot make a clear statement about the relation between the two radii, because the quantiles $c_k(\beta)$ for the Ehrenfest and uniform CPRF are not the same. In the simulation study in Chapter 6 we will observe that the uniform CPRF leads to smaller radii, which is an advantage.

Throughout our work, we assume that p is fixed. Nevertheless, we want to discuss how the radii behave for different p. If k is fixed, a greater p implies a smaller $\mathcal{V}_{\cap,k}$ and thus a smaller radius of the confidence bands. In the calculations for the Ehrenfest CPRF, this effect is captured by the lower bound

$$\mathcal{V}_{\cap,k} \geq 2^{-k} 2^{-p(p+1)((\Delta-B)+pB)}.$$

For the uniform CPRF, the bound from (3.25) shows a similar effect. The reason is that there are more possibilities for the two partitions to differ, leading to a smaller intersection.

Earlier we argued that one would need to increase k for a larger p to satisfy the assumptions with a similar quality in a finite sample case. This increase in k then increases the radius of the confidence bands. Because of the overlap of the two effects, we cannot make a clear statement about the overall effect on the radii.

We have already mentioned that the radius of the confidence bands is directly proportional to the variance of the errors. This leads us to the required estimation of σ when its value is unknown. However, if the true variance, denoted by σ^2, is known, then Theorem 5.1 holds with $\hat{\sigma}$ replaced by σ, and no estimator is required. One method of estimating the variance is to use the residuals of an estimator. The following lemma demonstrates that a variance estimator based on the residuals of the two random forest variants satisfies the assumptions of Theorem 5.1.

Lemma 5.4. *Assume that* $\mathbb{E}[\varepsilon_1^4] < \infty$ *and let*

$$\hat{\varepsilon}_i := Y_i - U^{(\mathrm{RF})}_{n,r_n,\omega}(X_i)$$
$$\text{and } \hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

There exists a constant C *independent of* n, k *and* r_n *such that*

$$\mathbb{P}(|\hat{\sigma}^2 - \sigma^2| > \kappa) \leq \frac{C}{\kappa}\Big(\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha]^2 + N_f(k)\binom{n}{r_n}^{-1} + (1 - c_X 2^{-k})^{r_n} + \frac{2^k}{n} + \frac{2^{2k}}{n^2} + n^{-1/2}\Big). \tag{5.14}$$

To fulfill the assumption on $\hat{\sigma}$ in Theorem 5.1 we need that the right hand side of (5.14) converges to zero for $\kappa = (\log n)^{-2}$. For the uniform CPRF (3.23) shows that $\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha] = \mathcal{O}(c_{p,\alpha}^k)$ for a constant $c_{p,\alpha} < 1$. Under the assumptions of Theorem 5.1 we have

$$(\log n)^2 c_{p,\alpha}^k \to 0.$$

For all the other terms from (5.14) we can argue similarly under the assumptions of Theorem 5.1.

In applying the confidence bands in practice, it is not necessary to use a variance estimator based on the residuals, despite the fact that it fulfills the assumption from Theorem 5.1. A noteworthy alternative are estimators that are not based on an estimator of m, see e.g. Müller et al. (2003) or Shen et al. (2020). The primary advantage of these estimators is that their quality is not tied to the quality of the estimator of the regression function. An estimator of this type will be utilized in the simulation study in Chapter 6.

So far in this chapter we have considered the complete generalized U-statistic. The following corollary applies to the incomplete generalized U-statistic.

Corollary 5.5. *In addition to the assumptions of Theorem 5.1 assume that*

$$N_f^2(k)\frac{n(\log n)^2}{N2^{2k}\mathcal{V}_{\cap,k}} \to 0. \tag{5.15}$$

Consider $U^{(\mathrm{RF})}_{n,r_n,N,\omega}(x)$ *from equation (2.11) on the event* $\{\hat{N} > 0\}$*. For*

$$\mathcal{C}_{n,N}(x) = \left[U^{(\mathrm{RF})}_{n,r_n,N,\omega}(x) - \hat{\sigma}c_k(\beta)\sqrt{\frac{\Psi_k(x)}{n}}, U^{(\mathrm{RF})}_{n,r_n,N,\omega}(x) + \hat{\sigma}c_k(\beta)\sqrt{\frac{\Psi_k(x)}{n}}\right]$$

it holds that

$$\liminf_{n\to\infty} \inf_{m\in\mathcal{H}(\alpha,C_H)} \mathbb{P}\left(m(x) \in \mathcal{C}_{n,N}(x),\, \forall x \in [0,1]^p\right) \geq 1-\beta.$$

We only consider the estimator on the event $\{\hat{N} > 0\}$ because otherwise we would use an empty random forest. The corollary shows that the incomplete version of the U-statistic can still be used to construct confidence bands provided that N is large enough. The assumption (5.15) on N is similar to its counterpart (4.11) in Theorem 4.4. The additional $N_f^2(k)(\log n)^2$ leads to a stronger assumption, which is necessary because the result is uniform. The term $N_f^2(k)$ is a consequence of the utilized union bound, and $(\log n)^2$ is required because the uniform error is larger than the pointwise error by a logarithm. For the Ehrenfest forest with independent covariates, the condition simplifies to

$$N_f^2(k)\frac{n(\log n)^2}{2^{2k}\mathcal{V}_{\cap,k}}\frac{1}{N} \lesssim \frac{2^k n(\log n)^2}{N} \to 0.$$

In practice, it may not be necessary to select N as large as the condition demands. The simulations in Chapter 6 will suggest that a smaller N is sufficient in practice.

The collection of auxiliary results leading to Theorem 5.1 implies some more results not directly related to confidence bands. For a fixed x_0, we obtain a different bound for the mean squared error and asymptotic normality under different assumptions than in Chapter 4. Furthermore, the results used for the confidence bands yield uniform convergence of the random forest estimator. We present the two pointwise results first, followed by the uniform convergence result.

Corollary 5.6. *Consider a CPRF from Definition 3.2 and suppose* $0 < c_X \leq f_X \leq C_X$*,* $\mathbb{E}[\varepsilon_1^2] < \infty$*,* $2^k \leq r_n < n$ *and* m *is Hölder continuous of order* $\alpha \in (0,1]$*. It then holds that*

$$\begin{aligned}\mathbb{E}\left[\left(U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0) - m(x_0)\right)^2\right] &= \mathcal{O}\left(\mathbb{E}\left[\mathfrak{d}(A_k(x_0,\omega))^\alpha\right]^2\right) + \mathcal{O}((1-c_X2^{-k})^{r_n}) \\ &\quad + \mathcal{O}\left(\frac{2^{2k}\mathcal{V}_{\cap,k}}{n}\right) + \mathcal{O}\left(\frac{2^{2k}}{r_n n}\right) + \mathcal{O}\left(\frac{2^k}{r_n}\left(\frac{r_n}{n}\right)^{r_n}\right). \end{aligned} \tag{5.16}$$

The result above is similar to Proposition 4.1. We additionally assume that $2^k \leq r_n$ which is a necessary condition for convergence anyway. The first two terms on the right hand side of (5.16), which correspond to the approximation error, are the same. The main difference is a different bound for the stochastic error that is captured in the last three terms. Compared to Proposition 4.1 the last two terms replace an $\mathcal{O}(2^k r_n/n^2)$ term. In the proof of Proposition 4.1, the bound for the stochastic error in (4.23) was

$$\operatorname{Var}\big(U^{(\varepsilon)}_{n,r_n,\omega}(x_0)\big) \leq \frac{r_n^2}{n}\zeta^n_{1,\omega}(x_0) + \frac{r_n^2}{n^2}\zeta^n_{r_n}(x_0) = \mathcal{O}\left(2^{2k}\mathcal{V}_{\cap,k}/n\right) + \mathcal{O}\left(2^k r_n/n^2\right). \tag{5.17}$$

The first term is the same in (5.16) and corresponds to the first order term of the Hoeffding-decomposition or the Hájek projection, respectively. The second term is a bound for the variances of the higher order terms from the Hoeffding-decomposition. In this chapter we use a more direct bound the difference between the stochastic error and its Hájek projection. We call their difference the projection error, which is further decomposed into two remainder terms $R^{(1)}_{n,r_n,\omega}(x_0)$ and $R^{(2)}_{n,r_n,\omega}(x_0)$. Lemma 5.18 and Lemma 5.19 provide bounds for the variances of these terms and lead to the result above. If $r_n/n < c < 1$, the last term in (4.23) is

$$\mathcal{O}\left(\frac{2^k}{r_n}\left(\frac{r_n}{n}\right)^{r_n}\right) = \mathcal{O}\left(\frac{2^k}{r_n}\exp(r_n \log c)\right),$$

which converges to zero faster than the other terms under appropriate assumptions. Thus, it remains to compare the term $\mathcal{O}(2^{2k}/(r_n n))$ from (4.23) with $\mathcal{O}(2^k r_n/n^2)$ from Proposition 4.1. Their difference is reflected by the difference between r_n/n and $2^k/r_n$ and in general it depends on r_n, which term is smaller. However, in the case $r_n = \Theta(n)$ the new result yields the better rate. In particular, it gives a rate for the stochastic error, that is better than the stochastic error rate of a single regression tree in this case, which was not the case in Proposition 4.1. When applied to the two specific types of CPRFs, the same choices for k as in Corollary 4.2 can be employed, but unlike there $r_n = \Theta(n)$ is a permissible option in both cases. We proceed with the central limit theorem.

Corollary 5.7. *For any fixed x_0 assume that $r_n/n \leq c$ for some $c < 1$ and $\mathbb{E}[\varepsilon_1^2] < \infty$. If*

$$\frac{1}{\mathcal{V}_{\cap,k} r_n} \to 0, \tag{5.18}$$

$$\frac{2^k}{r_n}\log\left(n 2^{-2k}\mathcal{V}^{-1}_{\cap,k}\right) \to 0, \tag{5.19}$$

$$\mathbb{E}\left[\mathfrak{d}(A_k(x_0,\omega))^\alpha\right]^2 \frac{n}{2^{2k}\mathcal{V}_{\cap,k}} \to 0, \tag{5.20}$$

it holds that

$$\sqrt{\frac{n}{\sigma^2\Psi_k(x_0)}}\left(U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0) - m(x_0)\right) \xrightarrow{d} \mathcal{N}(0,1).$$

Under the additional assumption

$$\frac{n}{N 2^{2k}\mathcal{V}_{\cap,k}} \to 0, \tag{5.21}$$

it holds that

$$\sqrt{\frac{n}{\sigma^2 \Psi_k(x_0)}} \left(U^{(\mathrm{RF})}_{n,r_n,N,\omega}(x_0) - m(x_0) \right) \xrightarrow{d} \mathcal{N}(0,1).$$

As was the case with the preceding corollary, the present result is compared with its counterpart, Theorem 4.4. The emphasis of the comparison is placed on the second assertion, as it relates to the incomplete U-statistic case. The assumptions (5.19), (5.20) and (5.21) are the same as in Theorem 4.4. Previously, we required $\mathbb{E}\left[\varepsilon_1^6\right] < \infty$, which was a consequence of the assumption on the U-statistic kernel in Theorem 2.13. Thus, the assumption on the moments of the errors is weaker here. In Theorem 4.4 we additionally assumed that

$$\frac{r_n}{n 2^k \mathcal{V}_{\cap,k}} \to 0, \tag{5.22}$$

and here we assume that $r_n/n \leq c < 1$ and $\mathcal{V}_{\cap,k}^{-1} r_n^{-1} \to 0$. The previous assumption implies that

$$\frac{r_n}{n} \leq \frac{r_n}{n 2^k \mathcal{V}_{\cap,k}} \to 0,$$

which is more restrictive than $r_n/n \leq c < 1$. The reason for the difference between (5.18) and (5.22) are the different bounds for the stochastic error employed in the previous chapter and here. In the discussion of Corollary 5.6 above we already explained their differences and the reasons. The bounds in (5.17) and (5.16) differ only for the variance of the higher order terms of the Hoeffding-decomposition. For the central limit theorem, dominance of the first order terms in the Hoeffding-decomposition is required. To ensure this dominance we need (5.22) with the bound in (5.17), as opposed to (5.18) with the new bound. To put (5.18) into perspective, we note that

$$\mathcal{V}_{\cap,k} \gtrsim 2^{-k} k^{-(p-1)},$$

for the uniform CPRF and $\mathcal{V}_{\cap,k} = \Theta(2^{-k})$ for the Ehrenfest CPRF. Thus, it implies

$$\frac{1}{\mathcal{V}_{\cap,k} r_n} \lesssim \frac{2^k k^{(p-1)}}{r_n} \to 0. \tag{5.23}$$

In both versions of the central limit theorem we need that

$$\frac{2^k}{r_n} \log\left(n 2^{-2k} \mathcal{V}_{\cap,k}^{-1}\right) \lesssim \frac{2^k}{r_n} \log\left(n 2^{-k} k^{(p-1)}\right) \to 0,$$

which already is an assumption on the rate of $2^k/r_n$. The new assumption (5.18), which leads to (5.23), is only more restrictive in the terms of order k or $\log n$, respectively. Therefore, one can argue that the collection of assumptions in Corollary 5.7 is less restrictive than that in Theorem 4.4, since $r_n/n < c < 1$ is eligible here. In Remark 4.5 we already discussed that this might be possible, because the main assumption on r_n and n in Theorem 2.13 by Peng et al. (2022) can be weakened. However, this seemed difficult due to the sharp bounds required throughout several proofs. The last result in this section is the uniform convergence result for CPRFs below.

Corollary 5.8 (Uniform convergence of a CPRF)**.** *Consider a centered purely random forest with at most $N_f(k)$ undividable cells. Assume $r_n/n \leq c < 1$ and $r_n/k \to \infty$. Let $\nu > 4$ with $\mathbb{E}[|\varepsilon_1|^\nu] < \infty$. For any $q_R \in \mathbb{N}$ with $\mathbb{E}[|\varepsilon_1|^{q_R}] < \infty$ it holds that*

$$\begin{aligned}
&\|U^{(\mathrm{RF})}_{n,r_n,\omega} - m\|_\infty \\
&\quad = \mathcal{O}_{\mathbb{P}}\left(2^k\left(\frac{(\log n)^{3/2}}{n^{1-1/\nu}} + \frac{\mathcal{V}_{\cap,k}^{1/4}(\log n)^{5/4}}{n^{3/4}} + \frac{\mathcal{V}_{\cap,k}^{1/3}\log n}{n^{2/3}}\right)\right) \\
&\qquad + \mathcal{O}\left(\sqrt{\frac{2^{2k}\mathcal{V}_{\cap,k}k}{n}}\right) + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{2^{2k}}{r_n n}}N_f(k)^{1/q_R}\right) \\
&\qquad + \mathcal{O}_{\mathbb{P}}\left(\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha] + N_f(k)\binom{n}{r_n}^{-1/2}\right) + \mathcal{O}_{\mathbb{P}}(2^k(1-c_X 2^{-k})^{r_n}).
\end{aligned}$$

This result is noteworthy because most convergence results for random forests algorithms in the literature are pointwise. The uniform convergence holds in more regimes than the ones in which we can construct confidence bands. We mainly need that $2^k \to \infty$, $2^k/n \to 0$ and $2^{2k}/(r_n n) = o(N_f(k)^{2/q_R})$ up to logarithmic factors. The latter depends on the moments of the errors. If all moments exist it suffices that $2^k/r_n = \mathcal{O}(1)$ and $2^k/n \to 0$ up to logarithms. The last term in the corollary corresponds to the union bound of the probabilities of empty cells in a regression tree. To ensure that this term converges to zero, it is sufficient that $2^k/r_n = o(k^{-1})$. Remark 5.21 below shows that this is a necessary condition.

Note that all but two of the terms are similar to those in Corollary 5.6. The first term has no counterpart, because it corresponds to the Gaussian approximation of the supremum. The other terms are slightly different, because this is a uniform result. The disparities are consequences of union bounds and the $\sqrt{k}$ in the first term is caused by the supremum of Gaussian random variables.

In the context of data dependent forests, where empty cells are typically not permitted, the last term will not occur and thus $2^k = \Theta(r_n)$ might be sufficient, as we already discussed earlier. This suggests the possibility that a random forest consisting of non-consistent trees can achieve uniform consistency.

5.2 Confidence bands for kernel random forests

In this section we present a confidence band result for the KeRF estimator

$$U^{(\mathrm{KRF})}_{n,r_n,\omega}(x_0) = \frac{\sum_{I\in B_{r_n,n}}\sum_{j\in I} Y_j \mathbb{I}\{X_j \in A_k(x_0,\omega_I)\}}{\sum_{I\in B_{r_n,n}}\sum_{j\in I}\mathbb{I}\{X_j \in A_k(x_0,\omega_I)\}}.$$

from Section 3.4. The different structure of the estimator compared to the random forest will in general lead to a different Gaussian process in the asymptotic distribution of the supremum. Let

$$\check{K}_k(x_0,x) = \mathbb{E}\left[\mathbb{I}\{x \in A_k(x_0,\omega)\}\right] p_{x_0}^{-1}$$

and

$$\Phi_k(x_0) := \mathbb{E}\left[\check{K}_k(x_0, X_1)^2\right] = \mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x, \omega_1) \cap A_k(x, \omega_2)\}\right] p_{x_0}^{-2}. \tag{5.24}$$

Further we define

$$\check{\mathcal{F}}_k := \{\check{f}_{x_0,k}(x, s) \mid x_0 \in [0,1]^p\}$$

for

$$\check{f}_{x_0,k}(x, s) := \sigma^{-1}\Phi_k^{-1/2}(x_0) s \check{K}_k(x_0, x).$$

Theorem 5.9. *Assume that the same conditions as in Theorem 5.1 hold, but with assumption (5.4) and assumption (5.5) replaced by*

$$\mathbb{E}\left[\sup_{x_0 \in [0,1]^p} \max_{I \in B_{r_n,n}} \mathfrak{d}(A_k(x_0, \omega_I))^\alpha\right]^2 \frac{n(\log n)^2}{2^{2k}\mathcal{V}_{\cap,k}} \to 0 \tag{5.25}$$

$$\text{and} \qquad N_f(k)^{2/q_R} \frac{2^k(\log n)^3}{n} \to 0. \tag{5.26}$$

Let $\mathbf{S}_k$ be a sequence of random variables with $\mathbf{S}_k \stackrel{d}{=} \sup_{x_0 \in [0,1]^p} |B_k \check{f}_{x_0,k}|$, where B_k is a sequence of centered Gaussian processes indexed by $\check{\mathcal{F}}_k$ and with covariance function

$$\operatorname{Cov}(B_k(\check{f}_{x_1,k}), B_k(\check{f}_{x_2,k})) = \Phi_k^{-1/2}(x_1)\Phi_k^{-1/2}(x_2)\mathbb{E}\left[\check{K}_k(x_1, X_1)\check{K}_k(x_2, X_1)\right].$$

For $c_k(\beta) = F_{\mathbf{S}_k}^{-1}(1-\beta)$ denote

$$\mathcal{C}_n^K(x) = \left[U_{n,r_n,\omega}^{(KRF)}(x) - \hat{\sigma} c_k(\beta)\sqrt{\frac{\Phi_k(x)}{n}}, U_{n,r_n,\omega}^{(KRF)}(x) + \hat{\sigma} c_k(\beta)\sqrt{\frac{\Phi_k(x)}{n}}\right].$$

It holds that

$$\liminf_{n\to\infty} \inf_{m \in \mathcal{H}(\alpha, C_H)} \mathbb{P}\left(m(x) \in \mathcal{C}_n^K(x),\, \forall x \in [0,1]^p\right) \geq 1 - \beta.$$

Assumption (5.25) that replaces (5.4) is more restrictive, because we need a bound for the expectation of the maximum diameter instead of a single diameter. In the proofs, we will see that the averaging over the ω_I in the weights of the standard random forest reduces the approximation error to the pointwise diameter. Due to the different weight structure discussed in Section 3.4, this is difficult for the KeRF. In particular, the assumption implies that with the result above the confidence bands of the KeRF do not work for the uniform CPRF.

Compared to Theorem 5.1, assumption (5.26) replaces assumption (5.5). The latter is needed to replace the random denominator in the proof strategy of Theorem 5.1. The random denominator for the KeRF is different and outside of both sums. We need the new assumption (5.26) to replace this new denominator with its expectation. This is captured in the term $\delta_n(x_0)$ in the proof. The new assumption is weaker because $2^k \leq \mathcal{V}_{\cap,k}^{-1}$ and $r_n \leq n$. For the term $\delta_n(x_0)$ we do not need to show that it converges to zero faster than the asymptotic standard deviation. It suffices to show that it converges at a logarithmic rate, because the term is multiplied by the other remainder terms. In the proof of Theorem 5.1,

the term $R^{(1)}_{n,r_n,\omega}$ is an additive remainder term and therefore needs the faster convergence to zero.

Further, we observe that the Gaussian processes and the confidence bands differ due to the replacement of Ψ_k by Φ_k in the definition of the function classes. In general, these functions are not the same because of the different estimator structure. In Section 3.4, we noted that Scornet (2016b) showed that the term in (3.34) converges to zero for centered random forests when X is uniformly distributed and the depth k satisfies $k = \Theta((\log_2 n)/3)$. Our results illustrate a similar effect in the asymptotic behavior. The uniform distribution of X_1 implies that

$$p_{x_0}(\omega) = p_{x_0} = 2^{-k}$$

and thus $\Psi_k = \Phi_k$. The equality of the functions means that Gaussian processes in Theorem 5.1 and Theorem 5.9 are the same.

5.3 Confidence bands for the histogram estimator

We want to add a result for the histogram estimator defined in (2.2). The result is not directly a corollary of the results for the random forest but one can use the same proof technique in a less complex case. We note that $\mathbb{V}(A_\delta(x_0)) = \delta^p$ for all x_0. Like a regression tree the estimator is piecewise constant and relies on a partition of the feature space, which leads to the similarity in the proof. We obtain the following confidence band result.

Theorem 5.10. *Let $\alpha \in (0, 1]$, consider a histogram regression estimator and assume that*

$$n\delta^{p+2\alpha}(\log n) \to 0. \tag{5.27}$$

Further let $q_R \in 2\mathbb{N}$ and $4 \leq q_G \in \mathbb{N}$ be such that

$$\frac{\log n}{\delta^{p(1+2/q_R)}n} \to 0 \tag{5.28}$$

$$n^{2/q_G}\frac{(\log n)^5}{\delta^p n} \to 0 \tag{5.29}$$

$$\mathbb{E}\left[|\varepsilon_1|^{\max\{q_R,q_G\}}\right] < \infty. \tag{5.30}$$

Let $(Z_j)_{j\in\mathbb{N}}$ be i.i.d. standard normal distributed random variables. Let $c_\delta(\beta)$ satisfy

$$\mathbb{P}\left(\max_{j=1,\dots,\delta^{-p}} |Z_j| \leq c_\delta(\beta)\right) = 1 - \beta$$

and let $\hat{\sigma}$ be an estimator of σ with

$$\mathbb{P}(|\hat{\sigma}^2 - \sigma^2| > (\log n)^{-2}) \to 0.$$

For $p_{x_0}(\delta) = \mathbb{P}(X_1 \in A_\delta(x_0))$ let

$$\mathcal{C}_n(x) = \left[\hat{m}_H(x) - \hat{\sigma}c_\delta(\beta)(p_x(\delta)n)^{-1/2}, \hat{m}_H(x) + \hat{\sigma}c_\delta(\beta)(p_x(\delta)n)^{-1/2}\right].$$

For $C_H > 0$ it holds that

$$\liminf_{n\to\infty} \inf_{m\in\mathcal{H}(\alpha,C_H)} \mathbb{P}(m(x) \in \mathcal{C}_n(x),\ \forall x \in [0,1]^p) \geq 1 - \beta.$$

The structure of the results is very similar to the random forest confidence bands. For the histogram the assumptions are less complex. Assumption (5.30) on the moments of the errors is similar to its counterpart in Theorem 5.1. If enough moments exist, the main assumptions are up to logarithms that

$$n\delta^{p+2\alpha} \to 0 \qquad \text{and}$$
$$n\delta^{p} \to \infty.$$

The first assumption is an undersmoothing assumption and the second ensures that the estimator is smooth enough to be consistent. These are not contradictory, but as p increases, the assumptions become more restrictive. In the random forest case, this is the same for 2^k and n. If $X_1 \sim U[0,1]^p$ we have $p_x(\delta) = \delta^p$ and hence the radius of the confidence band is

$$\sigma c_\delta(\beta)(n\delta^p)^{-1/2}.$$

We are not aware of any other confidence band result in the literature for the histogram estimator in the multivariate or univariate regression case. However, the structure of the histogram with fixed deterministic cells might imply that the difference in the proof from the univariate to the multivariate case is minor. The first result for the histogram is for univariate density estimation by Smirnov (1950). In addition to the the different estimation problem, this result differs from our result in another crucial aspect. The asymptotic bands rely on an extreme value distribution that is independent of δ. Instead our result provides bands that are contingent on δ, similar to the random forest bands that are dependent on k. Thus, the knowledge of the limit distribution is not necessary, and the accuracy of the bands does not depend on the rate of convergence to that limit distribution.

5.4 The distribution of $\mathbf{S}_k$

In this section we will analyze the distribution of $\mathbf{S}_k \stackrel{d}{=} \sup_{f\in\mathcal{F}_k} |B_k f|$. This distribution and especially its quantiles $c_k(\beta) = F^{-1}_{\mathbf{S}_k}(1-\beta)$ are of interest because (5.9) shows how they directly affect the radius of the confidence bands in Theorem 5.1. The two main objects that characterize the distribution of $\mathbf{S}_k$ are the function class $\mathcal{F}_k$ and the covariance function of the centered Gaussian process $B_k f$. The covariance function is of interest because it directly affects the distribution of the supremum. In one extreme case, the process consists of uncorrelated and thus independent Gaussian random variables. Consequently, the distribution of its maximum has heavier tails compared to the case of positively correlated random variables. Thus, it is evident that the covariance function affects the distribution and especially its quantiles.

We will first analyze the function class and then we will consider the covariance function in the case where X is uniformly distributed. The results regarding the covariance will be of use in Chapter 6, as the same uniformly distributed X will be employed therein. Especially, the results will be helpful to approximate the distribution of $\mathbf{S}_k$ by a Monte Carlo simulation. Figure 5.1 shows the estimated density of $\mathbf{S}_k$ for $k = 5$ and the densities of

$$\sqrt{\frac{n}{\sigma^2 2^{2k}\mathcal{V}_{\cap,k}}}\|U^{(\varepsilon)}_{n,r_n,\omega}\|_\infty$$

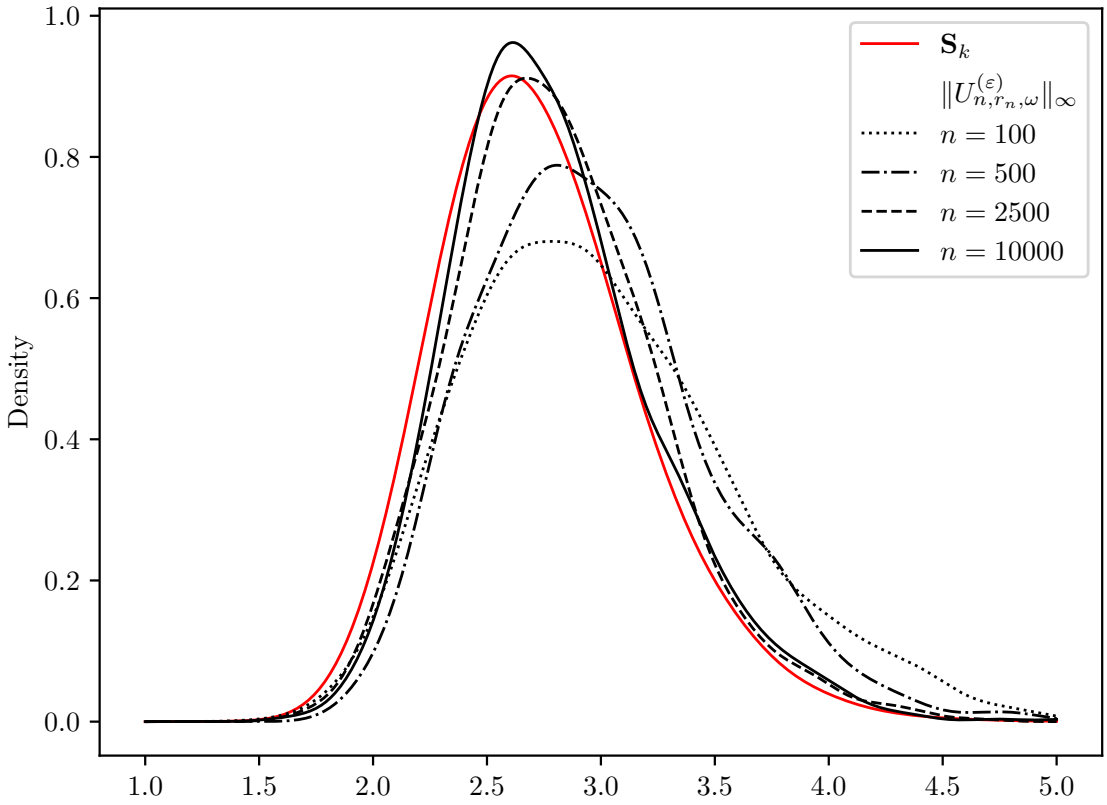

Figure 5.1: Estimated densities of standardized $\|U^{(\varepsilon)}_{n,r_n,\omega}\|_\infty$ for different n and $\mathbf{S}_k$, both for $k = 5$.

for $k = 5$ and different values of n. All densities are estimated by a Gaussian kernel density estimator with the same bandwidth. The density of $\|U^{(\varepsilon)}_{n,r_n,\omega}\|_\infty$ is estimated based on 1 000 copies of the supremum for each n. For $\mathbf{S}_k$, the more efficient simulation described in Chapter 6 allows to use 100 000 copies of $\mathbf{S}_k$ for the density estimation. We observe that the densities get closer to that of $\mathbf{S}_k$ as n increases. We omit to present the exact values, but the Cramér-von-Mises and Kolmogorov-Smirnov distances of the estimated densities to the estimated density of $\mathbf{S}_k$ both decrease monotonously in n.

5.4.1 The function class $\mathcal{F}_k$

For $f_{x_0,k} \in \mathcal{F}_k$ defined in (5.2) and (5.1) we have

$$f_{x_0,k}(x,s) = \sigma^{-1} s \Psi_k^{-1/2}(x_0) K_k(x_0,x) = \sigma^{-1} s \mathbb{E}[K_k(x_0,X_1)^2]^{-1/2} K_k(x_0,x)$$

with

$$K_k(x_0,x) = \mathbb{E}\left[\mathbb{I}\{x \in A_k(x_0,\omega)\} p_{x_0}(\omega)^{-1}\right].$$

This shows that the functions $f_{x_0,k}$ in the class $\mathcal{F}_k$ only depend on x_0 via the cells $A_k(x_0,\omega)$. More precisely, x_0 only occurs as an argument of A_k within the definition of $f_{x_0,k}$.

In Section 3.3.1.3 we studied the behavior of the undividable cells of centered purely random forests. Since there are finitely many of these cells the function class is, in fact, finite. This is the case because

$$A_k(x_1, \omega) = A_k(x_2, \omega) \qquad \forall \omega$$

if x_1 and x_2 are in the same undividable set. We choose the set $\mathcal{X}_k$ such that it contains exactly one element of each undividable cell. This implies that

$$\mathcal{F}_k = \{f_{x_0,k} \mid x_0 \in [0,1]^p\} = \{f_{x_0,k} \mid x_0 \in \mathcal{X}_k\}$$

for $\mathcal{X}_k$ from Section 3.3.1.3 and especially, this is why $|\mathcal{F}_k| = N_f(k)$. For different function classes that only depend on x_0 via the $A_k(x_0, \omega)$ the same arguments apply. We note that the function class can equivalently be identified by the set $\mathcal{X}_k$ and therefore, the process $B_k f_{x_0,k}$ can be interpreted as a process in $\mathcal{X}_k$.

The finite size of the function class will be used in the main proof. For instance, it can be used to bound the expectation of $\mathbf{S}_k$ which leads to a bound for its quantiles denoted in Theorem 5.1. Note that (5.43) from the proof of Theorem 5.1 yields that $c_k(\beta) = \mathcal{O}(\sqrt{k})$.

5.4.2 The covariance function

In general, the covariance of the Gaussian process is

$$\mathrm{Cov}(B_k(f_{x_1,k}), B_k(f_{x_2,k})) = \Psi_k^{-1/2}(x_1)\Psi_k^{-1/2}(x_2)\mathbb{E}\left[K_k(x_1, X_1)K_k(x_2, X_1)\right].$$

If X is uniformly distributed, i.e. $f_X = \mathbb{I}_{[0,1]^p}$, (4.5) and Remark 3.4 imply

$$p_{x_0}(\omega) = \int_{A_k(x_0,\omega)} f_X(x)dx = \mathbb{V}(A_k(x_0, \omega)) = 2^{-k}$$

and thus $p_{x_0} = 2^{-k}$. For K_k its definition from (4.8) implies

$$K_k(x_0, x) = \mathbb{E}\left[\mathbb{I}\{x \in A_k(x_0, \omega)\}\right] 2^k$$

and for Ψ_k its definition from (4.9) implies

$$\begin{aligned}
\Psi_k(x_0) &= \mathbb{E}\left[K_k(x_0, X_1)^2\right]. \\
&= \mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0, \omega_1) \cap A_k(x_0, \omega_2)\}\right] 2^{2k} \\
&= \mathcal{V}_{\cap,k} 2^{2k}.
\end{aligned}$$

Once more, employing the uniform distribution of X, the covariance is

$$\begin{aligned}
\mathrm{Cov}(B_k(f_{x_1,k}), B_k(f_{x_2,k})) &= \Psi_k^{-1/2}(x_1)\Psi_k^{-1/2}(x_2)\mathbb{E}\left[K_k(x_1, X_1)K_k(x_2, X_1)\right] \\
&= \mathcal{V}_{\cap,k}^{-1}\mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_1, \omega_1) \cap A_k(x_2, \omega_2)\}\right] \\
&= \mathcal{V}_{\cap,k}^{-1}\mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_1, \omega_1) \cap A_k(x_2, \omega_2)\} \mid \omega_1, \omega_2\right]\right] \\
&= \mathcal{V}_{\cap,k}^{-1}\mathbb{E}\left[\mathbb{V}\left(A_k(x_1, \omega_1) \cap A_k(x_2, \omega_2)\right)\right]. \qquad (5.31)
\end{aligned}$$

This means the covariance is equal to the expected volume of this intersection divided by $\mathcal{V}_{\cap,k}$ and most importantly does not depend on X_1. Following Definition 3.2, let $S_l(x_0, \omega)$ be the number of splits orthogonal to coordinate l used in the construction of the cell $A_k(x_0, \omega)$.

Lemma 5.11. *It holds that*

$$\begin{aligned}\mathbb{V}\left(A_k(x_1,\omega_1)\cap A_k(x_2,\omega_2)\right)&\\ = 2^{-\sum_{l=1}^{p}\max\{S_l(x_1,\omega_1),S_l(x_2,\omega_2)\}}&\\ \times\prod_{l=1}^{p}\mathbb{I}\left\{\lfloor x_1^{(l)}2^{\min\{S_l(x_1,\omega_1),S_l(x_2,\omega_2)\}}\rfloor=\lfloor x_2^{(l)}2^{\min\{S_l(x_1,\omega_1),S_l(x_2,\omega_2)\}}\rfloor\right\}.&\end{aligned}$$

The lemma implies that the size of the intersection is determined by the maximum number of cuts per direction, if the intersection is not empty. Further it is not empty if and only if the l-th components of x_1 and x_2 are in the same cell of the grid with cell size $2^{-\min(S_l(x_1,\omega_1),S_l(x_2,\omega_2)}$ for all $l\in\{1,\dots.p\}$. Using the lemma, we can calculate the size of the intersection based on the number of splits per direction in the partitions created by ω_1 and ω_2. This is helpful for the estimation of the covariances. For a fixed x_1,x_2 we only need to simulate the number of cuts per direction.

Now we want to identify the positions in the covariance matrix with equal entries. That means cases for $(x_i)_{i=1}^4$with

$$\mathbb{P}\left(X_1\in A_k(x_1,\omega_1)\cap A_k(x_2,\omega_2)\right)=\mathbb{P}\left(X_1\in A_k(x_3,\omega_1)\cap A_k(x_4,\omega_2)\right).$$

Lemma 5.12. *Let $\Pi(p)$ be the permutations of $\{1,\dots,p\}$. It holds that*

$$\begin{aligned}&\mathbb{P}\left(X_1\in A_k(x_1,\omega_1)\cap A_k(x_2,\omega_2)\right)=\mathbb{P}\left(X_1\in A_k(x_3,\omega_1)\cap A_k(x_4,\omega_2)\right)\\ &\Leftrightarrow\exists\pi\in\Pi(p):\left(\max\{t\in\{0,\dots k\}:\lfloor x_1^{(l)}2^t\rfloor=\lfloor x_2^{(l)}2^t\rfloor\}\right)_{l=1}^{p}\\ &\qquad=\left(\max\{t\in\{0,\dots k\}:\lfloor x_3^{(\pi(l))}2^t\rfloor=\lfloor x_4^{(\pi(l))}2^t\rfloor\}\right)_{l=1}^{p}.\end{aligned}$$

The existence of the permutation in the lemma implies that the covariance is determined by the order statistic of

$$\mathfrak{C}_k(x_1,x_2):=\left(\max\{t\in\{0,\dots k\}:\lfloor x_1^{(l)}2^t\rfloor=\lfloor x_2^{(l)}2^t\rfloor\}\right)_{l=1}^{p}\in\{0,\dots,k\}^p. \tag{5.32}$$

Let $\mathfrak{C}_k^{(\uparrow)}(x_1,x_2)$ denote the order statistic of the above vector, that is the increasingly ordered version of $\mathfrak{C}_k(x_1,x_2)$. In the unordered vector above each of the entries corresponds to the finest 2^{-t}-grid on $[0,1]$ in which both $x_1^{(l)}$ and $x_2^{(l)}$ are in the same cell. We can interpret the vector $\mathfrak{C}_k(x_1,x_2)$ as the componentwise closeness of x_1 and x_2 in this grid. Larger entries imply that they are closer and the maximum closeness is $\mathfrak{C}_k(x_0,x_0)=(k,\dots,k)$.

Let us denote

$$\Omega_S:=\{(t_1,\dots,t_p)\mid t_i\in\{0,\dots k\}\forall i\in\{1,\dots p\},t_1\le t_2\le\dots\le t_p\}.$$

We know that

$$\left|\Omega_S\right|=\binom{k+p}{p}.$$

Further, it holds that $\mathfrak{C}_k^{(\uparrow)}(x_1,x_2)\in\Omega_S$ for all $x_1,x_2\in[0,1]^p$. Hence there are at most $\binom{k+p}{p}$ distinct entries in the covariance matrix. Again this is helpful for the estimation of the matrix.

Another implication of these two lemmas is that the volume of the intersection does not directly depend on x_1 and x_2 but rather their relation to each other. By relation we mean the vector

$$\mathfrak{C}_k(x_1, x_2) = \left(\max\{t \in \{0, \dots k\} : \lfloor x_1^{(l)} 2^t \rfloor = \lfloor x_2^{(l)} 2^t \rfloor\}\right)_{l=1}^p .$$

The splits do not directly depend on x_1 and x_2 as well. We note that

$$\begin{aligned}
&\mathbb{V}\left(A_k(x_1, \omega_1) \cap A_k(x_2, \omega_2)\right) \\
&\quad = 2^{-\sum_{l=1}^p \max\{S_l(x_1,\omega_1), S_l(x_2,\omega_2)\}} \\
&\qquad \times \prod_{l=1}^p \mathbb{I}\left\{\lfloor x_1^{(l)} 2^{\min\{S_l(x_1,\omega_1), S_l(x_2,\omega_2)\}} \rfloor = \lfloor x_2^{(l)} 2^{\min\{S_l(x_1,\omega_1), S_l(x_2,\omega_2)\}} \rfloor\right\} \\
&\quad = 2^{-\sum_{l=1}^p \max\{S_l(x_1,\omega_1), S_l(x_2,\omega_2)\}} \prod_{l=1}^p \mathbb{I}\left\{\min\{S_l(x_1,\omega_1), S_l(x_2,\omega_2)\} \le \mathfrak{C}_k^{(l)}(x_1, x_2)\right\}
\end{aligned}$$

by the definition of $\mathfrak{C}_k(x_1, x_2)$. For any ω let $S(\omega) = (S_l(\omega))_{l=1}^p \stackrel{d}{=} (S_l(x_0, \omega))_{l=1}^p$. For a fixed closeness of two points represented by $\mathbf{c} \in \{0, \dots k\}^p$ we denote

$$\mathbb{V}_\cap\left(\mathbf{c}, S(\omega_1), S(\omega_2)\right) := 2^{-\sum_{l=1}^p \max\{S_l(\omega_1), S_l(\omega_2)\}} \prod_{l=1}^p \mathbb{I}\left\{\min\{S_l(\omega_1), S_l(\omega_2)\} \le \mathbf{c}_l\right\}. \quad (5.33)$$

We can use this to make another observation regarding the covariance.

Remark 5.13. We know that for x_1 and x_2 the covariance of the Gaussian process at these points is

$$\frac{1}{\mathcal{V}_{\cap,k}} \mathbb{P}\left(X_1 \in A_k(x_1, \omega_1) \cap A_k(x_2, \omega_2)\right) = \frac{1}{\mathcal{V}_{\cap,k}} \mathbb{E}\left[\mathbb{V}\left(A_k(x_1, \omega_1) \cap A_k(x_2, \omega_2)\right)\right].$$

If the covariance is equal to one we have

$$\mathbb{E}\left[\mathbb{V}\left(A_k(x_1, \omega_1) \cap A_k(x_2, \omega_2)\right)\right] = \mathcal{V}_{\cap,k} = \mathbb{E}\left[\mathbb{V}\left(A_k(x_0, \omega_1) \cap A_k(x_0, \omega_2)\right)\right].$$

Let the "relation" between $\mathfrak{C}_k(x_1, x_2) = \mathbf{c}$ be fixed. The above implies that

$$\begin{aligned}
&\mathbb{E}\left[2^{-\sum_{l=1}^p \max\{S_l(\omega_1), S_l(\omega_2)\}} \prod_{l=1}^p \mathbb{I}\left\{\min\{S_l(\omega_1), S_l(\omega_2)\} \le \mathbf{c}_l\right\}\right] \\
&\quad = \mathbb{E}\left[\mathbb{V}_\cap\left(\mathbf{c}, S(\omega_1), S(\omega_2)\right)\right] \\
&\quad = \mathbb{E}\left[\mathbb{V}\left(A_k(x_1, \omega_1) \cap A_k(x_2, \omega_2)\right)\right] \\
&\quad = \mathbb{E}\left[\mathbb{V}\left(A_k(x_0, \omega_1) \cap A_k(x_0, \omega_2)\right)\right] \\
&\quad = \mathbb{E}\left[\mathbb{V}_\cap\left(\{k\}^p, S(\omega_1), S(\omega_2)\right)\right] \\
&\quad = \mathbb{E}\left[2^{-\sum_{l=1}^p \max\{S_l(\omega_1), S_l(\omega_2)\}} \prod_{l=1}^p \mathbb{I}\left\{\min\{S_l(\omega_1), S_l(\omega_2)\} \le k\right\}\right] \\
&\quad = \mathbb{E}\left[2^{-\sum_{l=1}^p \max\{S_l(\omega_1), S_l(\omega_2)\}}\right].
\end{aligned}$$

This implies that

$$\mathbb{P}\left(\bigcap_{l=1}^{p}\{\min\{S_l(\omega_1), S_l(\omega_2)\} \leq \mathbf{c}_l\}\right) = 1$$

and hence

$$\mathbb{P}\left(\min\{S_l(\omega_1), S_l(\omega_2)\} \leq \mathbf{c}_l\right) = 1 \; \forall l \in \{1, \ldots, p\}.$$

Since ω_1 and ω_2 are independent we have

$$\mathbb{P}\left(S_l(\omega_1) \leq \mathbf{c}_l\right) = 1 \; \forall l \in \{1, \ldots, p\}.$$

Since the axes are interchangeable in the feature space and in the partitioning algorithm, there would be a $c^* \leq k$ with $\mathbf{c}_l = c^*$ for all $l \in \{1, \ldots p\}$. This directly implies that the RF estimator is constant on all cells in the 2^{-c^*}grid because these cells are never split by any partition. The limit Gaussian process is a process in the function class

$$\mathcal{F}_k = \{f_{x_0,k}(x,s) = \sigma^{-1}\mathcal{V}_{\cap,k}^{-1/2} s\mathbb{P}\left(x \in A_k(x_0,\omega)\right) \mid x_0 \in [0,1]^p\}.$$

If we know that the partition never splits certain cells it holds that $f_{x_1,k} = f_{x_2,k}$ almost surely if x_1 and x_2 are in one of these cells. Hence the two points in the feature space correspond to the same function in the class and it suffices to consider one point in each cell for the Gaussian process. This makes sense because for fixed k the process is the limit of a sequence of estimators that are constant on these cells. In other words, the estimation of the covariance gives as a method to determine the size of the undividable cells.

5.5 Proof strategy

Similar to Chapter 4, we use the decomposition of the error from (3.13). The important difference is that we need to handle the terms uniformly in x_0. The error can be bounded by

$$\begin{aligned}\sup_{x_0 \in [0,1]^p} |U_{n,r_n,\omega}^{(\mathrm{RF})}(x_0) - m(x_0)| &= \|U_{n,r_n,\omega}^{(\mathrm{RF})} - m\|_\infty \\ &\leq \|U_{n,r_n,\omega}^{(m)} - m\|_\infty + \|U_{n,r_n,\omega}^{(\varepsilon)}\|_\infty.\end{aligned}$$

For the stochastic error we will use a different proof strategy than in Chapter 4 which is necessary because Theorem 2.13 is pointwise. For

$$K_k(x_0, x) = \mathbb{E}\left[\mathbb{I}\{x \in A_k(x_0,\omega)\} p_{x_0}(\omega)^{-1}\right]$$

from (4.8) let us denote

$$\hat{U}_{n,r_n,\omega}^{(\varepsilon)}(x_0) = \frac{1}{n}\sum_{j=1}^{n} \varepsilon_j K_k(x_0, X_j). \tag{5.34}$$

This definition is based on the Hájek projection from (2.23) of the U-statistic $U^{(\varepsilon)}_{n,r_n,\omega}(x_0)$. The Hájek projection is

$$\frac{r_n}{n}\sum_{j=1}^{n} h^{(\varepsilon)}_{n,1}(X_j,\varepsilon_j) = \frac{1}{n}\sum_{j=1}^{n} \varepsilon_j \mathbb{E}\left[\mathbb{I}\{X_j \in A_k(x_0,\omega)\} p_{x_0}(\omega)^{-1}\lambda(r_n, p_{x_0}(\omega)) \mid X_j\right]$$

for

$$h^{(\varepsilon)}_{n,1}(x,s) = \frac{s}{r_n}\mathbb{E}\left[\mathbb{I}\{x \in A_k(x_0,\omega)\} p_{x_0}(\omega)^{-1}\lambda(r_n, p_{x_0}(\omega))\right]$$

from (4.26) in Lemma 4.11. The only difference to (5.34) is the term $\lambda(r_n, p_{x_0}(\omega))$, see (4.25), that converges to one uniformly under reasonable assumptions on k and r_n. Under the assumptions in Theorem 5.1, the term $\hat{U}^{(\varepsilon)}_{n,r_n,\omega}$ is the asymptotically leading term of $U^{(\varepsilon)}_{n,r_n,\omega}$. We call $U^{(\varepsilon)}_{n,r_n,\omega} - \hat{U}^{(\varepsilon)}_{n,r_n,\omega}$ the projection error due to its connection to the Hájek projection. Later we will decompose the projection error into two terms, which we will call remainder terms. It is similar to the proof of Theorem 2.13 that the Hájek projection is the leading term, but since we consider the terms uniformly in x_0 it is necessary to handle both, the leading term and the projection error differently.

The idea for the leading term is that $\hat{U}^{(\varepsilon)}_{n,r_n,\omega}$ is close to a Gaussian process in x_0 if n is large enough. Definition 2.1 of the empirical process applied to the observations $(X_j,\varepsilon_j)_{j=1}^n$ and the function class $\mathcal{F}_k$ from (5.1) yields

$$\begin{aligned}
\hat{U}^{(\varepsilon)}_{n,r_n,\omega}(x_0) &= \frac{1}{n}\sum_{j=1}^{n} \varepsilon_j K_k(x_0, X_j) \\
&= \frac{\sigma\sqrt{\Psi_k(x_0)}}{n}\sum_{j=1}^{n} \sigma^{-1}\Psi_k^{-1/2}(x_0)\varepsilon_j K_k(x_0, X_j) \\
&= \frac{\sqrt{\sigma^2\Psi_k(x_0)}}{n}\sum_{j=1}^{n} \left(f_{x_0,k}(X_j,\varepsilon_j) - \mathbb{E}\left[f_{x_0,k}(X_1,\varepsilon_1)\right]\right) \\
&= \sqrt{\frac{\sigma^2\Psi_k(x_0)}{n}}\mathbb{G}_n f_{x_0,k}.
\end{aligned}$$

We use Theorem 2.4 by Chernozhukov et al. (2014b) to approximate

$$\sup_{f\in\mathcal{F}_k} |\mathbb{G}_n f| = \sqrt{\frac{n}{\sigma^2}} \sup_{x_0\in[0,1]^p} |\Psi_k^{-1/2}(x_0)\hat{U}^{(\varepsilon)}_{n,r_n,\omega}(x_0)|$$

by a sequence of random variables $\mathbf{S}_k$ with $\mathbf{S}_k \overset{d}{=} \sup_{f\in\mathcal{F}_k}|B_k f|$. Therein B_k is a sequence of centered Gaussian processes indexed by $\mathcal{F}_k$ from Theorem 2.4 that has the covariance function

$$\begin{aligned}
\operatorname{Cov}(B_k(f_{x_1,k}), B_k(f_{x_2,k})) &= \mathbb{E}[f_{x_1,k}(X_1,\varepsilon_1) f_{x_2,k}(X_1,\varepsilon_1)] \\
&= \Psi_k^{-1/2}(x_1)\Psi_k^{-1/2}(x_2)\sigma^{-2}\mathbb{E}[\varepsilon_1 K_k(x_1,X_1)\varepsilon_1 K_k(x_2,X_1)] \\
&= \Psi_k^{-1/2}(x_1)\Psi_k^{-1/2}(x_2)\mathbb{E}[K_k(x_1,X_1)K_k(x_2,X_1)].
\end{aligned}$$

In particular, this implies $\operatorname{Var}(B_k(f)) = 1$. The application of Theorem 2.4 leads to the following theorem, whose proof can be found in Section 5.6.10.1.

Theorem 5.14. *Let $B_k(f)$, $f \in \mathcal{F}_k$, be a sequence of centered Gaussian processes with covariance*

$$\operatorname{Cov}(B_k(f_{x_1,k}), B_k(f_{x_2,k})) = \Psi_k^{-1/2}(x_1)\Psi_k^{-1/2}(x_2)\mathbb{E}\left[K_k(x_1, X_1)K_k(x_2, X_1)\right].$$

If $\nu \in [4, \infty)$ and $\mathbb{E}[|\varepsilon_1|^\nu] < \infty$ there exists a sequence of random variables $\mathbf{S}_k \stackrel{d}{=} \sup_{x_0 \in [0,1]^p} |B_k f_{x_0,k}|$ such that

$$\left| \sqrt{\frac{n}{\sigma^2}} \sup_{x_0 \in [0,1]^p} |\Psi_k^{-1/2}(x_0)\hat{U}^{(\varepsilon)}_{n,r_n,\omega}(x_0)| - \mathbf{S}_k \right| = \mathcal{O}_{\mathbb{P}}\left(\frac{(\log n)^{3/2}}{\mathcal{V}_{\cap,k}^{1/2} n^{1/2-1/\nu}} + \frac{(\log n)^{5/4}}{\mathcal{V}_{\cap,k}^{1/4} n^{1/4}} + \frac{\log n}{\mathcal{V}_{\cap,k}^{1/6} n^{1/6}} \right).$$

Using this direct approximation of the supremum it will not be necessary to approximate the whole empirical process uniformly. This is an important difference to previous results in the literature like those by Johnston (1982), Claeskens and Van Keilegom (2003) or Chao et al. (2017). With the above theorem for the leading term of the asymptotic distribution of

$$\sqrt{\frac{n}{\sigma^2}} \|\Psi_k^{-1/2} U^{(\varepsilon)}_{n,r_n,\omega}\|_\infty,$$

it remains to handle the projection error uniformly in x_0. As demonstrated in Section 5.4, the observations concerning the finite size of $\mathcal{F}_k$ for analogous terms imply that the supremum over $x_0 \in [0,1]^p$ of the projection error is, in fact, a maximum over $x_0 \in \mathcal{X}_k$. In particular, the argument implies that for any function g and any realizations $(w_j)_{j \in J}$ of random variables $(\omega_j)_{j \in J}$ it holds that

$$\sup_{x_0 \in [0,1]^p} g\big((A_k(x_0, w_j))_{j \in J}, \eta\big) = \max_{x_0 \in \mathcal{X}_k} g\big((A_k(x_0, w_j))_{j \in J}, \eta\big)$$

where η can be any arbitrary argument that does not depend on x_0. The remainder terms from the projection error are of the above form. This will allow the utilization of union bounds in conjunction with bounds of finite moments of the remainder terms, thereby enabling their uniform handling.

For the approximation error $U^{(m)}_{n,r_n,\omega} - m$ we cannot directly use this argument because m is continuous in x_0 and thus, the supremum is not reduced to a maximum. But after using the Hölder continuity the bound

$$|m(X_1) - m(x_0)|\mathbb{I}\{X_1 \in A_k(x_0, \omega)\} \leq C_H \mathfrak{d}(A_k(x_0, \omega))^\alpha$$

only depends on x_0 via the cell $A_k(x_0, \omega)$ and allows us to use the same argument. To bound the uniform approximation error we will use a bound for the expected diameter of the cells. The averaging over many trees in the random forests leads to

$$\frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \mathfrak{d}(A_k(x_0, \omega_I))^\alpha \approx \mathbb{E}[\mathfrak{d}(A_k(x_0, \omega))^\alpha].$$

Hence the expectation of the supremum will be reduced to supremum of the expectation. These expectations are all the same and we will be able to use a non uniform bound of the diameter. In the context of a single regression tree, this argument is invalid, implying the necessity for bounds on the expectation with the supremum inside. In general, these bounds are of larger magnitude, and as a result, the approximation error is also larger.

5.6 Proofs

In this section, the proofs of all results in this chapter are given.

5.6.1 Proof of Theorem 5.1

The following proposition is essential for the proof of Theorem 5.1. Its proof is postponed to Section 5.6.2.

Proposition 5.15. *Under the assumptions of Theorem 5.1 let B_k be the sequence of Gaussian processes defined in equation (5.8). For any $\nu \geq 4$ with $\mathbb{E}[|\varepsilon_1|^\nu] < \infty$ there exists a sequence of random variables $\mathbf{S}_k \stackrel{d}{=} \sup_{x_0 \in [0,1]^p} |B_k f_{x_0,k}|$ with*

$$\begin{aligned}
&|\sqrt{n}\|\Psi_k^{-1/2}(U^{(\mathrm{RF})}_{n,r_n,\omega} - m)\|_\infty - \sigma \mathbf{S}_k| \\
&\quad = \mathcal{O}_{\mathbb{P}}\left(\frac{(\log n)^{3/2}}{\mathcal{V}_{\cap,k}^{1/2} n^{1/2-1/\nu}} + \frac{(\log n)^{5/4}}{\mathcal{V}_{\cap,k}^{1/4} n^{1/4}} + \frac{\log n}{\mathcal{V}_{\cap,k}^{1/6} n^{1/6}} \right) + o_{\mathbb{P}}\left((\log n)^{-1} \right).
\end{aligned}$$

Proof of Theorem 5.1. We lower bound the coverage probability by

$$\begin{aligned}
&\mathbb{P}\left(m(x) \in \mathcal{C}_n(x),\, \forall x \in [0,1]^p\right) \\
&\quad = \mathbb{P}\left(\|\Psi_k^{-1/2}(U^{(\mathrm{RF})}_{n,r_n,\omega} - m)\|_\infty \leq \frac{\hat{\sigma} c_k(\beta)}{\sqrt{n}} \right) \\
&\quad = 1 - \mathbb{P}\left(\|\Psi_k^{-1/2}(U^{(\mathrm{RF})}_{n,r_n,\omega} - m)\|_\infty > \frac{\hat{\sigma} c_k(\beta)}{\sqrt{n}} \right) \\
&\quad = 1 - \mathbb{P}\left(\|\Psi_k^{-1/2}(U^{(\mathrm{RF})}_{n,r_n,\omega} - m)\|_\infty > \sigma \frac{\hat{\sigma}}{\sigma} \frac{c_k(\beta)}{\sqrt{n}}, |\hat{\sigma}/\sigma - 1| \leq (\log n)^{-2} \right) \\
&\qquad - \mathbb{P}\left(\|\Psi_k^{-1/2}(U^{(\mathrm{RF})}_{n,r_n,\omega} - m)\|_\infty > \sigma \frac{\hat{\sigma}}{\sigma} \frac{c_k(\beta)}{\sqrt{n}}, |\hat{\sigma}/\sigma - 1| > (\log n)^{-2} \right) \\
&\quad \geq 1 - \mathbb{P}\left(\|\Psi_k^{-1/2}(U^{(\mathrm{RF})}_{n,r_n,\omega} - m)\|_\infty > \sigma(1 - (\log n)^{-2}) \frac{c_k(\beta)}{\sqrt{n}}, |\hat{\sigma}/\sigma - 1| \leq (\log n)^{-2} \right) \\
&\qquad - \mathbb{P}\left(\|\Psi_k^{-1/2}(U^{(\mathrm{RF})}_{n,r_n,\omega} - m)\|_\infty > \sigma \frac{\hat{\sigma}}{\sigma} \frac{c_k(\beta)}{\sqrt{n}}, |\hat{\sigma}/\sigma - 1| > (\log n)^{-2} \right) \\
&\quad \geq 1 - \mathbb{P}\left(\|\Psi_k^{-1/2}(U^{(\mathrm{RF})}_{n,r_n,\omega} - m)\|_\infty > \sigma(1 - (\log n)^{-2}) \frac{c_k(\beta)}{\sqrt{n}} \right) \\
&\qquad - \mathbb{P}\left(|\hat{\sigma}/\sigma - 1| > (\log n)^{-2} \right) \\
&\quad \geq 1 - \mathbb{P}\left(\left|\sqrt{n}\|\Psi_k^{-1/2}(U^{(\mathrm{RF})}_{n,r_n,\omega} - m)\|_\infty - \sigma \mathbf{S}_k\right| + \sigma \mathbf{S}_k > \sigma(1 - (\log n)^{-2}) c_k(\beta) \right) \\
&\qquad - \mathbb{P}\left(|\hat{\sigma} - \sigma| > \sigma (\log n)^{-2} \right) \\
&\quad \geq 1 - \mathbb{P}\left(\sigma \mathbf{S}_k > \sigma(1 - (\log n)^{-2}) c_k(\beta) - (\log n)^{-1} \right) - \mathbb{P}\left(|\hat{\sigma} - \sigma| > \sigma (\log n)^{-2} \right) \\
&\qquad - \mathbb{P}\left(\left|\sqrt{n}\|\Psi_k^{-1/2}(U^{(\mathrm{RF})}_{n,r_n,\omega} - m)\|_\infty - \sigma \mathbf{S}_k\right| > (\log n)^{-1} \right). \qquad (5.35)
\end{aligned}$$

Assumption (5.7) implies $\mathbb{E}[|\varepsilon_1|^{q_G}] < \infty$ for $q_G \geq 4$, therefore Proposition 5.15 yields

$$\begin{aligned}
&|\sqrt{n}\|\Psi_k^{-1/2}(U_{n,r_n,\omega}^{(\mathrm{RF})} - m)\|_\infty - \sigma \mathbf{S}_k| \\
&\quad = \mathcal{O}_{\mathbb{P}}\left(\frac{(\log n)^{3/2}}{\mathcal{V}_{\cap,k}^{1/2} n^{1/2-1/q_G}} + \frac{(\log n)^{5/4}}{\mathcal{V}_{\cap,k}^{1/4} n^{1/4}} + \frac{\log n}{\mathcal{V}_{\cap,k}^{1/6} n^{1/6}} \right) + o_{\mathbb{P}}\left((\log n)^{-1}\right).
\end{aligned}$$

We argue why these terms are $o_{\mathbb{P}}((\log n)^{-1})$. With (5.6) we have

$$\frac{(\log n)^{5/2}}{\mathcal{V}_{\cap,k}^{1/2} n^{1/2-1/q_G}} = \Big(\frac{(\log n)^5}{\mathcal{V}_{\cap,k} n} n^{2/q_G} \Big)^{1/2} \to 0.$$

For the second term we have

$$\frac{(\log n)^{9/4}}{\mathcal{V}_{\cap,k}^{1/4} n^{1/4}} = \Big(\frac{(\log n)^9}{\mathcal{V}_{\cap,k} n} \Big)^{1/4} = \Big(\frac{(\log n)^5}{\mathcal{V}_{\cap,k} n} n^{2/q_G} \Big)^{1/4} n^{-1/(2q_G)} \log n \to 0.$$

The third term works the same. Hence

$$\mathbb{P}\big(|\sqrt{n}\|\Psi_k^{-1/2}(U_{n,r_n,\omega}^{(\mathrm{RF})} - m)\|_\infty - \sigma \mathbf{S}_k| > (\log n)^{-1} \big) \to 0. \tag{5.36}$$

Further we have

$$\begin{aligned}
\mathbb{P}\big(|\hat{\sigma} - \sigma| > \sigma (\log n)^{-2} \big) &= \mathbb{P}\left(\frac{|\hat{\sigma}^2 - \sigma^2|}{\hat{\sigma} + \sigma} > \sigma (\log n)^{-2} \right) \\
&\leq \mathbb{P}\big(|\hat{\sigma}^2 - \sigma^2| > \sigma^2 (\log n)^{-2} \big) \to 0
\end{aligned} \tag{5.37}$$

by assumption on $\hat{\sigma}$. It remains to handle

$$\begin{aligned}
&1 - \mathbb{P}\left(\sigma \mathbf{S}_k > \sigma(1 - (\log n)^{-2}) c_k(\beta) - (\log n)^{-1} \right) \\
&\quad = \mathbb{P}\left(\sigma \mathbf{S}_k \leq \sigma c_k(\beta) - \sigma (\log n)^{-2} c_k(\beta) - (\log n)^{-1} \right) \\
&\quad = \mathbb{P}\left(\mathbf{S}_k \leq c_k(\beta) \right) - \mathbb{P}\left(\sigma c_k(\beta) - \sigma (\log n)^{-2} c_k(\beta) - (\log n)^{-1} < \sigma \mathbf{S}_k \leq \sigma c_k(\beta) \right) \\
&\quad = 1 - \beta - \mathbb{P}\left(-(\log n)^{-2} c_k(\beta) - (\log n)^{-1} \sigma^{-1} < \mathbf{S}_k - c_k(\beta) \leq 0 \right) \\
&\quad \geq 1 - \beta - \mathbb{P}\left(|\mathbf{S}_k - c_k(\beta)| \leq (\log n)^{-2} c_k(\beta) + (\log n)^{-1} \sigma^{-1} \right).
\end{aligned} \tag{5.38}$$

Corollary 2.1. by Chernozhukov et al. (2014a) yields that

$$\sup_{\xi \in \mathbb{R}} \mathbb{P}\left(|\mathbf{S}_k - \xi| \leq \kappa \right) \leq 4\kappa \left(\mathbb{E}\left[\sup_{f \in \mathcal{F}_k} |B_k f| \right] + 1 \right).$$

We apply Corollary 2.2.8. by van der Vaart and Wellner (1996) to bound the above expectation. Let d be the standard deviation semimetric on $\mathcal{F}_k$ that is involved in this corollary. It holds that

$$\begin{aligned}
d(f_{x_1,k}, f_{x_2,k}) &:= \mathbb{E}\left[(B_k f_{x_1,k} - B_k f_{x_2,k})^2 \right]^{1/2} \\
&= \mathbb{E}\left[(B_k f_{x_1,k})^2 - 2 B_k f_{x_1,k} B_k f_{x_2,k} + (B_k f_{x_2,k})^2 \right]^{1/2}
\end{aligned}$$

$$
\begin{aligned}
&= (2 - 2\mathbb{E}\left[B_k f_{x_1,k} B_k f_{x_2,k}\right])^{1/2} \\
&= \sqrt{2}\left(1 - \mathrm{Cov}(B_k(f_{x_1,k}), B_k(f_{x_2,k}))\right)^{1/2} \\
&= \sqrt{2}\left(1 - \Psi_k^{-1/2}(x_1)\Psi_k^{-1/2}(x_2)\mathbb{E}\left[K_k(x_1, X_1)K_k(x_2, X_1)\right]\right)^{1/2} \\
&\leq \sqrt{2}. \qquad (5.39)
\end{aligned}
$$

The finite size of $\mathcal{F}_k$, that is $N_f(k) \leq 2^{kp}$, implies that the packing number from Definition 2.2 is bounded by 2^{kp} for any semimetric. For the semimetric d from above $D(\mathcal{F}_k, d, \epsilon) \leq N_f(k) \leq 2^{kp}$ is also implied by its equality to zero if x_1 and x_2 are in one undividable cell. Note that (5.39) further implies that $D(\mathcal{F}_k, d, \epsilon) = 1$ if $\epsilon > \sqrt{2}$. Using these observations Corollary 2.2.8. by van der Vaart and Wellner (1996) yields for a universal constant K that

$$
\begin{aligned}
\mathbb{E}\left[\sup_{f \in \mathcal{F}_k} |B_k f|\right] &\leq \mathbb{E}\left[|B_k f_{x_0,k}|\right] + K \int_0^\infty \sqrt{\log D(\mathcal{F}_k, d, \epsilon)} d\epsilon \\
&\leq \sigma + K \int_0^{\sqrt{2}} \sqrt{\log 2^{kp}} d\epsilon \\
&\leq \sigma + K\sqrt{2}\sqrt{pk \log 2} \\
&\lesssim \sqrt{k}. \qquad (5.40)
\end{aligned}
$$

With Chernozhukov et al. (2014a, Theorem 2.1) we obtain

$$
\sup_{\xi \in \mathbb{R}} \mathbb{P}(|\mathbf{S}_k - \xi| \leq \kappa) \leq 4\kappa \left(\mathbb{E}\left[\sup_{f \in \mathcal{F}_k} |B_k f|\right] + 1\right) \lesssim \kappa\sqrt{k}. \qquad (5.41)
$$

Using (5.41) we obtain

$$
\mathbb{P}\left(|\mathbf{S}_k - c_k(\beta)| \leq (\log n)^{-2} c_k(\beta) + (\log n)^{-1}\sigma^{-1}\right) \lesssim \left((\log n)^{-2} c_k(\beta) + (\log n)^{-1}\sigma^{-1}\right)\sqrt{k}. \qquad (5.42)
$$

Thus, we need an upper bound for $c_k(\beta)$. Again using (5.40), yields that

$$
\mathbb{P}(\mathbf{S}_k \leq \xi) = 1 - \mathbb{P}(\mathbf{S}_k > \xi) \geq 1 - \frac{1}{\xi}\mathbb{E}[\mathbf{S}_k] \geq 1 - \xi^{-1} C\sqrt{k}
$$

for some constant C. We recall that $c_k(\beta) = \inf\{\xi \in \mathbb{R} : \mathbb{P}(\mathbf{S}_k \leq \xi) \geq 1 - \beta\}$. For $\xi = C\sqrt{k}\beta^{-1}$ we get

$$
c_k(\beta) \leq C\beta^{-1}\sqrt{k} \lesssim \sqrt{k}. \qquad (5.43)
$$

With (5.3) and (5.5) which implies $2^k = o(n)$ it holds that $k = \mathcal{O}(\log n)$ and therefore (5.42) is $o(1)$. Thus (5.38) yields

$$
\begin{aligned}
&\mathbb{P}\left(\sigma\mathbf{S}_k \leq \sigma c_k(\beta) - \sigma(\log n)^{-2} c_k(\beta) - (\log n)^{-1}\right) \\
&\geq 1 - \beta - \mathbb{P}\left(|\mathbf{S}_k - c_k(\beta)| \leq (\log n)^{-2} c_k(\beta) + (\log n)^{-1}\sigma^{-1}\right) \\
&= 1 - \beta - o(1).
\end{aligned}
$$

Together with (5.35), (5.36) and (5.37) we obtain

$$
\liminf_{n \to \infty} \inf_{m \in \mathcal{H}(\alpha, C_H)} \mathbb{P}\left(m(x) \in \hat{\mathcal{C}}_n(x),\, \forall x \in [0,1]^p\right) \geq 1 - \beta. \qquad \square
$$

5.6.2 Proof of Proposition 5.15

To prove the proposition we use Theorem 5.14 and the auxiliary results below. Their proofs can be found in Section 5.6.10. We first present the results that handle $U^{(m)}_{n,r_n,\omega}(x_0) - m(x_0)$. For

$$U^{(1)}_{n,r_n,\omega}(x_0) := \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \mathbb{I}\{\exists j \in I : X_j \in A_k(x_0, \omega_I)\}, \tag{5.44}$$

we have

$$U^{(m)}_{n,r_n,\omega}(x_0) - m(x_0) = U^{(m)}_{n,r_n,\omega}(x_0) - m(x_0)U^{(1)}_{n,r_n,\omega}(x_0) + m(x_0)(U^{(1)}_{n,r_n,\omega}(x_0) - 1). \tag{5.45}$$

The two results below handle both terms from this decomposition. Their proofs are postponed to Section 5.6.10.2.

Lemma 5.16. *For an α-Hölder continuous $m : [0,1]^p \to \mathbb{R}$ with Hölder constant C_H it holds that*

$$\mathbb{E}\big[\|U^{(m)}_{n,r_n,\omega} - mU^{(1)}_{n,r_n,\omega}\|_\infty\big] \le C_H \left(\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha] + p^{\alpha/2} N_f(k) \binom{n}{r_n}^{-1/2} \right)$$

for an arbitrary $x \in [0,1]^p$.

Lemma 5.17. *For a bounded $m : [0,1]^p \to \mathbb{R}$ it holds that*

$$\mathbb{E}\big[\|m(U^{(1)}_{n,r_n,\omega} - 1)\|_\infty\big] \le \|m\|_\infty 2^k (1 - c_X 2^{-k})^{r_n}.$$

The next results will handle the projection error introduced in Section 5.5. We have

$$\begin{aligned} U^{(\varepsilon)}_{n,r_n,\omega}(x_0) &= \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \sum_{j \in I} \varepsilon_j \frac{\mathbb{I}\{X_j \in A_k(x_0,\omega_I)\}}{\sum_{i \in I} \mathbb{I}\{X_i \in A_k(x_0,\omega_I)\}} \\ &= \frac{r_n}{n} \sum_{j=1}^n \varepsilon_j \frac{1}{\binom{n-1}{r_n-1}} \sum_{I \in B_{r_n,n} : j \in I} \frac{\mathbb{I}\{X_j \in A_k(x_0,\omega_I)\}}{\sum_{i \in I} \mathbb{I}\{X_i \in A_k(x_0,\omega_I)\}}. \end{aligned}$$

For $\hat{U}^{(\varepsilon)}_{n,r_n,\omega}$ from (5.34) we obtain

$$\begin{aligned} &U^{(\varepsilon)}_{n,r_n,\omega}(x_0) - \hat{U}^{(\varepsilon)}_{n,r_n,\omega}(x_0) \\ &= \frac{1}{n} \sum_{j=1}^n \varepsilon_j \left(\frac{r_n}{\binom{n-1}{r_n-1}} \sum_{I \in B_{r_n,n} : j \in I} \frac{\mathbb{I}\{X_j \in A_k(x_0,\omega_I)\}}{\sum_{i \in I} \mathbb{I}\{X_i \in A_k(x_0,\omega_I)\}} - K_k(x_0, X_j) \right) \\ &= \frac{1}{n} \sum_{j=1}^n \varepsilon_j \frac{r_n}{\binom{n-1}{r_n-1}} \sum_{I \in B_{r_n,n} : j \in I} \left(\frac{\mathbb{I}\{X_j \in A_k(x_0,\omega_I)\}}{\sum_{i \in I} \mathbb{I}\{X_i \in A_k(x_0,\omega_I)\}} - \frac{1}{r_n p_{x_0}(\omega_I)} \mathbb{I}\{X_j \in A_k(x_0,\omega_I)\} \right) \\ &\quad + \frac{1}{n} \sum_{j=1}^n \varepsilon_j \left(\frac{r_n}{\binom{n-1}{r_n-1}} \sum_{I \in B_{r_n,n} : j \in I} \frac{1}{r_n p_{x_0}(\omega_I)} \mathbb{I}\{X_j \in A_k(x_0,\omega_I)\} - K_k(x_0, X_j) \right) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \sum_{j \in I} \varepsilon_j \mathbb{I}\{X_j \in A_k(x_0, \omega_I)\} \left(\frac{1}{\sum_{i \in I} \mathbb{I}\{X_i \in A_k(x_0, \omega_I)\}} - \frac{1}{r_n p_{x_0}(\omega_I)} \right) \\
&\quad + \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j \frac{1}{\binom{n-1}{r_n-1}} \sum_{I \in B_{r_n,n} : j \in I} \left(p_{x_0}(\omega_I)^{-1} \mathbb{I}\{X_j \in A_k(x_0, \omega_I)\} - K_k(x_0, X_j) \right) \\
&=: R^{(1)}_{n,r_n,\omega}(x_0) + R^{(2)}_{n,r_n,\omega}(x_0) \qquad (5.46)
\end{aligned}
$$

The lemmas below handle the two remainder terms $R^{(1)}_{n,r_n,\omega}(x_0)$ and $R^{(2)}_{n,r_n,\omega}(x_0)$ separately. Their proofs are again postponed and can be found in Section 5.6.10.3.

Lemma 5.18. *Let $q \in 2\mathbb{N}$ be fixed with $q \leq r_n$ and $r_n \to \infty$. Assume that $\mathbb{E}\left[|\varepsilon_1|^q\right] < \infty$ and $2^k \leq r_n$, then there exists a constant C that depends on q but not on n and k such that*

$$
\mathbb{E}\left[R^{(1)}_{n,r_n,\omega}(x_0)^q\right] \leq C \left(\frac{2^{2k}}{r_n n} \right)^{q/2}.
$$

Lemma 5.19. *For $R^{(2)}_{n,r_n,\omega}(x_0)$ it holds that*

$$
\mathbb{E}\left[R^{(2)}_{n,r_n,\omega}(x_0)^2\right] = \operatorname{Var}\left(R^{(2)}_{n,r_n,\omega}(x_0)\right) \leq \frac{\sigma^2 C_X}{c_X^2} \frac{2^k}{r_n} \left(\frac{r_n}{n} \right)^{r_n}.
$$

Similar to Lemma 5.19, we can prove a corollary that will be used for a another remainder term that appears in the proof for the KeRF. We will state this corollary before the proof for the KeRF where it is used. Combining all these results, we are able to prove the Proposition 5.15.

Proof of Proposition 5.15. Using the reverse triangle inequality we get

$$
\begin{aligned}
&\left| \|\Psi_k^{-1/2} \hat{U}^{(\varepsilon)}_{n,r_n,\omega}\|_\infty - \|\Psi_k^{-1/2}(U^{(\mathrm{RF})}_{n,r_n,\omega} - m)\|_\infty \right| \\
&\qquad \leq \|\Psi_k^{-1/2}(\hat{U}^{(\varepsilon)}_{n,r_n,\omega} - (U^{(\mathrm{RF})}_{n,r_n,\omega} - m))\|_\infty \\
&\qquad = \|\Psi_k^{-1/2}(\hat{U}^{(\varepsilon)}_{n,r_n,\omega} - U^{(\varepsilon)}_{n,r_n,\omega} - U^{(m)}_{n,r_n,\omega} + m)\|_\infty \\
&\qquad \leq \|\Psi_k^{-1/2}(\hat{U}^{(\varepsilon)}_{n,r_n,\omega} - U^{(\varepsilon)}_{n,r_n,\omega})\|_\infty + \|\Psi_k^{-1/2}(U^{(m)}_{n,r_n,\omega} - m)\|_\infty \\
&\qquad \leq \|\Psi_k^{-1/2} R^{(1)}_{n,r_n,\omega}\|_\infty + \|\Psi_k^{-1/2} R^{(2)}_{n,r_n,\omega}\|_\infty + \|\Psi_k^{-1/2}(U^{(m)}_{n,r_n,\omega} - m)\|_\infty
\end{aligned}
$$

Hence, for the sequence of random variables $\mathbf{S}_k$ as defined in the theorem we get

$$
\begin{aligned}
&|\sqrt{n}\|\Psi_k^{-1/2}(U^{(\mathrm{RF})}_{n,r_n,\omega} - m)\|_\infty - \sigma \mathbf{S}_k| \\
&\quad = \left| \sqrt{n} \left(\|\Psi_k^{-1/2}(U^{(\mathrm{RF})}_{n,r_n,\omega} - m)\|_\infty - \|\Psi_k^{-1/2} \hat{U}^{(\varepsilon)}_{n,r_n,\omega}\|_\infty + \|\Psi_k^{-1/2} \hat{U}^{(\varepsilon)}_{n,r_n,\omega}\|_\infty \right) - \sigma \mathbf{S}_k \right| \\
&\quad \leq \sqrt{n} \left| \|\Psi_k^{-1/2}(U^{(\mathrm{RF})}_{n,r_n,\omega} - m)\|_\infty - \|\Psi_k^{-1/2} \hat{U}^{(\varepsilon)}_{n,r_n,\omega}\|_\infty \right| + \left| \sqrt{n} \|\Psi_k^{-1/2} \hat{U}^{(\varepsilon)}_{n,r_n,\omega}\|_\infty - \sigma \mathbf{S}_k \right| \\
&\quad \leq \sqrt{n} \|\Psi_k^{-1/2}\|_\infty \left(\|U^{(m)}_{n,r_n,\omega} - m\|_\infty + \|R^{(1)}_{n,r_n,\omega}\|_\infty + \|R^{(2)}_{n,r_n,\omega}\|_\infty \right) \\
&\qquad + |\sqrt{n} \|\Psi_k^{-1/2} \hat{U}^{(\varepsilon)}_{n,r_n,\omega}\|_\infty - \sigma \mathbf{S}_k| \\
&\quad \lesssim \left(\frac{n}{2^{2k} \mathcal{V}_{\cap,k}} \right)^{1/2} \left(\|U^{(m)}_{n,r_n,\omega} - m\|_\infty + \|R^{(1)}_{n,r_n,\omega}\|_\infty + \|R^{(2)}_{n,r_n,\omega}\|_\infty \right)
\end{aligned}
$$

$$+ |\sqrt{n}\|\Psi_k^{-1/2}\hat{U}^{(\varepsilon)}_{n,r_n,\omega}\|_\infty - \sigma \mathbf{S}_k| \tag{5.47}$$

because

$$\|\Psi_k^{-1/2}\|_\infty \leq \left(\frac{c_X}{C_X^2} 2^{2k} \mathcal{V}_{\cap,k}\right)^{-1/2}$$

owing to (4.20). We omit the constants for a shorter notation. In the rest of the proof we will show that all terms from (5.47) except the last are $o_{\mathbb{P}}((\log n)^{-1})$. For the last term we apply Theorem 5.14.

For the approximation error we have with (5.45) that

$$\|U^{(m)}_{n,r_n,\omega} - m\|_\infty \leq \|U^{(m)}_{n,r_n,\omega} - mU^{(1)}_{n,r_n,\omega}\|_\infty + \|m(U^{(1)}_{n,r_n,\omega} - 1)\|_\infty.$$

Using Lemma 5.16 we get

$$\begin{aligned}
&\mathbb{P}\left(\|U^{(m)}_{n,r_n,\omega} - mU^{(1)}_{n,r_n,\omega}\|_\infty \geq \kappa(\log n)^{-1}\Big(\frac{n}{2^{2k}\mathcal{V}_{\cap,k}}\Big)^{-1/2}\right) \\
&\leq \left(\frac{n(\log n)^2}{\kappa^2 2^{2k}\mathcal{V}_{\cap,k}}\right)^{1/2} \mathbb{E}\left[\|U^{(m)}_{n,r_n,\omega} - mU^{(1)}_{n,r_n,\omega}\|_\infty\right] \\
&\leq C_H \left(\frac{n(\log n)^2}{\kappa^2 2^{2k}\mathcal{V}_{\cap,k}}\right)^{1/2} \left(\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha] + p^{\alpha/2} N_f(k) \binom{n}{r_n}^{-1/2}\right)
\end{aligned} \tag{5.48}$$

for all $\kappa > 0$. Assumption (5.4) yields

$$C_H\big(\frac{n(\log n)^2}{\kappa^2 2^{2k}\mathcal{V}_{\cap,k}}\big)^{1/2}\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha] \to 0$$

for all $\kappa > 0$. The fact that $\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha] \geq p^{\alpha/2}2^{-\alpha k/p}$ and $\mathcal{V}_{\cap,k} \leq 2^{-k}$ together with (5.4) imply

$$p^\alpha 2^{-2\alpha k/p}\frac{n}{2^k} \leq \mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha]^2 \frac{n(\log n)^2}{2^{2k}\mathcal{V}_{\cap,k}}.$$

Hence $n = o(2^{k(1+2\alpha/p)})$. Again using $\mathcal{V}_{\cap,k} \leq 2^{-k}$ assumption (5.5) implies with (5.13) that $2^k = o(r_n)$. For the squared second term from (5.48) we obtain with (5.3) that

$$\begin{aligned}
C_H^2 \frac{n(\log n)^2}{\kappa^2 2^{2k}\mathcal{V}_{\cap,k}} p^{2\alpha} N_f(k)^2 \binom{n}{r_n}^{-1} &\leq \frac{C_H^2 p^{2\alpha}}{\kappa^2} n(\log n)^2 2^{2kp}\big(\frac{r_n}{n}\big)^{r_n} \\
&= o\big(2^{k(2p+2+2\alpha/p)} c^{r_n}\big) \\
&= o\big(r_n^{(2p+2+2\alpha/p)} \exp(r_n \log(c))\big) \to 0
\end{aligned} \tag{5.49}$$

because $(\log n)^2 = o(2^k)$ and $\log(c) < 0$. Lemma 5.17 yields

$$\begin{aligned}
&\mathbb{P}\left(\|m(U^{(1)}_{n,r_n,\omega} - 1)\|_\infty \geq \kappa(\log n)^{-1}\Big(\frac{n}{2^{2k}\mathcal{V}_{\cap,k}}\Big)^{-1/2}\right) \\
&\leq \kappa^{-1}\left(\frac{n(\log n)^2}{2^{2k}\mathcal{V}_{\cap,k}}\right)^{1/2} \mathbb{E}\left[\|m(U^{(1)}_{n,r_n,\omega} - 1)\|_\infty\right]
\end{aligned}$$

$$\leq \kappa^{-1} \left(\frac{n(\log n)^2}{2^{2k} \mathcal{V}_{\cap,k}} \right)^{1/2} \|m\|_\infty 2^k (1 - c_X 2^{-k})^{r_n}$$
$$= \kappa^{-1} \left(\frac{n(\log n)^2}{\mathcal{V}_{\cap,k}} \right)^{1/2} \|m\|_\infty (1 - c_X 2^{-k})^{r_n} \to 0$$

for all $\kappa > 0$ with (5.13).

Assumption (5.7) implies that $\mathbb{E}[|\varepsilon_1|^{q_R}] < \infty$. Hence for every $\kappa > 0$ Lemma 5.18 yields

$$\begin{aligned}
\mathbb{P}\left(\|R^{(1)}_{n,r_n,\omega}\|_\infty \geq \kappa (\log n)^{-1} \Big(\frac{n}{2^{2k}\mathcal{V}_{\cap,k}} \Big)^{-1/2} \right) &\leq \mathbb{E}\left[\|R^{(1)}_{n,r_n,\omega}\|^{q_R}_\infty \right] \left(\frac{n(\log n)^2}{\kappa^2 2^{2k} \mathcal{V}_{\cap,k}} \right)^{q_R/2} \\
&\lesssim \sum_{x_0 \in \mathcal{X}_k} \mathbb{E}\left[R^{(1)}_{n,r_n,\omega}(x_0)^{q_R} \right] \left(\frac{n(\log n)^2}{2^{2k} \mathcal{V}_{\cap,k}} \right)^{q_R/2} \\
&\leq N_f(k) C \left(\frac{2^{2k}}{r_n n} \right)^{q_R/2} \left(\frac{n(\log n)^2}{2^{2k} \mathcal{V}_{\cap,k}} \right)^{q_R/2} \\
&\lesssim N_f(k) \left(\frac{(\log n)^2}{r_n \mathcal{V}_{\cap,k}} \right)^{q_R/2} \to 0
\end{aligned}$$

due to Assumption (5.5). Lemma 5.19 implies with a union bound and by using $N_f(k) \leq 2^{kp}$ and $\mathcal{V}_{\cap,k} \geq 2^{-2k}$ that

$$\begin{aligned}
\mathbb{P}\left(\|R^{(2)}_{n,r_n,\omega}\|_\infty \geq \kappa (\log n)^{-1} \Big(\frac{n}{2^{2k}\mathcal{V}_{\cap,k}} \Big)^{-1/2} \right) &\leq \mathbb{E}\left[\|R^{(2)}_{n,r_n,\omega}\|^2_\infty \right] \frac{n(\log n)^2}{\kappa^2 2^{2k} \mathcal{V}_{\cap,k}} \\
&\leq \frac{n(\log n)^2}{\kappa^2 2^{2k} \mathcal{V}_{\cap,k}} \sum_{x_0 \in \mathcal{X}_k} \mathbb{E}\left[R^{(2)}_{n,r_n,\omega}(x_0)^2 \right] \\
&\leq N_f(k) \frac{\sigma^2 C_X}{c_X^2} \frac{2^k}{r_n} \frac{n(\log n)^2}{2^{2k} \mathcal{V}_{\cap,k}} \left(\frac{r_n}{n} \right)^{r_n} \\
&= N_f(k) \frac{\sigma^2 C_X}{c_X^2 \kappa^2} \frac{(\log n)^2}{2^k \mathcal{V}_{\cap,k}} \left(\frac{r_n}{n} \right)^{r_n - 1} \\
&\leq \frac{\sigma^2 C_X}{c_X^2 \kappa^2} 2^{k(p+1)} (\log n)^2 c^{r_n - 1} \to 0
\end{aligned}$$

with the same argument we already used for (5.49). Applying (5.47) and Theorem 5.14 for any $\nu \geq 4$ with $\mathbb{E}[|\varepsilon_1|^\nu] < \infty$ we end up with

$$\begin{aligned}
&\left| \sqrt{n} \|\Psi_k^{-1/2} (U^{(\mathrm{RF})}_{n,r_n,\omega} - m)\|_\infty - \sigma \mathbf{S}_k \right| \\
&\quad \lesssim \Big(\frac{n}{2^{2k}\mathcal{V}_{\cap,k}} \Big)^{1/2} \left(\|U^{(m)}_{n,r_n,\omega} - m\|_\infty + \|R^{(1)}_{n,r_n,\omega}\|_\infty + \|R^{(2)}_{n,r_n,\omega}\|_\infty \right) \\
&\qquad + |\sqrt{n} \|\Psi_k^{-1/2} \hat{U}^{(\varepsilon)}_{n,r_n,\omega}\|_\infty - \sigma \mathbf{S}_k| \\
&\quad = o_{\mathbb{P}}\left((\log n)^{-1} \right) + \mathcal{O}_{\mathbb{P}}\left(\frac{(\log n)^{3/2}}{\mathcal{V}_{\cap,k}^{1/2} n^{1/2 - 1/\nu}} + \frac{(\log n)^{5/4}}{\mathcal{V}_{\cap,k}^{1/4} n^{1/4}} + \frac{\log n}{\mathcal{V}_{\cap,k}^{1/6} n^{1/6}} \right). \qquad \square
\end{aligned}$$

5.6.3 Proof of Lemma 5.4

The estimation error can be bounded by

$$\begin{aligned}
|\hat{\sigma}^2 - \sigma^2| &= \left| \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\varepsilon}_i^2 - \sigma^2 \right) \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^{n} \left(\left(Y_i - U_{n,r_n,\omega}^{(\mathrm{RF})}(X_i) \right)^2 - \sigma^2 \right) \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^{n} \left(\left(m(X_i) + \varepsilon_i - U_{n,r_n,\omega}^{(\mathrm{RF})}(X_i) \right)^2 - \sigma^2 \right) \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^{n} \left(\left(m(X_i) - U_{n,r_n,\omega}^{(\mathrm{RF})}(X_i) \right)^2 + 2\varepsilon_i \left(m(X_i) - U_{n,r_n,\omega}^{(\mathrm{RF})}(X_i) \right) + \varepsilon_i^2 - \sigma^2 \right) \right| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \left(m(X_i) - U_{n,r_n,\omega}^{(\mathrm{RF})}(X_i) \right)^2 + 2 \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left(m(X_i) - U_{n,r_n,\omega}^{(\mathrm{RF})}(X_i) \right) \right| \\
&\quad + \left| \frac{1}{n} \sum_{i=1}^{n} \left(\varepsilon_i^2 - \sigma^2 \right) \right| \\
&=: E_1 + 2E_2 + E_3.
\end{aligned}$$

Markov's inequality implies

$$\begin{aligned}
\mathbb{P}\left(|\hat{\sigma}^2 - \sigma^2| > \kappa \right) &\leq \mathbb{P}\left(E_1 > \kappa/3 \right) + \mathbb{P}\left(E_2 > \kappa/3 \right) + \mathbb{P}\left(E_3 > \kappa/3 \right) \\
&\leq \frac{3}{\kappa} \mathbb{E}\left[E_1 \right] + \frac{6}{\kappa} \mathbb{E}\left[E_2 \right] + \frac{3}{\kappa} \mathbb{E}\left[E_3 \right]. \qquad (5.50)
\end{aligned}$$

The first expectation is bounded by

$$\begin{aligned}
\mathbb{E}\left[E_1 \right] &= \mathbb{E}\left[\left(m(X_1) - U_{n,r_n,\omega}^{(\mathrm{RF})}(X_1) \right)^2 \right] \\
&= \mathbb{E}\left[\left(m(X_1) - U_{n,r_n,\omega}^{(m)}(X_1) - U_{n,r_n,\omega}^{(\varepsilon)}(X_1) \right)^2 \right] \\
&\leq 2\mathbb{E}\left[\left(m(X_1) - U_{n,r_n,\omega}^{(m)}(X_1) \right)^2 \right] + 2\mathbb{E}\left[U_{n,r_n,\omega}^{(\varepsilon)}(X_1)^2 \right]. \qquad (5.51)
\end{aligned}$$

Using (5.45) we get

$$\begin{aligned}
&\mathbb{E}\left[\left(m(X_1) - U_{n,r_n,\omega}^{(m)}(X_1) \right)^2 \right] \\
&\lesssim \mathbb{E}\left[\left(U_{n,r_n,\omega}^{(m)}(X_1) - m(X_1) U_{n,r_n,\omega}^{(1)}(X_1) \right)^2 \right] + \mathbb{E}\left[m(X_1)^2 \left(U_{n,r_n,\omega}^{(1)}(X_1) - 1 \right)^2 \right] \\
&\leq \mathbb{E}\left[\| U_{n,r_n,\omega}^{(m)} - m U_{n,r_n,\omega}^{(1)} \|_\infty^2 \right] + \|m\|_\infty^2 \mathbb{E}\left[\left(U_{n,r_n,\omega}^{(1)}(X_1) - 1 \right)^2 \right]. \qquad (5.52)
\end{aligned}$$

Using the same arguments as in Lemma 5.16, but instead for the expectation of the squared expression we get

$$\mathbb{E}\left[\| U_{n,r_n,\omega}^{(m)} - m U_{n,r_n,\omega}^{(1)} \|_\infty^2 \right] \leq C_H^2 \left(\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha]^2 + p^\alpha N_f(k) \binom{n}{r_n}^{-1} \right). \qquad (5.53)$$

With the definition of $U^{(1)}_{n,r_n,\omega}$ from (5.44) we get

$$\begin{aligned}
\left(U^{(1)}_{n,r_n,\omega}(X_1)-1\right)^2 &= \left(\frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}}\left(\mathbb{I}\{\exists j\in I: X_j\in A_k(X_1,\omega_I)\}-1\right)\right)^2\\
&= \left(\frac{n-r_n}{n}\frac{1}{\binom{n-1}{r_n}}\sum_{I\in B_{r_n,n}:1\notin I}\left(\mathbb{I}\{\exists j\in I: X_j\in A_k(X_1,\omega_I)\}-1\right)\right)^2\\
&\le \frac{1}{\binom{n-1}{r_n}}\sum_{I\in B_{r_n,n}:1\notin I}\left(\mathbb{I}\{\exists j\in I: X_j\in A_k(X_1,\omega_I)\}-1\right)^2\\
&= \frac{1}{\binom{n-1}{r_n}}\sum_{I\in B_{r_n,n}:1\notin I}\mathbb{I}\{\nexists j\in I: X_j\in A_k(X_1,\omega_I)\}.
\end{aligned}$$

Thus, with (4.6) we obtain

$$\begin{aligned}
\mathbb{E}\left[\left(U^{(1)}_{n,r_n,\omega}(X_1)-1\right)^2\right] &\le \mathbb{E}\left[\mathbb{I}\{\nexists j\in\{2,\dots,r_n+1\}: X_j\in A_k(X_1,\omega)\}\right]\\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\{\nexists j\in\{2,\dots,r_n+1\}: X_j\in A_k(X_1,\omega)\}\mid X_1,\omega\right]\right]\\
&= \mathbb{E}\left[\mathbb{E}\left[(1-p_{X_1}(\omega))^{r_n}\mid X_1,\omega\right]\right]\\
&\le (1-c_X2^{-k})^{r_n}. \qquad (5.54)
\end{aligned}$$

The equations (5.52), (5.53) and (5.54) yield

$$\begin{aligned}
\mathbb{E}\left[\left(m(X_1)-U^{(m)}_{n,r_n,\omega}(X_1)\right)^2\right] \lesssim C_H^2\left(\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha]^2+p^\alpha N_f(k)\binom{n}{r_n}^{-1}\right)\\
+\|m\|_\infty^2(1-c_X2^{-k})^{r_n}. \qquad (5.55)
\end{aligned}$$

We continue with the second term from (5.51), which can be decomposed into

$$\begin{aligned}
U^{(\varepsilon)}_{n,r_n,\omega}(X_1) &= \frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}:1\notin I}\sum_{j\in I}\varepsilon_j\frac{\mathbb{I}\{X_j\in A_k(X_1,\omega_I)\}}{\sum_{i\in I}\mathbb{I}\{X_i\in A_k(X_1,\omega_I)\}}\\
&\quad+\frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}:1\in I}\sum_{j\in I}\varepsilon_j\frac{\mathbb{I}\{X_j\in A_k(X_1,\omega_I)\}}{1+\sum_{i\in I\setminus\{1\}}\mathbb{I}\{X_i\in A_k(X_1,\omega_I)\}}\\
&= \frac{n-r_n}{n}\frac{1}{\binom{n-1}{r_n}}\sum_{I\in B_{r_n,n}:1\notin I}\sum_{j\in I}\varepsilon_j\frac{\mathbb{I}\{X_j\in A_k(X_1,\omega_I)\}}{\sum_{i\in I}\mathbb{I}\{X_i\in A_k(X_1,\omega_I)\}}\\
&\quad+\frac{r_n}{n}\frac{1}{\binom{n-1}{r_n-1}}\sum_{I\in B_{r_n,n}:1\in I}\sum_{j\in I}\varepsilon_j\frac{\mathbb{I}\{X_j\in A_k(X_1,\omega_I)\}}{1+\sum_{i\in I\setminus\{1\}}\mathbb{I}\{X_i\in A_k(X_1,\omega_I)\}}\\
&=: \frac{n-r_n}{n}U_1+\frac{r_n}{n}U_2. \qquad (5.56)
\end{aligned}$$

U_1 conditional X_1 is a generalized U-statistic analogue to $U^{(\varepsilon)}_{n-1,r_n,\omega}$ on the sample with index set $\{2,\dots,n\}$. We apply Lemma 2.11 and Lemma 2.15 with (4.6) to obtain

$$\mathbb{E}\left[U_1^2\mid X_1\right]\le\frac{r_n}{n-1}\mathbb{E}\left[\left(\sum_{j=2}^{r_n+1}\varepsilon_j\frac{\mathbb{I}\{X_j\in A_k(X_1,\omega)\}}{\sum_{i=2}^{r_n+1}\mathbb{I}\{X_i\in A_k(X_1,\omega)\}}\right)^2\,\Big|\,X_1\right]$$

$$
\begin{aligned}
&\leq \sigma^2 \frac{r_n^2}{n-1} \mathbb{E}\left[\frac{\mathbb{I}\{X_2 \in A_k(X_1,\omega)\}}{\left(1+\sum_{i=3}^{r_n+1} \mathbb{I}\{X_i \in A_k(X_1,\omega)\}\right)^2} \,\middle|\, X_1 \right] \\
&= \sigma^2 \frac{r_n^2}{n-1} \mathbb{E}\left[\frac{\mathbb{E}\left[\mathbb{I}\{X_2 \in A_k(X_1,\omega)\} \mid X_1, (X_i)_{i=3}^{r_n+1}, \omega\right]}{\left(1+\sum_{i=3}^{r_n+1} \mathbb{I}\{X_i \in A_k(X_1,\omega)\}\right)^2} \,\middle|\, X_1 \right] \\
&= \sigma^2 \frac{r_n^2}{n-1} \mathbb{E}\left[\frac{p_{X_1}(\omega)}{\left(1+\sum_{i=3}^{r_n+1} \mathbb{I}\{X_i \in A_k(X_1,\omega)\}\right)^2} \,\middle|\, X_1 \right] \\
&\leq \sigma^2 C_X \frac{r_n^2}{(n-1)2^k} \mathbb{E}\left[\left(1+\sum_{i=2}^{r_n+1} \mathbb{I}\{X_i \in A_k(X_1,\omega)\}\right)^{-2} \,\middle|\, X_1 \right] \\
&\leq \sigma^2 C_X \frac{r_n^2}{(n-1)2^k} 2 \frac{2^{2k}}{c_X^2 r_n^2} \\
&= 2\sigma^2 \frac{C_X}{c_X^2} \frac{2^k}{(n-1)}. \qquad (5.57)
\end{aligned}
$$

For U_2 we use Jensen's inequality and again Lemma 2.15 with (4.6) to obtain

$$
\begin{aligned}
\mathbb{E}\left[U_2^2\right] &= \mathbb{E}\left[\left(\frac{1}{\binom{n-1}{r_n-1}} \sum_{I \in B_{r_n,n}: 1 \in I} \sum_{j \in I} \varepsilon_j \frac{\mathbb{I}\{X_j \in A_k(X_1,\omega_I)\}}{1+\sum_{i \in I \setminus \{1\}} \mathbb{I}\{X_i \in A_k(X_1,\omega_I)\}} \right)^2 \right] \\
&\leq \mathbb{E}\left[\left(\sum_{j=1}^{r_n} \varepsilon_j \frac{\mathbb{I}\{X_j \in A_k(X_1,\omega)\}}{1+\sum_{i=2}^{r_n} \mathbb{I}\{X_i \in A_k(X_1,\omega)\}} \right)^2 \right] \\
&= \sigma^2 \sum_{j=1}^{r_n} \mathbb{E}\left[\left(\frac{\mathbb{I}\{X_j \in A_k(X_1,\omega)\}}{1+\sum_{i=2}^{r_n} \mathbb{I}\{X_i \in A_k(X_1,\omega)\}} \right)^2 \right] \\
&= \sigma^2 (r_n-1) \mathbb{E}\left[\left(\frac{\mathbb{I}\{X_2 \in A_k(X_1,\omega)\}}{2+\sum_{i=3}^{r_n} \mathbb{I}\{X_i \in A_k(X_1,\omega)\}} \right)^2 \right] \\
&\qquad + \sigma^2 \mathbb{E}\left[\left(1+\sum_{i=2}^{r_n} \mathbb{I}\{X_i \in A_k(X_1,\omega)\}\right)^{-2} \right] \\
&\leq \sigma^2 (r_n-1) C_X 2^{-k} \mathbb{E}\left[\left(2+\sum_{i=3}^{r_n} \mathbb{I}\{X_i \in A_k(X_1,\omega)\}\right)^2 \right] \\
&\qquad + \sigma^2 \mathbb{E}\left[\left(1+\sum_{i=2}^{r_n} \mathbb{I}\{X_i \in A_k(X_1,\omega)\}\right)^{-2} \right] \\
&\leq \sigma^2 \left((r_n-1) C_X 2^{-k} + 1 \right) \mathbb{E}\left[\left(1+\sum_{i=2}^{r_n} \mathbb{I}\{X_i \in A_k(X_1,\omega)\}\right)^{-2} \right] \\
&\leq 2\sigma^2 \left((r_n-1) C_X 2^{-k} + 1 \right) \frac{2^{2k}}{c_X^2 r_n^2}. \qquad (5.58)
\end{aligned}
$$

Together with (5.56), (5.57) and (5.58) yield

$$\begin{aligned}
\mathbb{E}\left[U^{(\varepsilon)}_{n,r_n,\omega}(X_1)^2\right] &\leq 2\mathbb{E}\left[\left(\frac{n-r_n}{n}U_1\right)^2\right] + 2\mathbb{E}\left[\left(\frac{r_n}{n}U_2\right)^2\right] \\
&\leq 2\mathbb{E}\left[\mathbb{E}\left[U_1^2 \mid X_1\right]\right] + 2\left(\frac{r_n}{n}\right)^2 \mathbb{E}\left[U_2^2\right] \\
&\leq 4\sigma^2 \frac{C_X}{c_X^2}\frac{2^k}{(n-1)} + 4\left(\frac{2^k}{c_X n}\right)^2 \sigma^2\left((r_n-1)C_X 2^{-k}+1\right) \\
&= \mathcal{O}\left(2^k/n\right) + \mathcal{O}\left(\frac{2^k r_n}{n^2}\right) + \mathcal{O}\left(\frac{2^{2k}}{n^2}\right) \\
&= \mathcal{O}\left(2^k/n\right) + \mathcal{O}\left(\frac{2^{2k}}{n^2}\right) \qquad (5.59)
\end{aligned}$$

(5.51), (5.55) and (5.59) imply

$$\begin{aligned}
\mathbb{E}\left[E_1\right] &\leq 2\mathbb{E}\left[\left(m(X_1) - U^{(m)}_{n,r_n,\omega}(X_1)\right)^2\right] + 2\mathbb{E}\left[U^{(\varepsilon)}_{n,r_n,\omega}(X_1)^2\right] \\
&= \mathcal{O}\left(\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha]^2\right) + \mathcal{O}\left(N_f(k)\binom{n}{r_n}^{-1}\right) + \mathcal{O}\left((1-c_X 2^{-k})^{r_n}\right) \\
&\quad + \mathcal{O}\left(2^k/n\right) + \mathcal{O}\left(\frac{2^{2k}}{n^2}\right). \qquad (5.60)
\end{aligned}$$

We proceed with E_2.

$$\mathbb{E}\left[E_2\right] \leq \mathbb{E}\left[\left|\frac{1}{n}\sum_{l=1}^n \varepsilon_l\left(m(X_l) - U^{(m)}_{n,r_n,\omega}(X_l)\right)\right|\right] + \mathbb{E}\left[\left|\frac{1}{n}\sum_{l=1}^n \varepsilon_l U^{(\varepsilon)}_{n,r_n,\omega}(X_l)\right|\right] \qquad (5.61)$$

We use the independence of $(m(X_l), U^{(m)}_{n,r_n,\omega}(X_l))$ and ε_l in conjunction with $\|U^{(m)}_{n,r_n,N,\omega}\|_\infty \leq \|m\|_\infty$ to obtain

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{1}{n}\sum_{l=1}^n \varepsilon_l\left(m(X_l) - U^{(m)}_{n,r_n,\omega}(X_l)\right)\right|\right]^2 &\leq \mathbb{E}\left[\left(\frac{1}{n}\sum_{l=1}^n \varepsilon_l\left(m(X_l) - U^{(m)}_{n,r_n,\omega}(X_l)\right)\right)^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\varepsilon_1^2\left(m(X_1) - U^{(m)}_{n,r_n,\omega}(X_1)\right)^2\right] \\
&\leq \frac{1}{n}\sigma^2 4\|m\|_\infty^2. \qquad (5.62)
\end{aligned}$$

The second part from 5.61) satisfies

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{1}{n}\sum_{l=1}^n \varepsilon_l U^{(\varepsilon)}_{n,r_n,\omega}(X_l)\right|\right]^2 &\leq \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{l=1}^n \varepsilon_l U^{(\varepsilon)}_{n,r_n,\omega}(X_l)\right)^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\varepsilon_1^2 U^{(\varepsilon)}_{n,r_n,\omega}(X_1)^2\right]
\end{aligned}$$

$$+ \frac{(n-1)}{n}\mathbb{E}\left[\varepsilon_1\varepsilon_2 U^{(\varepsilon)}_{n,r_n,\omega}(X_1)U^{(\varepsilon)}_{n,r_n,\omega}(X_2)\right] \tag{5.63}$$

Using the decomposition from (5.56) we obtain

$$\varepsilon_1^2 U^{(\varepsilon)}_{n,r_n,\omega}(X_1)^2 = \varepsilon_1^2\left(\frac{n-r_n}{n}U_1 + \frac{r_n}{n}U_2\right)^2 \leq 2\varepsilon_1^2\left(\frac{n-r_n}{n}U_1\right)^2 + 2\varepsilon_1^2\left(\frac{r_n}{n}U_2\right)^2. \tag{5.64}$$

Analogously to (5.58) we obtain

$$\begin{aligned}
\mathbb{E}\left[\varepsilon_1^2 U_2^2\right] &= \mathbb{E}\left[\varepsilon_1^2\left(\frac{1}{\binom{n-1}{r_n-1}}\sum_{I\in B_{r_n,n}:1\in I}\sum_{j\in I}\varepsilon_j\frac{\mathbb{I}\{X_j\in A_k(X_1,\omega_I)\}}{1+\sum_{i\in I\setminus\{1\}}\mathbb{I}\{X_i\in A_k(X_1,\omega_I)\}}\right)^2\right] \\
&\leq \mathbb{E}\left[\varepsilon_1^2\left(\sum_{j=1}^{r_n}\varepsilon_j\frac{\mathbb{I}\{X_j\in A_k(X_1,\omega)\}}{1+\sum_{i=2}^{r_n}\mathbb{I}\{X_i\in A_k(X_1,\omega)\}}\right)^2\right] \\
&\leq \mathbb{E}\left[\varepsilon_1^4\right]\sum_{j=1}^{r_n}\mathbb{E}\left[\left(\frac{\mathbb{I}\{X_j\in A_k(X_1,\omega)\}}{1+\sum_{i=2}^{r_n}\mathbb{I}\{X_i\in A_k(X_1,\omega)\}}\right)^2\right] \\
&\leq 2\mathbb{E}\left[\varepsilon_1^4\right]\left((r_n-1)C_X 2^{-k}+1\right)\frac{2^{2k}}{c_X^2 r_n^2}.
\end{aligned} \tag{5.65}$$

The independence of U_1 and ε_1 together with (5.64), (5.65) and (5.57) yield

$$\begin{aligned}
\mathbb{E}\left[\varepsilon_1^2 U^{(\varepsilon)}_{n,r_n,\omega}(X_1)^2\right] &\leq \mathbb{E}\left[2\varepsilon_1^2\left(\frac{n-r_n}{n}U_1\right)^2\right] + \mathbb{E}\left[2\varepsilon_1^2\left(\frac{r_n}{n}U_2\right)^2\right] \\
&= 2\left(\frac{n-r_n}{n}\right)^2\mathbb{E}\left[\varepsilon_1^2\right]\mathbb{E}\left[U_1^2\right] + 2\left(\frac{r_n}{n}\right)^2\mathbb{E}\left[\varepsilon_1^2U_2^2\right] \\
&\leq 4\frac{C_X}{c_X^2}\sigma^4\frac{2^k}{(n-1)} + 4\frac{C_X}{c_X^2}\mathbb{E}\left[\varepsilon_1^4\right]\frac{2^k r_n}{n^2} + 4\left(\frac{2^k}{c_X n}\right)^2\mathbb{E}\left[\varepsilon_1^4\right] \\
&= \mathcal{O}\left(2^k/n\right).
\end{aligned} \tag{5.66}$$

Using that

$$U^{(\varepsilon)}_{n,r_n,\omega}(X_l) = \sum_{j=1}^{n}\varepsilon_j\frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}:j\in I}\frac{\mathbb{I}\{X_j\in A_k(X_l,\omega_I)\}}{\sum_{i\in I}\mathbb{I}\{X_i\in A_k(X_l,\omega_I)\}} =: \sum_{j=1}^{n}\varepsilon_j\tilde{W}_j(X_l),$$

where $\tilde{W}_j(X_l)$ is independent of $(\varepsilon_j)_{j=1}^n$ we obtain with the Cauchy-Schwarz inequality that

$$\begin{aligned}
\mathbb{E}\left[\varepsilon_1\varepsilon_2 U^{(\varepsilon)}_{n,r_n,\omega}(X_1)U^{(\varepsilon)}_{n,r_n,\omega}(X_2)\right] &= \mathbb{E}\left[\varepsilon_1\varepsilon_2\sum_{j=1}^{n}\varepsilon_j\tilde{W}_j(X_1)\sum_{i=1}^{n}\varepsilon_i\tilde{W}_i(X_2)\right] \\
&= \mathbb{E}\left[\varepsilon_1^2\varepsilon_2^2\tilde{W}_1(X_1)\tilde{W}_2(X_2)\right] + \mathbb{E}\left[\varepsilon_1^2\varepsilon_2^2\tilde{W}_2(X_1)\tilde{W}_1(X_2)\right] \\
&= \sigma^4\mathbb{E}\left[\tilde{W}_1(X_1)\tilde{W}_2(X_2)\right] + \sigma^4\mathbb{E}\left[\tilde{W}_2(X_1)\tilde{W}_1(X_2)\right]
\end{aligned}$$

$$\leq \sigma^4 \mathbb{E}\left[\tilde{W}_1(X_1)^2\right] + \sigma^4 \mathbb{E}\left[\tilde{W}_1(X_2)^2\right]. \tag{5.67}$$

We omit the usual step involving the tower rule with the conditional expectation given ω, and Lemma 2.15 with (4.6) yields

$$\begin{aligned}
\mathbb{E}\left[\tilde{W}_1(X_1)^2\right] &= \mathbb{E}\left[\left(\frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}: 1 \in I} \frac{\mathbb{I}\{X_1 \in A_k(X_1, \omega_I)\}}{\sum_{i \in I} \mathbb{I}\{X_i \in A_k(X_1, \omega_I)\}}\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{1}{\binom{n-1}{r_n-1}} \sum_{I \in B_{r_n,n}: 1 \in I} \frac{r_n}{n} \frac{1}{1 + \sum_{i \in I \setminus \{1\}} \mathbb{I}\{X_i \in A_k(X_1, \omega_I)\}}\right)^2\right] \\
&\leq \frac{r_n^2}{n^2} \mathbb{E}\left[\left(1 + \sum_{i=2}^{r_n} \mathbb{I}\{X_i \in A_k(X_1, \omega)\}\right)^{-2}\right] \\
&\leq \frac{r_n^2}{n^2} \frac{2^{2k+1}}{c_X^2 r_n^2} \\
&= 2\frac{2^{2k}}{c_X^2 n^2},
\end{aligned}$$

and similarly

$$\begin{aligned}
\mathbb{E}\left[\tilde{W}_1(X_2)^2\right] &= \mathbb{E}\left[\left(\frac{1}{\binom{n-1}{r_n-1}} \sum_{I \in B_{r_n,n}: 1 \in I} \frac{r_n}{n} \frac{\mathbb{I}\{X_1 \in A_k(X_2, \omega_I)\}}{1 + \sum_{i \in I \setminus \{1\}} \mathbb{I}\{X_i \in A_k(X_2, \omega_I)\}}\right)^2\right] \\
&\leq \frac{r_n^2}{n^2} \frac{1}{\binom{n-1}{r_n-1}} \sum_{I \in B_{r_n,n}: 1 \in I} \mathbb{E}\left[\left(\frac{\mathbb{I}\{X_1 \in A_k(X_2, \omega_I)\}}{1 + \sum_{i \in I \setminus \{1\}} \mathbb{I}\{X_i \in A_k(X_2, \omega_I)\}}\right)^2\right] \\
&\leq \frac{r_n^2}{n^2} \mathbb{E}\left[\left(1 + \sum_{i=3}^{r_n+1} \mathbb{I}\{X_i \in A_k(X_2, \omega)\}\right)^{-2}\right] \\
&\leq 2\frac{2^{2k}}{c_X^2 n^2}.
\end{aligned}$$

Thus, with (5.67) we have

$$\mathbb{E}\left[\varepsilon_1 \varepsilon_2 U^{(\varepsilon)}_{n,r_n,\omega}(X_1) U^{(\varepsilon)}_{n,r_n,\omega}(X_2)\right] \leq \sigma^4 \mathbb{E}\left[\tilde{W}_1(X_1)^2\right] + \sigma^4 \mathbb{E}\left[\tilde{W}_1(X_2)^2\right] \leq 4\sigma^4 \frac{2^{2k}}{c_X^2 n^2}.$$

With (5.63) and (5.66) this implies

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{1}{n} \sum_{l=1}^{n} \varepsilon_l U^{(\varepsilon)}_{n,r_n,\omega}(X_l)\right|\right]^2 &\leq \frac{1}{n} \mathbb{E}\left[\varepsilon_1^2 U^{(\varepsilon)}_{n,r_n,\omega}(X_1)^2\right] + \frac{(n-1)}{n} \mathbb{E}\left[\varepsilon_1 \varepsilon_2 U^{(\varepsilon)}_{n,r_n,\omega}(X_1) U^{(\varepsilon)}_{n,r_n,\omega}(X_2)\right] \\
&= \mathcal{O}\left(\frac{2^k}{n^2}\right) + \mathcal{O}\left(\frac{2^{2k}}{n^2}\right) = \mathcal{O}\left(\frac{2^{2k}}{n^2}\right),
\end{aligned}$$

and hence with (5.61) and (5.62) we have

$$\mathbb{E}\left[E_2\right] = \mathcal{O}\left(n^{-1/2}\right) + \mathcal{O}\left(2^k/n\right). \tag{5.68}$$

The third expectation from (5.50) satisfies

$$\mathbb{E}\left[E_3\right] = \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\left(\varepsilon_i^2 - \sigma^2\right)\right|\right] \leq \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\right)^{1/2} = \frac{1}{\sqrt{n}}\|\varepsilon_1\|_4^2 = \mathcal{O}\left(n^{-1/2}\right).$$

With (5.50), (5.60) and (5.68) this finally yields

$$\begin{aligned}
&\mathbb{P}\left(|\hat{\sigma}^2 - \sigma^2| > \kappa\right) \\
&\quad\leq \frac{3}{\kappa}\mathbb{E}\left[E_1\right] + \frac{6}{\kappa}\mathbb{E}\left[E_2\right] + \frac{3}{\kappa}\mathbb{E}\left[E_3\right] \\
&\quad\leq \frac{C}{\kappa}\left(\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha]^2 + N_f(k)\binom{n}{r_n}^{-1} + (1 - c_X 2^{-k})^{r_n} + \frac{2^k}{n} + \frac{2^{2k}}{n^2} + n^{-1/2}\right)
\end{aligned}$$

for a suitable constant C. □

5.6.4 Proof of Corollary 5.5

We prove the corollary on the event that $\{\hat{N} > 0\}$. The case $\hat{N} = 0$ is not of interest as it corresponds to an empty random forest. We omit the indicator by bounding it by one whenever possible. The incomplete generalized U-statistic is

$$U_{n,r_n,N,\omega}^{(\mathrm{RF})}(x_0) = \frac{1}{\hat{N}}\sum_{I\in B_{r_n,n}} \rho_I \sum_{j\in I} Y_j W_{j,k}(x_0, I)$$

with

$$W_{j,k}(x_0, I) = \frac{\mathbb{I}\{X_j \in A_k(x_0, \omega_I)\}}{\sum_{i\in I}\mathbb{I}\{X_i \in A_k(x_0, \omega_I)\}}$$

from (3.10). We decompose its difference to the complete U-statistic

$$\begin{aligned}
&U_{n,r_n,N,\omega}^{(\mathrm{RF})}(x_0) - U_{n,r_n,\omega}^{(\mathrm{RF})}(x_0) \\
&\qquad= \frac{1}{\hat{N}}\sum_{I\in B_{r_n,n}} \rho_I \sum_{j\in I} Y_j W_{j,k}(x_0, I) - \frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}}\sum_{j\in I} Y_j W_{j,k}(x_0, I) \\
&\qquad= \left(\frac{1}{\hat{N}} - \frac{1}{N}\right)\sum_{I\in B_{r_n,n}} \rho_I \sum_{j\in I} Y_j W_{j,k}(x_0, I) \qquad (5.69) \\
&\qquad\quad + \sum_{I\in B_{r_n,n}}\left(\frac{\rho_I}{N} - \frac{1}{\binom{n}{r_n}}\right)\sum_{j\in I} Y_j W_{j,k}(x_0, I). \qquad (5.70)
\end{aligned}$$

We start with a bound that will be used for both terms. We exploit that the sum of the $W_{j,k}$ is either zero or one. In the first case the following bound holds trivially. We recall

the abbreviated notation $W_{j,k}(x_0, \omega)$ for the set $[r_n]$ from (3.10). In the second case we use Jensen's inequality, Lemma 2.14 and the bound for $p_{x_0}(\omega)$ from (4.6) to obtain

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{j=1}^{r_n} Y_j W_{j,k}(x_0,\omega)\right)^2\right] \\
&\leq \mathbb{E}\left[\sum_{j=1}^{r_n} Y_j^2 W_{j,k}(x_0,\omega)\right] \\
&= r_n \mathbb{E}\left[Y_1^2 \frac{\mathbb{I}\{X_1 \in A_k(x_0,\omega)\}}{\sum_{i=1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0,\omega)\}}\right] \\
&= r_n \mathbb{E}\left[(m(X_1)+\varepsilon_1)^2 \mathbb{I}\{X_1 \in A_k(x_0,\omega)\} \mathbb{E}\left[\frac{1}{1+\sum_{i=2}^{r_n} \mathbb{I}\{X_i \in A_k(x_0,\omega)\}} \mid X_1, \varepsilon_1, \omega\right]\right] \\
&\leq r_n 2 \mathbb{E}\left[(m(X_1)^2+\varepsilon_1^2)\mathbb{I}\{X_1 \in A_k(x_0,\omega)\} \frac{1}{r_n p_{x_0}(\omega)}\right] \\
&\leq 2 c_X^{-1} 2^k \mathbb{E}\left[(\|m\|_\infty^2 + \varepsilon_1^2)\mathbb{I}\{X_1 \in A_k(x_0,\omega)\}\right] \\
&= 2 c_X^{-1}(\|m\|_\infty^2 + \sigma^2). \qquad (5.71)
\end{aligned}$$

Using that the ρ_I are i.i.d. and independent of all the other random variables and (5.71) we get for the second term from (5.69) that

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{I \in B_{r_n,n}} \left(\rho_I \frac{1}{N} - \frac{1}{\binom{n}{r_n}}\right) \sum_{j \in I} Y_j W_{j,k}(x_0, I)\right)^2\right] \\
&= \sum_{I \in B_{r_n,n}} \mathbb{E}\left[\left(\rho_I \frac{1}{N} - \frac{1}{\binom{n}{r_n}}\right)^2\right] \mathbb{E}\left[\left(\sum_{j=1}^{r_n} Y_j W_{j,k}(x_0,\omega)\right)^2\right] \\
&= \binom{n}{r_n} \frac{1}{N^2} \mathrm{Var}(\rho_I) 2 c_X^{-1}(\|m\|_\infty^2 + \sigma^2) \\
&= \binom{n}{r_n} \frac{1}{N^2} \frac{N}{\binom{n}{r_n}} \left(1 - \frac{N}{\binom{n}{r_n}}\right) 2 c_X^{-1}(\|m\|_\infty^2 + \sigma^2) \\
&\leq \frac{1}{N} 2 c_X^{-1}(\|m\|_\infty^2 + \sigma^2) \\
&= \mathcal{O}(N^{-1}). \qquad (5.72)
\end{aligned}$$

With (5.72) and (5.71) we further obtain

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{I \in B_{r_n,n}} \rho_I \sum_{j \in I} Y_j W_{j,k}(x_0, I)\right)^2\right] \\
&\leq 2\mathbb{E}\left[\left(\sum_{I \in B_{r_n,n}} \left(\rho_I - \frac{N}{\binom{n}{r_n}}\right) \sum_{j \in I} Y_j W_{j,k}(x_0, I)\right)^2\right]
\end{aligned}$$

$$
\begin{aligned}
&\quad + 2\mathbb{E}\left[\left(\sum_{I\in B_{r_n,n}} \frac{N}{\binom{n}{r_n}} \sum_{j\in I} Y_j W_{j,k}(x_0, I)\right)^2\right] \\
&= 2N^2\mathbb{E}\left[\left(\sum_{I\in B_{r_n,n}} \left(\frac{\rho_I}{N} - \frac{1}{\binom{n}{r_n}}\right) \sum_{j\in I} Y_j W_{j,k}(x_0, I)\right)^2\right] \\
&\quad + 2N^2\mathbb{E}\left[\left(\frac{1}{\binom{n}{r_n}} \sum_{I\in B_{r_n,n}} \sum_{j\in I} Y_j W_{j,k}(x_0, I)\right)^2\right] \\
&= \mathcal{O}(N) + \mathcal{O}(N^2). \qquad (5.73)
\end{aligned}
$$

We note that $\hat{N} \sim \mathrm{Bin}(\binom{n}{r_n}, N/\binom{n}{r_n})$. Thus, for the first term from (5.69) we get with the Cauchy-Schwarz inequality, Lemma 2.18 and (5.73) that

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{I}\{\hat{N} > 0\}\left|\left(\frac{1}{\hat{N}} - \frac{1}{N}\right) \sum_{I\in B_{r_n,n}} \rho_I \sum_{j\in I} Y_j W_{j,k}(x_0, I)\right|\right]^2 \\
&\leq \mathbb{E}\left[\mathbb{I}\{\hat{N} > 0\}\left(\frac{1}{\hat{N}} - \frac{1}{N}\right)^2\right] \mathbb{E}\left[\left(\sum_{I\in B_{r_n,n}} \rho_I \sum_{j\in I} Y_j W_{j,k}(x_0, I)\right)^2\right] \\
&= \mathcal{O}(N^{-3})\mathcal{O}(N^2) \\
&= \mathcal{O}(N^{-1})
\end{aligned}
$$

Together with (5.69) and (5.72) this implies

$$
\mathbb{E}\left[\mathbb{I}\{\hat{N} > 0\}|U^{(\mathrm{RF})}_{n,r_n,N,\omega}(x_0) - U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0)|\right] = \mathcal{O}(N^{-1/2}).
$$

We obtain

$$
\begin{aligned}
&\mathbb{P}\left(\mathbb{I}\{\hat{N} > 0\}\sqrt{n} \sup_{x_0\in[0,1]^p} \Psi_k^{-1/2}(x_0)|U^{(\mathrm{RF})}_{n,r_n,N,\omega}(x_0) - U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0)| > (\log n)^{-1}\right) \\
&\leq \log n\mathbb{E}\left[\mathbb{I}\{\hat{N} > 0\}\sqrt{n} \sup_{x_0\in[0,1]^p} \Psi_k^{-1/2}(x_0)|U^{(\mathrm{RF})}_{n,r_n,N,\omega}(x_0) - U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0)|\right] \\
&\leq \log n\sqrt{\frac{n}{2^{2k}\mathcal{V}_{\cap,k}}} \sum_{x_0\in\mathcal{X}_k} \mathbb{E}\left[\mathbb{I}\{\hat{N} > 0\}|U^{(\mathrm{RF})}_{n,r_n,N,\omega}(x_0) - U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0)|\right] \\
&= \mathcal{O}\left(\log n\sqrt{\frac{n}{2^{2k}\mathcal{V}_{\cap,k}}}\frac{N_f(k)}{N^{1/2}}\right) \to 0.
\end{aligned}
$$

Including this in the decomposition in the proof of Proposition 5.15 yields

$$
\begin{aligned}
&\mathbb{I}\{\hat{N} > 0\}|\sqrt{n}\|\Psi_k^{-1/2}(U^{(\mathrm{RF})}_{n,r_n,N,\omega} - m)\|_\infty - \sigma\mathbf{S}_k| \\
&\quad = \mathcal{O}_{\mathbb{P}}\left(\frac{(\log n)^{3/2}}{\mathcal{V}_{\cap,k}^{1/2} n^{1/2-1/\nu}} + \frac{(\log n)^{5/4}}{\mathcal{V}_{\cap,k}^{1/4} n^{1/4}} + \frac{\log n}{\mathcal{V}_{\cap,k}^{1/6} n^{1/6}}\right) + o_{\mathbb{P}}\left((\log n)^{-1}\right).
\end{aligned}
$$

Using this instead of Proposition 5.15 in the proof of Theorem 5.1 yields the result. □

5.6.5 Proof of Corollary 5.6

We only consider the stochastic error from

$$\mathbb{E}\left[\left(U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0) - m(x_0)\right)^2\right] = \mathbb{E}\left[\left(U^{(m)}_{n,r_n,\omega}(x_0) - m(x_0)\right)^2\right] + \mathbb{E}\left[U^{(\varepsilon)}_{n,r_n,\omega}(x_0)^2\right].$$

For the approximation error we use Proposition 4.9, as we did in the proof of Proposition 4.1. With (5.34) and the decomposition from (5.46) we have

$$U^{(\varepsilon)}_{n,r_n,\omega}(x_0) = \hat{U}^{(\varepsilon)}_{n,r_n,\omega}(x_0) + R^{(1)}_{n,r_n,\omega}(x_0) + R^{(2)}_{n,r_n,\omega}(x_0).$$

The definition of K_k in (4.8) and (4.20) yield

$$\mathbb{E}\left[\hat{U}^{(\varepsilon)}_{n,r_n,\omega}(x_0)^2\right] = \frac{1}{n}\mathbb{E}\left[\varepsilon_1^2 K_k(x_0, X_1)^2\right] = \frac{\sigma^2}{n}\Psi_k(x_0) = \Theta\left(2^{2k}\mathcal{V}_{\cap,k}/n\right).$$

Lemma 5.18 implies that

$$\mathbb{E}\left[R^{(1)}_{n,r_n,\omega}(x_0)^2\right] = \mathcal{O}\left(\frac{2^{2k}}{r_n n}\right),$$

and Lemma 5.19 implies

$$\mathbb{E}\left[R^{(2)}_{n,r_n,\omega}(x_0)^2\right] = \mathcal{O}\left(\frac{2^k}{r_n}\left(\frac{r_n}{n}\right)^{r_n}\right).$$

Together this yields the claim of the corollary. □

5.6.6 Proof of Corollary 5.7

First, we prove that the leading term

$$\hat{U}^{(\varepsilon)}_{n,r_n,\omega}(x_0) = \frac{1}{n}\sum_{j=1}^{n}\varepsilon_j K_k(x_0, X_j)$$

converges in distribution to a normal distribution if it is appropriately standardized. The definition of Ψ_k in (4.9) implies

$$\mathrm{Var}\left(\varepsilon_j K_k(x_0, X_j)\right) = \sigma^2\Psi(x_0).$$

We apply the Lindeberg-Feller central limit theorem to $\hat{U}^{(\varepsilon)}_{n,r_n,\omega}(x_0)$. Before we prove the Lindeberg condition we note that (4.6) implies

$$\begin{aligned} K_k(x_0, X_1)^2 &= \mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega)\}p_{x_0}(\omega)^{-1} \mid X_1\right]^2 \\ &\leq \mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega)\} \mid X_1\right]^2 c_X^{-2}2^{2k}, \end{aligned}$$

and (4.20) yields $\Psi_k(x_0) = \Theta(2^{2k}\mathcal{V}_{\cap,k})$. Using these observations, the dominated convergence theorem yields that

$$\lim_{n\to\infty}\frac{1}{n\sigma^2\Psi_k(x_0)}\sum_{j=1}^{n}\mathbb{E}\left[\varepsilon_j^2 K_k^2(x_0, X_j)\mathbb{I}\left\{\varepsilon_j^2 K_k^2(x_0, X_j) > \kappa^2 n\sigma^2\Psi_k(x_0)\right\}\right]$$

$$
\begin{aligned}
&= \lim_{n\to\infty} \frac{1}{\sigma^2\Psi_k(x_0)}\mathbb{E}\left[\varepsilon_1^2 K_k^2(x_0,X_1)\mathbb{I}\left\{\sigma^{-2}\varepsilon_1^2 K_k^2(x_0,X_1)\Psi_k^{-1}(x_0) > \kappa^2 n\right\}\right] \\
&\lesssim \lim_{n\to\infty} \frac{1}{2^{2k}\mathcal{V}_{\cap,k}}\mathbb{E}\Big[\varepsilon_1^2\mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega)\} \mid X_1\right]^2 2^{2k} \\
&\qquad \times \mathbb{I}\left\{\sigma^{-2}\varepsilon_1^2\mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega)\} \mid X_1\right]^2 \mathcal{V}_{\cap,k}^{-1} > \kappa^2 n\right\}\Big] \\
&\leq \lim_{n\to\infty} \frac{1}{\mathcal{V}_{\cap,k}}\mathbb{E}\left[\varepsilon_1^2\mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega)\} \mid X_1\right]^2 \mathbb{I}\left\{\varepsilon_1^2 > \kappa^2\sigma^2\mathcal{V}_{\cap,k} n\right\}\right] \\
&= \lim_{n\to\infty} \frac{1}{\mathcal{V}_{\cap,k}}\mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0,\omega)\} \mid X_1\right]^2\right]\mathbb{E}\left[\varepsilon_1^2\mathbb{I}\left\{\varepsilon_1^2 > \kappa^2\sigma^2\mathcal{V}_{\cap,k} n\right\}\right] \\
&= \lim_{n\to\infty} \mathbb{E}\left[\varepsilon_1^2\mathbb{I}\left\{\varepsilon_1^2 > \kappa^2\sigma^2\mathcal{V}_{\cap,k} n\right\}\right] \to 0,
\end{aligned}
$$

because $n\mathcal{V}_{\cap,k} \to \infty$ is implied by (5.18) and $\mathbb{E}[\varepsilon_1^2] < \infty$. Thus, the Lindeberg condition is fulfilled and we obtain

$$
\sqrt{\frac{n}{\sigma^2\Psi_k(x_0)}}\hat{U}_{n,r_n,\omega}^{(\varepsilon)}(x_0) \to \mathcal{N}(0,1). \tag{5.74}
$$

It remains to show that the remainder terms are negligible. We need to prove that all other terms are $o_{\mathbb{P}}(\sqrt{2^{2k}\mathcal{V}_{\cap,k}/n})$ because $\Psi_k(x_0) = \Theta(2^{2k}\mathcal{V}_{\cap,k})$. Proposition 4.9, (5.20) and (5.19) yield

$$
\begin{aligned}
\mathbb{P}\left(|U_{n,r_n,\omega}^{(m)}(x_0) - m(x_0)| > \kappa\sqrt{2^{2k}\mathcal{V}_{\cap,k}/n}\right) &\leq \mathbb{E}\left[|U_{n,r_n,\omega}^{(m)}(x_0) - m(x_0)|\right]\sqrt{\frac{n}{2^{2k}\mathcal{V}_{\cap,k}\kappa^2}} \\
&\leq \sqrt{\frac{n}{2^{2k}\mathcal{V}_{\cap,k}\kappa^2}} C_H \mathbb{E}\left[\mathfrak{d}(A_k(x_0,\omega))^\alpha\right] \\
&\quad + \sqrt{\frac{n}{2^{2k}\mathcal{V}_{\cap,k}\kappa^2}}|m(x_0)|(1 - c_X 2^{-k})^{r_n} \to 0,
\end{aligned} \tag{5.75}
$$

for any $\kappa > 0$. Using that $\mathbb{E}[\varepsilon_1^2] < \infty$, Lemma 5.18 and (5.18) imply for any $\kappa > 0$ that

$$
\begin{aligned}
\mathbb{P}\left(|R_{n,r_n,\omega}^{(1)}(x_0)| > \kappa\sqrt{2^{2k}\mathcal{V}_{\cap,k}/n}\right) &\leq \mathbb{E}\left[R_{n,r_n,\omega}^{(1)}(x_0)^2\right]\frac{n}{2^{2k}\mathcal{V}_{\cap,k}\kappa^2} \\
&= \mathcal{O}\left(r_n^{-1}\mathcal{V}_{\cap,k}^{-1}\right) \to 0.
\end{aligned} \tag{5.76}
$$

With Lemma 5.19 we obtain

$$
\begin{aligned}
\mathbb{P}\left(|R_{n,r_n,\omega}^{(2)}(x_0)| > \kappa\sqrt{2^{2k}\mathcal{V}_{\cap,k}/n}\right) &\leq \mathbb{E}\left[R_{n,r_n,\omega}^{(2)}(x_0)^2\right]\frac{n}{2^{2k}\mathcal{V}_{\cap,k}\kappa^2} \\
&\leq \frac{\sigma^2 C_X}{c_X^2}\frac{2^k}{r_n}\left(\frac{r_n}{n}\right)^{r_n}\frac{n}{2^{2k}\mathcal{V}_{\cap,k}\kappa^2} \\
&= \mathcal{O}\left(\left(\frac{r_n}{n}\right)^{r_n-1} 2^{-k}\mathcal{V}_{\cap,k}^{-1}\right).
\end{aligned}
$$

Together (3.20) and (5.18) imply that $2^k/r_n \to 0$. Using the assumption $r_n/n \leq c < 1$ and $\mathcal{V}_{\cap,k} \geq 2^{-2k}$ this yields

$$\mathcal{O}\left(\left(\frac{r_n}{n}\right)^{r_n-1} 2^{-k}\mathcal{V}_{\cap,k}^{-1}\right) = \mathcal{O}\left(c^{r_n-1}2^k\right) = \mathcal{O}\left(\exp\left((r_n-1)\log c + k\log 2\right)\right) = o(1).$$

In combination with (5.74), (5.75) and (5.76) we obtain

$$\sqrt{\frac{n}{\sigma^2\Psi(x_0)}}\left(U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0) - m(x_0)\right) \xrightarrow{d} \mathcal{N}(0,1),$$

which is the first assertion of the corollary. To prove the second assertion, it remains to show that

$$\sqrt{\frac{n}{2^{2k}\mathcal{V}_{\cap,k}}}\left(U^{(\mathrm{RF})}_{n,r_n,N,\omega}(x_0) - U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0)\right) \xrightarrow{\mathbb{P}} 0.$$

We use the decomposition

$$\begin{aligned}
&U^{(\mathrm{RF})}_{n,r_n,N,\omega}(x_0) - U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0) \\
&\quad = \left(\frac{1}{\hat{N}} - \frac{1}{N}\right)\sum_{I\in B_{r_n,n}} \rho_I \sum_{j\in I} Y_j W_{j,k}(x_0, I) + \sum_{I\in B_{r_n,n}} \left(\frac{\rho_I}{N} - \frac{1}{\binom{n}{r_n}}\right)\sum_{j\in I} Y_j W_{j,k}(x_0, I)
\end{aligned}$$

from (5.69) in the proof of Corollary 5.5. For the second term, (5.72) from the same proof yields

$$\mathbb{E}\left[\left|\sum_{I\in B_{r_n,n}}\left(\rho_I\frac{1}{N} - \frac{1}{\binom{n}{r_n}}\right)\sum_{j\in I} Y_j W_{j,k}(x_0, I)\right|\right] = \mathcal{O}(N^{-1/2}),$$

and thus, (5.21) implies

$$\sqrt{\frac{n}{2^{2k}\mathcal{V}_{\cap,k}}}\sum_{I\in B_{r_n,n}}\left(\frac{\rho_I}{N} - \frac{1}{\binom{n}{r_n}}\right)\sum_{j\in I} Y_j W_{j,k}(x_0, I) \xrightarrow{\mathbb{P}} 0. \tag{5.77}$$

For the first term from the decomposition we apply the Cauchy-Schwarz inequality and (5.73) implies

$$\begin{aligned}
&\mathbb{E}\left[\left|\left(\frac{1}{N} - \frac{\hat{N}}{N^2}\right)\sum_{I\in B_{r_n,n}} \rho_I \sum_{j\in I} Y_j W_{j,k}(x_0, I)\right|\right]^2 \\
&\quad \leq \mathbb{E}\left[\left(\frac{1}{N} - \frac{\hat{N}}{N^2}\right)^2\right]\mathbb{E}\left[\left(\sum_{I\in B_{r_n,n}} \rho_I \sum_{j\in I} Y_j W_{j,k}(x_0, I)\right)^2\right] \\
&\quad = \frac{1}{N^4}\mathbb{E}\left[\left(N - \hat{N}\right)^2\right]\mathcal{O}(N^2) \\
&\quad = \mathcal{O}\left(\binom{n}{r_n}^{-1}\right) \\
&\quad = \mathcal{O}\left(N^{-1}\right),
\end{aligned}$$

because $\hat{N} \sim \mathrm{Bin}(\binom{n}{r_n}, N/\binom{n}{r_n})$. Noting that $N/\hat{N} \xrightarrow{\mathbb{P}} 1$, Slutsky's theorem and (5.21) yield

$$\sqrt{\frac{n}{2^{2k}\mathcal{V}_{\cap,k}}}\frac{N}{\hat{N}}\left(\frac{1}{N}-\frac{\hat{N}}{N^2}\right)\sum_{I\in B_{r_n,n}}\rho_I\sum_{j\in I}Y_jW_{j,k}(x_0,I)\xrightarrow{\mathbb{P}}0.$$

Together with the decomposition above and (5.77), this proves the second assertion of the corollary. □

5.6.7 Proof of Corollary 5.8

Using the triangle inequality we have

$$\begin{aligned}\|U^{(\mathrm{RF})}_{n,r_n,\omega}-m\|_\infty &\le \|\hat{U}^{(\varepsilon)}_{n,r_n,\omega}\|_\infty+\|\hat{U}^{(\varepsilon)}_{n,r_n,\omega}-U^{(\varepsilon)}_{n,r_n,\omega}\|_\infty+\|U^{(m)}_{n,r_n,\omega}-m\|_\infty\\ &\le \|\Psi_k^{1/2}\|_\infty\|\Psi_k^{-1/2}\hat{U}^{(\varepsilon)}_{n,r_n,\omega}\|_\infty+\|\hat{U}^{(\varepsilon)}_{n,r_n,\omega}-U^{(\varepsilon)}_{n,r_n,\omega}\|_\infty+\|U^{(m)}_{n,r_n,\omega}-m\|_\infty\\ &\le \|\Psi_k^{1/2}\|_\infty\sqrt{\frac{\sigma^2}{n}}\mathbf{S}_k+\|\Psi_k^{1/2}\|_\infty\left|\|\Psi_k^{-1/2}\hat{U}^{(\varepsilon)}_{n,r_n,\omega}\|_\infty-\sqrt{\frac{\sigma^2}{n}}\mathbf{S}_k\right|\\ &\quad+\|R^{(1)}_{n,r_n,\omega}\|_\infty+\|R^{(2)}_{n,r_n,\omega}\|_\infty+\|U^{(m)}_{n,r_n,\omega}-m\|_\infty. \end{aligned}\tag{5.78}$$

We handle the terms one by one. We note that (4.20) implies

$$\|\Psi_k^{1/2}\|_\infty=\Theta(2^k\mathcal{V}_{\cap,k}^{1/2}).\tag{5.79}$$

With (5.40) we get

$$\mathbb{P}\left(\mathbf{S}_k>\kappa\right)\le\frac{1}{\kappa}\mathbb{E}\left[\mathbf{S}_k\right]\lesssim\frac{1}{\kappa}\sqrt{k},$$

and hence $\mathbf{S}_k=\mathcal{O}_{\mathbb{P}}(\sqrt{k})$. Thus, the first term satisfies

$$\|\Psi_k^{1/2}\|_\infty\sqrt{\frac{\sigma^2}{n}}\mathbf{S}_k=\mathcal{O}\left(\sqrt{2^{2k}\mathcal{V}_{\cap,k}k/n}\right).\tag{5.80}$$

Theorem 5.14 and (5.79) yield

$$\begin{aligned}&\|\Psi_k^{1/2}\|_\infty\left|\|\Psi_k^{-1/2}\hat{U}^{(\varepsilon)}_{n,r_n,\omega}\|_\infty-\sqrt{\frac{\sigma^2}{n}}\mathbf{S}_k\right|\\ &\quad=\sqrt{\sigma^2/n}\|\Psi_k^{1/2}\|_\infty\left|\sqrt{n/\sigma^2}\|\Psi_k^{-1/2}\hat{U}^{(\varepsilon)}_{n,r_n,\omega}\|_\infty-\mathbf{S}_k\right|\\ &\quad=\Theta(n^{-1/2}2^k\mathcal{V}_{\cap,k}^{1/2})\mathcal{O}_{\mathbb{P}}\left(\frac{(\log n)^{3/2}}{\mathcal{V}_{\cap,k}^{1/2}n^{1/2-1/\nu}}+\frac{(\log n)^{5/4}}{\mathcal{V}_{\cap,k}^{1/4}n^{1/4}}+\frac{\log n}{\mathcal{V}_{\cap,k}^{1/6}n^{1/6}}\right)\\ &\quad=\mathcal{O}_{\mathbb{P}}\left(2^k\left(\frac{(\log n)^{3/2}}{n^{1-1/\nu}}+\frac{\mathcal{V}_{\cap,k}^{1/4}(\log n)^{5/4}}{n^{3/4}}+\frac{\mathcal{V}_{\cap,k}^{1/3}\log n}{n^{2/3}}\right)\right).\end{aligned}\tag{5.81}$$

Lemma 5.18 yields for q_R with $\mathbb{E}[|\varepsilon_1|^{q_R}]<\infty$ that

$$\mathbb{P}\left(\|R^{(1)}_{n,r_n,\omega}\|_\infty>\kappa\right)\le\sum_{x_0\in\mathcal{X}_k}\frac{1}{\kappa^{q_R}}\mathbb{E}\left[\left(R^{(1)}_{n,r_n,\omega}(x_0)\right)^{q_R}\right]\le N_f(k)\frac{1}{\kappa^{q_R}}C\left(\frac{2^{2k}}{r_nn}\right)^{q_R/2}.$$

For $\epsilon > 0$ let $\kappa > (\epsilon/C)^{-1/q_R}$, we obtain

$$\mathbb{P}\left(\|R^{(1)}_{n,r_n,\omega}\|_\infty > \kappa \sqrt{\frac{2^{2k}}{r_n n}} N_f(k)^{1/q_R} \right) \leq C \frac{1}{\kappa^{q_R}} < \epsilon,$$

and subsequently we have

$$\|R^{(1)}_{n,r_n,\omega}\|_\infty = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{2^{2k}}{r_n n}} N_f(k)^{1/q_R} \right). \tag{5.82}$$

With Lemma 5.19, $N_f(k) \leq 2^{kp}$, $\mathcal{V}_{\cap,k} \geq 2^{-2k}$ and the assumption $r_n/n \leq c < 1$ we get

$$\begin{aligned}
&\mathbb{P}\left(\|R^{(2)}_{n,r_n,\omega}\|_\infty > \sqrt{2^{2k}\mathcal{V}_{\cap,k} k/n} \right) \\
&\leq \frac{n}{2^{2k}\mathcal{V}_{\cap,k} k} \mathbb{E}\left[\|R^{(2)}_{n,r_n,\omega}\|^2_\infty \right] \\
&\leq \frac{n}{2^{2k}\mathcal{V}_{\cap,k} k} \sum_{x_0 \in \mathcal{X}_k} \mathbb{E}\left[R^{(2)}_{n,r_n,\omega}(x_0)^2 \right] \\
&\leq \frac{n}{2^{2k}\mathcal{V}_{\cap,k} k} N_f(k) \frac{\sigma^2 C_X}{c_X^2} \frac{2^k}{r_n} \left(\frac{r_n}{n} \right)^{r_n} \\
&\leq \frac{\sigma^2 C_X}{c_X^2} 2^{k(p+1)} \frac{1}{k} c^{r_n - 1} \\
&= \frac{\sigma^2 C_X}{c_X^2} \exp\left(k(p+1)\log 2 - \log k + (r_n - 1)\log c \right) \to 0
\end{aligned}$$

because we assumed that $r_n/k \to \infty$. Hence

$$\|R^{(2)}_{n,r_n,\omega}\|_\infty = o_{\mathbb{P}}\left(\sqrt{2^{2k}\mathcal{V}_{\cap,k} k/n} \right) \tag{5.83}$$

which is negligible compared to (5.80). We note that

$$\|U^{(m)}_{n,r_n,\omega} - m\|_\infty \leq \|U^{(m)}_{n,r_n,\omega} - m U^{(1)}_{n,r_n,\omega}\|_\infty + \|m(U^{(1)}_{n,r_n,\omega} - 1)\|_\infty.$$

Let $\epsilon > 0$ be arbitrary and choose $\kappa > C_H/\epsilon$. Lemma 5.16 yields

$$\begin{aligned}
&\mathbb{P}\left(\|U^{(m)}_{n,r_n,\omega} - m U^{(1)}_{n,r_n,\omega}\|_\infty > \kappa\big(\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha] + p^{\alpha/2} N_f(k) \binom{n}{r_n}^{-1/2} \big) \right) \\
&\leq \frac{\mathbb{E}\left[\|U^{(m)}_{n,r_n,\omega} - m U^{(1)}_{n,r_n,\omega}\|_\infty \right]}{\kappa\big(\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha] + p^{\alpha/2} N_f(k) \binom{n}{r_n}^{-1/2} \big)} \\
&\leq \frac{C_H}{\kappa} < \epsilon
\end{aligned}$$

and thus,

$$\|U^{(m)}_{n,r_n,\omega} - mU^{(1)}_{n,r_n,\omega}\|_\infty = \mathcal{O}_\mathbb{P}\left(\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha] + N_f(k)\binom{n}{r_n}^{-1/2}\right). \tag{5.84}$$

Lemma 5.17 directly implies

$$\|m(U^{(1)}_{n,r_n,\omega} - 1)\|_\infty = \mathcal{O}_\mathbb{P}(2^k(1 - c_X 2^{-k})^{r_n}). \tag{5.85}$$

In conjunction (5.78), (5.80), (5.81), (5.82), (5.83), (5.84) and (5.85) yield

$$\begin{aligned}
\|U^{(\mathrm{RF})}_{n,r_n,\omega} - m\|_\infty \leq{}& \|\Psi_k^{1/2}\|_\infty \left| \|\Psi_k^{-1/2}\hat{U}^{(\varepsilon)}_{n,r_n,\omega}\|_\infty - \sqrt{\frac{\sigma^2}{n}}\mathbf{S}_k \right| \\
&+ \sqrt{\frac{\sigma^2}{n}}\mathbf{S}_k\|\Psi_k^{1/2}\|_\infty + \|R^{(1)}_{n,r_n,\omega}\|_\infty + \|R^{(2)}_{n,r_n,\omega}\|_\infty + \|U^{(m)}_{n,r_n,\omega} - m\|_\infty \\
={}& \mathcal{O}_\mathbb{P}\left(2^k\left(\frac{(\log n)^{3/2}}{n^{1-1/\nu}} + \frac{\mathcal{V}_{\cap,k}^{1/4}(\log n)^{5/4}}{n^{3/4}} + \frac{\mathcal{V}_{\cap,k}^{1/3}\log n}{n^{2/3}}\right)\right) \\
&+ \mathcal{O}\left(\sqrt{\frac{2^{2k}\mathcal{V}_{\cap,k}k}{n}}\right) + \mathcal{O}_\mathbb{P}\left(\sqrt{\frac{2^{2k}}{r_n n}}N_f(k)^{1/q_R}\right) \\
&+ \mathcal{O}_\mathbb{P}\left(\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha] + N_f(k)\binom{n}{r_n}^{-1/2}\right) + \mathcal{O}_\mathbb{P}(2^k(1 - c_X 2^{-k})^{r_n}),
\end{aligned}$$

which completes the proof. □

5.6.8 Proof of Theorem 5.9

The proof strategy is similar to the one for the normal RF. We use the decomposition

$$\begin{aligned}
U^{(\mathrm{KRF})}_{n,r_n,\omega}(x_0) &= \frac{\sum_{I\in B_{r_n,n}}\sum_{j\in I} m(X_j)\mathbb{I}\{X_j \in A_k(x_0,\omega_I)\}}{\sum_{I\in B_{r_n,n}}\sum_{j\in I}\mathbb{I}\{X_j \in A_k(x_0,\omega_I)\}} \\
&\quad + \frac{\sum_{I\in B_{r_n,n}}\sum_{j\in I}\varepsilon_j\mathbb{I}\{X_j \in A_k(x_0,\omega_I)\}}{\sum_{I\in B_{r_n,n}}\sum_{j\in I}\mathbb{I}\{X_j \in A_k(x_0,\omega_I)\}} \\
&=: U^{(K,m)}_{n,r_n,\omega}(x_0) + U^{(K,\varepsilon)}_{n,r_n,\omega}(x_0).
\end{aligned} \tag{5.86}$$

The different structure leads to different but similar proof methods for the approximation error and the remainder terms. We will need the following corollary, which is similar to Lemma 5.19 and whose proof can be found in Section 5.6.10.3.

Corollary 5.20. *For*

$$R^{(K)}_{n,r_n,\omega}(x_0) = \frac{1}{np_{x_0}}\sum_{j=1}^n \varepsilon_j \frac{1}{\binom{n-1}{r_n-1}} \sum_{I\in B_{r_n,n}: j\in I} (\mathbb{I}\{X_j \in A_k(x_0,\omega_I)\} - \mathbb{P}(X_j \in A_k(x_0,\omega) \mid X_j))$$

it holds that

$$\mathbb{E}\left[R^{(K)}_{n,r_n,\omega}(x_0)^2\right] \leq \frac{\sigma^2}{c_X}\frac{2^k}{r_n}\left(\frac{r_n}{n}\right)^{r_n}.$$

Proof of Theorem 5.9. We will prove that

$$\begin{aligned}&\left|\sqrt{n}\|\Phi_k^{-1/2}(U_{n,r_n,\omega}^{(\mathrm{KRF})}-m)\|_\infty-\sigma\mathbf{S}_k\right|\\&\quad=\mathcal{O}_{\mathbb{P}}\left(\frac{(\log n)^{3/2}}{\mathcal{V}_{\cap,k}^{1/2}n^{1/2-1/\nu}}+\frac{(\log n)^{5/4}}{\mathcal{V}_{\cap,k}^{1/4}n^{1/4}}+\frac{\log n}{\mathcal{V}_{\cap,k}^{1/6}n^{1/6}}\right)+o_{\mathbb{P}}\left((\log n)^{-1}\right).\end{aligned}$$

Then the claim follows directly analogue as in the proof of Theorem 5.1, the only difference being the different Gaussian process. Let us define

$$\check{U}_{n,r_n,\omega}^{(K,\varepsilon)}(x_0)=\frac{1}{r_np_{x_0}}\frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}}\sum_{j\in I}\varepsilon_j\mathbb{I}\{X_j\in A_k(x_0,\omega_I)\}. \tag{5.87}$$

This leads to

$$\begin{aligned}&U_{n,r_n,\omega}^{(K,\varepsilon)}(x_0)-\check{U}_{n,r_n,\omega}^{(K,\varepsilon)}(x_0)\\&\quad=\sum_{I\in B_{r_n,n}}\sum_{j\in I}\varepsilon_j\mathbb{I}\{X_j\in A_k(x_0,\omega_I)\}\left(\frac{1}{\sum_{I\in B_{r_n,n}}\sum_{j\in I}\mathbb{I}\{X_j\in A_k(x_0,\omega_I)\}}-\frac{1}{\binom{n}{r_n}r_np_{x_0}}\right)\\&\quad=\check{U}_{n,r_n,\omega}^{(K,\varepsilon)}(x_0)\left(r_n\binom{n}{r_n}p_{x_0}\frac{1}{\sum_{I\in B_{r_n,n}}\sum_{j\in I}\mathbb{I}\{X_j\in A_k(x_0,\omega_I)\}}-1\right)\\&\quad=\check{U}_{n,r_n,\omega}^{(K,\varepsilon)}(x_0)\delta_n(x_0),\end{aligned} \tag{5.88}$$

for

$$\delta_n(x_0):=\left(\frac{1}{r_n\binom{n}{r_n}p_{x_0}}\sum_{I\in B_{r_n,n}}\sum_{j\in I}\mathbb{I}\{X_j\in A_k(x_0,\omega_I)\}\right)^{-1}-1. \tag{5.89}$$

Further we denote

$$\hat{U}_{n,r_n,\omega}^{(K,\varepsilon)}(x_0)=\frac{1}{np_{x_0}}\sum_{j=1}^{n}\varepsilon_j\mathbb{P}(X_j\in A_k(x_0,\omega)\mid X_j)$$

and with (5.87) we obtain

$$\begin{aligned}&\check{U}_{n,r_n,\omega}^{(K,\varepsilon)}(x_0)-\hat{U}_{n,r_n,\omega}^{(K,\varepsilon)}(x_0)\\&\qquad=\frac{1}{np_{x_0}}\sum_{j=1}^{n}\varepsilon_j\left(\frac{1}{\binom{n-1}{r_n-1}}\sum_{I\in B_{r_n,n}:j\in I}\mathbb{I}\{X_j\in A_k(x_0,\omega_I)\}-\mathbb{P}(X_j\in A_k(x_0,\omega)\mid X_j)\right)\\&\qquad=:R_{n,r_n,\omega}^{(K)}(x_0).\end{aligned} \tag{5.90}$$

(5.86), (5.88) and (5.90) lead to

$$\begin{aligned}U_{n,r_n,\omega}^{(\mathrm{KRF})}(x_0)-m(x_0)&=U_{n,r_n,\omega}^{(K,m)}(x_0)-m(x_0)+U_{n,r_n,\omega}^{(K,\varepsilon)}(x_0)\\&=U_{n,r_n,\omega}^{(K,m)}(x_0)-m(x_0)+\check{U}_{n,r_n,\omega}^{(\varepsilon)}(x_0)\left(1+\delta_n(x_0)\right)\\&=U_{n,r_n,\omega}^{(K,m)}(x_0)-m(x_0)+R_{n,r_n,\omega}^{(K)}(x_0)\left(1+\delta_n(x_0)\right)\end{aligned}$$

$$+ \hat{U}^{(K,\varepsilon)}_{n,r_n,\omega}(x_0)\,(1+\delta_n(x_0))\,.$$

With the same arguments that lead to (4.20) for Ψ_k, we get for Φ_k from (5.24) that

$$\frac{c_X}{C_X^2} 2^{2k}\mathcal{V}_{\cap,k} \le \Phi_k(x_0) \le \frac{C_X}{c_X^2} 2^{2k}\mathcal{V}_{\cap,k}$$

holds uniformly in x_0. Using this we obtain with the same arguments as in Section 5.6.2 that

$$\begin{aligned}
&|\sqrt{n}\|\Phi_k^{-1/2}(U^{(\mathrm{KRF})}_{n,r_n,\omega} - m)\|_\infty - \sigma\mathbf{S}_k| \\
&\quad\le \sqrt{n}\big(\|\Phi_k^{-1/2}(U^{(K,m)}_{n,r_n,\omega} - m)\|_\infty + \|\Phi_k^{-1/2}R^{(K)}_{n,r_n,\omega}(1+\delta_n)\|_\infty + \|\Phi_k^{-1/2}\hat{U}^{(K,\varepsilon)}_{n,r_n,\omega}\delta_n\|_\infty\big) \\
&\qquad + |\sqrt{n}\|\Phi_k^{-1/2}\hat{U}^{(K,\varepsilon)}_{n,r_n,\omega}\|_\infty - \sigma\mathbf{S}_k| \\
&\quad\lesssim \sqrt{\frac{n}{2^{2k}\mathcal{V}_{\cap,k}}}\big(\|U^{(K,m)}_{n,r_n,\omega} - m\|_\infty + \|R^{(K)}_{n,r_n,\omega}\|_\infty(1+\|\delta_n\|_\infty)\big) \\
&\qquad + \big|\sqrt{n}\|\Phi_k^{-1/2}\hat{U}^{(K,\varepsilon)}_{n,r_n,\omega}\|_\infty - \sigma\mathbf{S}_k\big|(1+\|\delta_n\|_\infty) + \sigma\mathbf{S}_k\|\delta_n\|_\infty. \qquad (5.91)
\end{aligned}$$

We will bound these terms one by one. Let

$$U^{(K,1)}_{n,r_n,\omega}(x_0) := \mathbb{I}\left\{\exists I \in B_{r_n,n} : \exists j \in I : X_j \in A_k(x_0,\omega_I)\right\} \qquad (5.92)$$

We get the decomposition

$$U^{(K,m)}_{n,r_n,\omega}(x_0) - m(x_0) = U^{(K,m)}_{n,r_n,\omega}(x_0) - m(x_0)U^{(K,1)}_{n,r_n,\omega}(x_0) + m(x_0)(U^{(K,1)}_{n,r_n,\omega}(x_0) - 1).$$

For the first term we obtain

$$\begin{aligned}
&|U^{(K,m)}_{n,r_n,\omega}(x_0) - m(x_0)U^{(K,1)}_{n,r_n,\omega}(x_0)| \\
&\qquad = \left|\frac{\sum_{I\in B_{r_n,n}}\sum_{j\in I}(m(X_j)-m(x_0))\mathbb{I}\{X_j\in A_k(x_0,\omega_I)\}}{\sum_{I\in B_{r_n,n}}\sum_{j\in I}\mathbb{I}\{X_j\in A_k(x_0,\omega_I)\}}\right| \\
&\qquad \le C_H\frac{\sum_{I\in B_{r_n,n}}\sum_{j\in I}\mathfrak{d}(A_k(x_0,\omega_I))^\alpha\mathbb{I}\{X_j\in A_k(x_0,\omega_I)\}}{\sum_{I\in B_{r_n,n}}\sum_{j\in I}\mathbb{I}\{X_j\in A_k(x_0,\omega_I)\}} \\
&\qquad = C_H\sum_{I\in B_{r_n,n}}\mathfrak{d}(A_k(x_0,\omega_I))^\alpha\frac{\sum_{j\in I}\mathbb{I}\{X_j\in A_k(x_0,\omega_I)\}}{\sum_{J\in B_{r_n,n}}\sum_{j\in J}\mathbb{I}\{X_j\in A_k(x_0,\omega_J)\}} \\
&\qquad \le C_H\max_{I\in B_{r_n,n}}\mathfrak{d}(A_k(x_0,\omega_I))^\alpha.
\end{aligned}$$

This yields

$$\begin{aligned}
&\mathbb{P}\left(\|U^{(K,m)}_{n,r_n,\omega} - mU^{(K,1)}_{n,r_n,\omega}\|_\infty > \kappa(\log n)^{-1}\sqrt{\frac{2^{2k}\mathcal{V}_{\cap,k}}{n}}\right) \\
&\quad\le \kappa^{-1}\left(\frac{n(\log n)^2}{2^{2k}\mathcal{V}_{\cap,k}}\right)^{1/2}\mathbb{E}\left[\sup_{x_0\in[0,1]^p}\max_{I\in B_{r_n,n}}\mathfrak{d}(A_k(x_0,\omega_I))^\alpha\right] \to 0 \qquad (5.93)
\end{aligned}$$

due to (5.25). With (5.92) we have

$$\begin{aligned}|U^{(K,1)}_{n,r_n,\omega}(x_0) - 1| &= \mathbb{I}\big\{\nexists I \in B_{r_n,n} : \exists j \in I : X_j \in A_k(x_0,\omega_I)\big\}\\ &= \mathbb{I}\big\{\forall I \in B_{r_n,n} : \nexists j \in I : X_j \in A_k(x_0,\omega_I)\big\}\\ &= \prod_{I\in B_{r_n,n}} \mathbb{I}\big\{\nexists j \in I : X_j \in A_k(x_0,\omega_I)\big\}\\ &= \min_{I\in B_{r_n,n}} \mathbb{I}\big\{\nexists j \in I : X_j \in A_k(x_0,\omega_I)\big\}.\end{aligned}$$

We obtain

$$\begin{aligned}&\mathbb{E}\left[\sup_{x_0\in[0,1]^p} |m(x_0)(U^{(K,1)}_{n,r_n,\omega}(x_0) - 1)|\right]\\ &\le \|m\|_\infty \mathbb{E}\left[\sup_{x_0\in[0,1]^p} \min_{I\in B_{r_n,n}} \mathbb{I}\big\{\nexists j \in I : X_j \in A_k(x_0,\omega_I)\big\}\right]\\ &\le \|m\|_\infty \mathbb{E}\left[\sup_{x_0\in[0,1]^p} \mathbb{I}\big\{\nexists j \in [r_n] : X_j \in A_k(x_0,\omega)\big\}\right]\\ &\le \|m\|_\infty 2^k (1 - c_X 2^{-k})^{r_n},\end{aligned}$$

where the last inequality follows from the proof of Lemma 5.17. Hence

$$\begin{aligned}&\mathbb{P}\left(\|m(U^{(K,1)}_{n,r_n,\omega} - 1)\|_\infty > \kappa(\log n)^{-1}\sqrt{\frac{2^{2k}\mathcal{V}_{\cap,k}}{n}}\right)\\ &\le \kappa^{-1}\left(\frac{n(\log n)^2}{2^{2k}\mathcal{V}_{\cap,k}}\right)^{1/2} \|m\|_\infty 2^k(1 - c_X 2^{-k})^{r_n} \to 0 \qquad (5.94)\end{aligned}$$

by (5.13). In conjunction (5.93) and (5.94) yield

$$\sqrt{\frac{n}{2^{2k}\mathcal{V}_{\cap,k}}}\|U^{(K,m)}_{n,r_n,\omega} - m\|_\infty = o_{\mathbb{P}}\left((\log n)^{-1}\right). \qquad (5.95)$$

For δ_n from (5.89) we will prove that

$$\|\delta_n\|_\infty = o_{\mathbb{P}}\left((\log n)^{-3/2}\right).$$

For $g(x) = 1/x$ and

$$\begin{aligned}\xi(x_0) &= \frac{1}{r_n\binom{n}{r_n}p_{x_0}} \sum_{I\in B_{r_n,n}} \sum_{j\in I} \mathbb{I}\{X_j \in A_k(x_0,\omega_I)\}\\ &= \frac{1}{r_n\binom{n}{r_n}p_{x_0}} \sum_{I\in B_{r_n,n}} \sum_{j\in I} \left(\mathbb{I}\{X_j \in A_k(x_0,\omega_I)\} - p_{x_0}(\omega_I)\right)\\ &\quad + \frac{1}{r_n\binom{n}{r_n}p_{x_0}} \sum_{I\in B_{r_n,n}} \sum_{j\in I} p_{x_0}(\omega_I)\end{aligned}$$

we have for $\kappa > 0$

$$\begin{aligned}
&\mathbb{P}\left(\sup_{x_0\in[0,1]^p} |\delta_n(x_0)| > \kappa\right) \\
&\leq \sum_{x_0\in\mathcal{X}_k} \mathbb{P}(|\delta_n(x_0)| > \kappa) \\
&= \sum_{x_0\in\mathcal{X}_k} \mathbb{P}\left(|g\left(\xi(x_0)\right) - g(1)| > \kappa\right) \\
&= \sum_{x_0\in\mathcal{X}_k} \mathbb{P}\left(g\left(\xi(x_0)\right) \notin [1-\kappa, 1+\kappa]\right) \\
&= \sum_{x_0\in\mathcal{X}_k} \mathbb{P}\left(\xi(x_0) \notin [g^{-1}(1+\kappa), g^{-1}(1-\kappa)]\right) \\
&\leq \sum_{x_0\in\mathcal{X}_k} \mathbb{P}\left(|\xi(x_0) - 1| > |1 - g^{-1}(1+\kappa)|\right) \\
&\leq \sum_{x_0\in\mathcal{X}_k} \mathbb{P}\left(\left|\frac{1}{r_n\binom{n}{r_n}p_{x_0}} \sum_{I\in B_{r_n,n}} \sum_{j\in I} \left(\mathbb{I}\{X_j \in A_k(x_0,\omega_I)\} - p_{x_0}(\omega_I)\right)\right| > \frac{\kappa}{2(\kappa+1)}\right) \\
&\qquad + \sum_{x_0\in\mathcal{X}_k} \mathbb{P}\left(\left|\frac{1}{r_n\binom{n}{r_n}p_{x_0}} \sum_{I\in B_{r_n,n}} \sum_{j\in I} p_{x_0}(\omega_I) - 1\right| > \frac{\kappa}{2(\kappa+1)}\right) \qquad (5.96)
\end{aligned}$$

because

$$|1 - g^{-1}(1+\kappa)| = |1 - \frac{1}{1+\kappa}| = \frac{\kappa}{\kappa+1} = \frac{1}{1+\kappa^{-1}}.$$

With Lemma 2.11, Proposition 2.16, Remark 2.17 and (4.6) we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{r_n p_{x_0}} \frac{1}{\binom{n}{r_n}} \sum_{I\in B_{r_n,n}} \sum_{j\in I} \left(\mathbb{I}\{X_j \in A_k(x_0,\omega_I)\} - p_{x_0}(\omega_I)\right)\right|^q\right] \\
&= \left(\frac{1}{r_n p_{x_0}}\right)^q \mathbb{E}\left[\left|\frac{1}{\binom{n}{r_n}} \sum_{I\in B_{r_n,n}} \sum_{j\in I} \left(\mathbb{I}\{X_j \in A_k(x_0,\omega_I)\} - p_{x_0}(\omega_I)\right)\right|^q\right] \\
&\lesssim \left(\frac{2^k}{r_n}\right)^q \left(\frac{r_n}{n}\right)^{q/2} \mathbb{E}\left[\left|\sum_{j=1}^{r_n} \left(\mathbb{I}\{X_j \in A_k(x_0,\omega)\} - p_{x_0}(\omega)\right)\right|^q\right] \\
&\lesssim \left(\frac{2^k}{r_n}\right)^q \left(\frac{r_n}{n}\right)^{q/2} \mathbb{E}\left[\left(r_n p_{x_0}(\omega)\right)^{q/2}\right] \\
&\lesssim \left(\frac{2^k}{n}\right)^{q/2}. \qquad (5.97)
\end{aligned}$$

For the second term from (5.96) we note that

$$\frac{1}{r_n\binom{n}{r_n}p_{x_0}} \sum_{I\in B_{r_n,n}} \sum_{j\in I} p_{x_0}(\omega_I) - 1 = \frac{1}{\binom{n}{r_n}p_{x_0}} \sum_{I\in B_{r_n,n}} (p_{x_0}(\omega_I) - p_{x_0}).$$

Hence we obtain

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{r_n\binom{n}{r_n}p_{x_0}}\sum_{I\in B_{r_n,n}}\sum_{j\in I}p_{x_0}(\omega_I)-1\right)^2\right] &= \mathbb{E}\left[\left(\frac{1}{p_{x_0}\binom{n}{r_n}}\sum_{I\in B_{r_n,n}}\left(p_{x_0}(\omega_I)-p_{x_0}\right)\right)^2\right] \\
&= \frac{1}{p_{x_0}^2\binom{n}{r_n}}\mathbb{E}\left[\left(p_{x_0}(\omega)-p_{x_0}\right)^2\right] \\
&\leq \frac{1}{p_{x_0}^2\binom{n}{r_n}}\mathbb{E}\left[p_{x_0}^2(\omega)\right] \\
&= \frac{1}{p_{x_0}^2\binom{n}{r_n}}\mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\{X_1\in A_k(x_0,\omega)\}\mid\omega\right]^2\right] \\
&\leq \frac{1}{p_{x_0}\binom{n}{r_n}} \\
&\lesssim \frac{2^k}{\binom{n}{r_n}} \qquad (5.98)
\end{aligned}$$

because

$$\operatorname{Cov}\left(p_{x_0}(\omega_1),p_{x_0}(\omega_2)\right)=0.$$

Together (5.96), (5.97) and (5.98) yield

$$\begin{aligned}
&\mathbb{P}\left(\sup_{x_0\in[0,1]^p}|\delta_n(x_0)|>\kappa\right) \\
&\leq \sum_{x_0\in\mathcal{X}_k}\mathbb{P}\left(\left|\frac{1}{r_n\binom{n}{r_n}p_{x_0}}\sum_{I\in B_{r_n,n}}\sum_{j\in I}\left(\mathbb{I}\{X_j\in A_k(x_0,\omega_I)\}-p_{x_0}(\omega_I)\right)\right|>\frac{\kappa}{2(\kappa+1)}\right) \\
&\quad+\sum_{x_0\in\mathcal{X}_k}\mathbb{P}\left(\left|\frac{1}{r_n\binom{n}{r_n}p_{x_0}}\sum_{I\in B_{r_n,n}}\sum_{j\in I}p_{x_0}(\omega_I)-1\right|>\frac{\kappa}{2(\kappa+1)}\right). \\
&\leq \sum_{x_0\in\mathcal{X}_k}\mathbb{E}\left[\left|\frac{1}{r_np_{x_0}}\frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}}\sum_{j\in I}\left(\mathbb{I}\{X_j\in A_k(x_0,\omega_I)\}-p_{x_0}(\omega_I)\right)\right|^{q_R}\right]\left(\frac{\kappa}{2(\kappa+1)}\right)^{-q_R} \\
&\quad+\sum_{x_0\in\mathcal{X}_k}\mathbb{E}\left[\left|\frac{1}{r_n\binom{n}{r_n}p_{x_0}}\sum_{I\in B_{r_n,n}}\sum_{j\in I}p_{x_0}(\omega_I)-1\right|^2\right]\left(\frac{\kappa}{2(\kappa+1)}\right)^{-2} \\
&\lesssim N_f(k)\left(\frac{2^k}{n}\right)^{q_R/2}(1+\kappa^{-1})^{q_R}+N_f(k)\frac{2^k}{\binom{n}{r_n}}(1+\kappa^{-1})^2.
\end{aligned}$$

This implies

$$\mathbb{P}\left(\|\delta_n\|_\infty>\kappa(\log n)^{-3/2}\right)\lesssim N_f(k)\left(\frac{2^k}{n}\right)^{q_R/2}(1+(\log n)^{3/2})^{q_R}+N_f(k)\frac{2^k}{\binom{n}{r_n}}(1+(\log n)^{3/2})^2.$$

For the first term we have

$$N_f(k)\left(\frac{2^k}{n}\right)^{q_R/2}(1+(\log n)^{3/2})^{q_R} \lesssim N_f(k)\left(\frac{2^k(\log n)^3}{n}\right)^{q_R/2} \to 0$$

with assumption (5.26) and for the second one we have

$$N_f(k)\frac{2^k}{\binom{n}{r_n}}(1+(\log n)^{3/2})^2 \lesssim 2^{k(p+1)}\big(\frac{r_n}{n}\big)^{r_n}(\log n)^3 \to 0$$

by assumption (5.3). Hence $\|\delta_n\|_\infty = o_{\mathbb{P}}((\log n)^{-3/2})$. This also implies that $1+\|\delta_n\|_\infty = \mathcal{O}(1)$. Corollary 5.20 yields with a union bound

$$\begin{aligned}\mathbb{P}\left(\|R^{(K)}_{n,r_n,\omega}\|_\infty > \kappa(\log n)^{-1}\sqrt{\frac{2^{2k}}{n}\mathcal{V}_{\cap,k}}\right) &\le \kappa^{-2}\frac{(\log n)^2 n}{2^{2k}\mathcal{V}_{\cap,k}}\sum_{x_0\in\mathcal{X}_k}\mathbb{E}\left[R^{(K)}_{n,r_n,\omega}(x_0)^2\right]\\ &\le N_f(k)\frac{\sigma^2}{c_X\kappa^2}\frac{(\log n)^2}{2^k\mathcal{V}_{\cap,k}}\left(\frac{r_n}{n}\right)^{r_n-1}\\ &\le \frac{\sigma^2}{c_X\kappa^2}2^{k(p+1)}(\log n)^2 c^{r_n-1}\to 0\end{aligned}$$

for all $\kappa>0$ by assumption (5.3). Hence

$$\sqrt{\frac{n}{2^{2k}\mathcal{V}_{\cap,k}}}\|R^{(K)}_{n,r_n,\omega}\|_\infty(1+\|\delta_n\|_\infty) = \sqrt{\frac{n}{2^{2k}\mathcal{V}_{\cap,k}}}\|R^{(K)}_{n,r_n,\omega}\|_\infty\mathcal{O}(1) = o_{\mathbb{P}}\left((\log n)^{-1}\right). \quad (5.99)$$

For any $\nu\ge 4$ with $\mathbb{E}[|\varepsilon_1|^\nu]<\infty$ we get analogue to Theorem 5.14 that

$$\begin{aligned}&\Big|\sqrt{\frac{n}{\sigma^2}}\sup_{x_0\in[0,1]^p}|\Phi_k^{-1/2}(x_0)\hat{U}^{(K,\varepsilon)}_{n,r_n,\omega}(x_0)| - \mathbf{S}_k\Big|\,(1+\|\delta_n\|_\infty)\\ &\qquad = \mathcal{O}_{\mathbb{P}}\left(\frac{(\log n)^{3/2}}{\mathcal{V}_{\cap,k}^{1/2}n^{1/2-1/\nu}} + \frac{(\log n)^{5/4}}{\mathcal{V}_{\cap,k}^{1/4}n^{1/4}} + \frac{\log n}{\mathcal{V}_{\cap,k}^{1/6}n^{1/6}}\right). \quad (5.100)\end{aligned}$$

It remains to show that $\mathbf{S}_k\|\delta_n\|_\infty = o_{\mathbb{P}}((\log n)^{-1})$. Regardless of the Gaussian process being a different one, we get analogously to equation (5.40) that

$$\mathbb{P}\left(\mathbf{S}_k>\kappa\right) \le \kappa^{-1}\mathbb{E}\left[\sup_{x_0\in[0,1]^p}|B_k\check{f}_{x_0,k}|\right] \lesssim \kappa^{-1}\sqrt{k}.$$

This implies $\mathbf{S}_k = \mathcal{O}_{\mathbb{P}}(\sqrt{k})$ and hence

$$\mathbf{S}_k\|\delta_n\|_\infty = \mathcal{O}_{\mathbb{P}}\left(\sqrt{k}\right)o_{\mathbb{P}}\left((\log n)^{-3/2}\right) = o_{\mathbb{P}}\left((\log n)^{-1}\right) \quad (5.101)$$

since $2^k\le n$. In total (5.91), (5.95), (5.99), (5.100) and (5.101) yield

$$\begin{aligned}&\left|\sqrt{n}\|\Phi_k^{-1/2}(U^{(\mathrm{KRF})}_{n,r_n,\omega}-m)\|_\infty - \sigma\mathbf{S}_k\right|\\ &\qquad = \mathcal{O}_{\mathbb{P}}\left(\frac{(\log n)^{3/2}}{\mathcal{V}_{\cap,k}^{1/2}n^{1/2-1/\nu}} + \frac{(\log n)^{5/4}}{\mathcal{V}_{\cap,k}^{1/4}n^{1/4}} + \frac{\log n}{\mathcal{V}_{\cap,k}^{1/6}n^{1/6}}\right) + o_{\mathbb{P}}\left((\log n)^{-1}\right)\end{aligned}$$

and the claim follows with the same arguments as in the proof of Theorem 5.1. □

5.6.9 Proofs for Section 5.4

5.6.9.1 Proof of Lemma 5.11

We start with the case $A_k(x_1, \omega_1) \cap A_k(x_2, \omega_2) \neq \emptyset$. Thus, there exists $x_0 \in A_k(x_1, \omega_1) \cap A_k(x_2, \omega_2)$. Equation (3.19) yields

$$\begin{aligned}\mathbb{V}\left(A_k(x_1, \omega_1) \cap A_k(x_2, \omega_2)\right) &= \mathbb{V}\left(A_k(x_0, \omega_1) \cap A_k(x_0, \omega_2)\right) \\ &= 2^{-\sum_{l=1}^{p} \max\{S_l(x_0,\omega_1), S_l(x_0,\omega_2)\}} \\ &= 2^{-\sum_{l=1}^{p} \max\{S_l(x_1,\omega_1), S_l(x_2,\omega_2)\}}.\end{aligned}$$

It remains to prove that

$$\begin{aligned}&\mathbb{I}\left\{A_k(x_1, \omega_1) \cap A_k(x_2, \omega_2) \neq \emptyset\right\} \\ &\quad = \prod_{l=1}^{p} \mathbb{I}\left\{\lfloor x_1^{(l)} 2^{\min\{S_l(x_1,\omega_1), S_l(x_2,\omega_2)\}} \rfloor = \lfloor x_2^{(l)} 2^{\min\{S_l(x_1,\omega_1), S_l(x_2,\omega_2)\}} \rfloor\right\}. \qquad (5.102)\end{aligned}$$

For the projections on the coordinates$A_k^{(l)}$ from (3.6) it holds that

$$\mathbb{I}\left\{A_k(x_1, \omega_1) \cap A_k(x_2, \omega_2) \neq \emptyset\right\} = \prod_{l=1}^{p} \mathbb{I}\left\{A_k^{(l)}(x_1, \omega_1) \cap A_k^{(l)}(x_2, \omega_2) \neq \emptyset\right\}. \qquad (5.103)$$

Similar to (3.18) the exact form of the $A_k^{(l)}$ described in Remark 3.5 implies that

$$\begin{aligned}&A_k^{(l)}(x_1, \omega_1) \cap A_k^{(l)}(x_2, \omega_2) \neq \emptyset \\ &\quad \Leftrightarrow \left(A_k^{(l)}(x_1, \omega_1) \subset A_k^{(l)}(x_2, \omega_2)\right) \vee \left(A_k^{(l)}(x_1, \omega_1) \supset A_k^{(l)}(x_2, \omega_2)\right).\end{aligned}$$

The form further implies that one interval being a subset of the other is equivalent to both $x_1^{(l)}$ and $x_2^{(l)}$, being in the larger set that is equal to the union. For the union, (3.7) yields

$$\begin{aligned}&A_k^{(l)}(x_1, \omega_1) \cup A_k^{(l)}(x_2, \omega_2) \\ &\quad = 2^{-\min\{S_l(x_1,\omega_1), S_l(x_2,\omega_2)\}} \left[\lfloor x_1^{(l)} 2^{\min\{S_l(x_1,\omega_1), S_l(x_2,\omega_2)\}} \rfloor, \lfloor x_1^{(l)} 2^{\min\{S_l(x_1,\omega_1), S_l(x_2,\omega_2)\}} \rfloor + 1\right).\end{aligned}$$

Together this implies

$$\begin{aligned}&A_k^{(l)}(x_1, \omega_1) \cap A_k^{(l)}(x_2, \omega_2) \neq \emptyset \\ &\quad \Leftrightarrow \left(A_k^{(l)}(x_1, \omega_1) \subset A_k^{(l)}(x_2, \omega_2)\right) \vee \left(A_k^{(l)}(x_1, \omega_1) \supset A_k^{(l)}(x_2, \omega_2)\right) \\ &\quad \Leftrightarrow x_1^{(l)}, x_2^{(l)} \in 2^{-\min\{S_l(x_1,\omega_1), S_l(x_2,\omega_2)\}} \\ &\qquad\quad \times \left[\lfloor x_1^{(l)} 2^{\min\{S_l(x_1,\omega_1), S_l(x_2,\omega_2)\}} \rfloor, \lfloor x_1^{(l)} 2^{\min\{S_l(x_1,\omega_1), S_l(x_2,\omega_2)\}} \rfloor + 1\right) \\ &\quad \Leftrightarrow \lfloor x_1^{(l)} 2^{\min\{S_l(x_1,\omega_1), S_l(x_2,\omega_2)\}} \rfloor = \lfloor x_2^{(l)} 2^{\min\{S_l(x_1,\omega_1), S_l(x_2,\omega_2)\}} \rfloor.\end{aligned}$$

Plugging this into (5.103) we obtain (5.102). □

5.6.9.2 Proof of Lemma 5.12

We can assume that

$$\begin{aligned}&\left(\max\{t \in \{0,\dots k\} : \lfloor x_1^{(l)} 2^t \rfloor = \lfloor x_2^{(l)} 2^t \rfloor\}\right)_{l=1}^p \\ &\quad = \left(\max\{t \in \{0,\dots k\} : \lfloor x_3^{(l)} 2^t \rfloor = \lfloor x_4^{(l)} 2^t \rfloor\}\right)_{l=1}^p\end{aligned}$$

without loss of generality because the CPRF is symmetric. For $s_1, s_2 \in \{0,\dots,k\}^p$ we denote $s_1 = (s_1(l))_{l=1}^p$ and s_2 analogously. For $S(x,\omega) = (S_l(x,\omega))_{l=1}^p$ we obtain

$$\begin{aligned}&\mathbb{P}\left(X_1 \in A_k(x_1,\omega_1) \cap A_k(x_2,\omega_2)\right) \\ &\quad = \mathbb{E}\left[\mathbb{V}\left(A_k(x_1,\omega_1) \cap A_k(x_2,\omega_2)\right)\right] \\ &\quad = \mathbb{E}\Bigg[2^{-\sum_{l=1}^p \max\{S_l(x_1,\omega_1),S_l(x_2,\omega_2)\}} \\ &\qquad \times \prod_{l=1}^p \mathbb{I}\left\{\lfloor x_1^{(l)} 2^{\min\{S_l(x_1,\omega_1),S_l(x_2,\omega_2)\}} \rfloor = \lfloor x_2^{(l)} 2^{\min\{S_l(x_1,\omega_1),S_l(x_2,\omega_2)\}} \rfloor\right\}\Bigg] \\ &\quad = \sum_{s_1,s_2 \in \{0,\dots,k\}^p} \mathbb{P}\left(S(x_1,\omega_1) = s_1, S(x_2,\omega_2) = s_2\right) \\ &\qquad \times 2^{-\sum_{l=1}^p \max\{s_1(l),s_2(l)\}} \prod_{l=1}^p \mathbb{I}\left\{\lfloor x_1^{(l)} 2^{\min\{s_1(l),s_2(l)\}} \rfloor = \lfloor x_2^{(l)} 2^{\min\{s_1(l),s_2(l)\}} \rfloor\right\} \\ &\quad = \sum_{s_1,s_2 \in \{0,\dots,k\}^p} \mathbb{P}\left(S(x_3,\omega_1) = s_1, S(x_4,\omega_2) = s_2\right) \\ &\qquad \times 2^{-\sum_{l=1}^p \max\{s_1(l),s_2(l)\}} \prod_{l=1}^p \mathbb{I}\left\{\lfloor x_3^{(l)} 2^{\min\{s_1(l),s_2(l)\}} \rfloor = \lfloor x_4^{(l)} 2^{\min\{s_1(l),s_2(l)\}} \rfloor\right\} \\ &\quad = \mathbb{P}\left(X_1 \in A_k(x_3,\omega_1) \cap A_k(x_4,\omega_2)\right)\end{aligned}$$

because for all $l \in [p]$ and $t \in \{0,\dots k\}$ we know that

$$\mathbb{I}\left\{\lfloor x_1^{(l)} 2^t \rfloor = \lfloor x_2^{(l)} 2^t \rfloor\right\} = \mathbb{I}\left\{\lfloor x_1^{(l)} 2^t \rfloor = \lfloor x_2^{(l)} 2^t \rfloor\right\}$$

and in particular,

$$\begin{aligned}&\mathbb{I}\left\{\lfloor x_1^{(l)} 2^{\min\{s_1(l),s_2(l)\}} \rfloor = \lfloor x_2^{(l)} 2^{\min\{s_1(l),s_2(l)\}} \rfloor\right\} \\ &\quad = \mathbb{I}\left\{\lfloor x_1^{(l)} 2^{\min\{s_1(l),s_2(l)\}} \rfloor = \lfloor x_2^{(l)} 2^{\min\{s_1(l),s_2(l)\}} \rfloor\right\}.\square\end{aligned}$$

5.6.10 Proofs of the auxiliary results

In this section we collect the proofs of the auxiliary results that have been used throughout the chapter so far. We start with the proof of Theorem 5.14, continue with the proofs for the approximation error results, and conclude the section with the proofs for the remainder terms that arise from the projection error.

5.6.10.1 Proof of Theorem 5.14

We want to apply Theorem 2.4. We note that the function class in the theorem is not dependent on n. However, we can apply the theorem to $\mathcal{F}_k$ for every n or k, respectively. Also the constant in the theorem does not depend on n. This is why we get a sequence of random variables in the claim of the theorem. Chernozhukov et al. (2014b) also point this out in their Remark 2.1. We consider the function class $\mathcal{F}_k$ from (5.1). To get the claimed result we need to consider $\mathcal{F}_k \cup -\mathcal{F}_k$. Due to Chernozhukov et al. (2014b, Corollary 2.1) we consider $\mathcal{F}_k$ without loss of generality.

In Section 5.4 we already explained why $\mathcal{F}_k$ is finite. Thus, it is pointwise measurable because the $f_{x_0,k}$ are measurable. Now, we prove that $\mathcal{F}_k$ is a VC type class satisfying Definition 2.3. Using the definition of K_k in (4.8) together with the bound for $p_{x_0}(\omega)$ in (4.6) and $\Psi_k(x_0) \geq c_X C_X^{-2} 2^{2k} \mathcal{V}_{\cap,k}$ from (4.20) we obtain

$$\begin{aligned}
\sup_{x_0 \in [0,1]^p} |f_{x_0,k}(x,s)| &= \sup_{x_0 \in [0,1]^p} \sigma^{-1} |s| \Psi_k^{-1/2}(x_0) K_k(x_0, x) \\
&\leq \sigma^{-1} |s| \frac{C_X}{c_X^{1/2}} 2^{-k} \mathcal{V}_{\cap,k}^{-1/2} \sup_{x_0 \in [0,1]^p} \mathbb{E}\left[\mathbb{I}\{x \in A_k(x_0, \omega)\} p_{x_0}(\omega)^{-1}\right] \\
&\leq \sigma^{-1} |s| \frac{C_X}{c_X^{1/2}} 2^{-k} \mathcal{V}_{\cap,k}^{-1/2} c_X^{-1} 2^k \\
&= \sigma^{-1} |s| \frac{C_X}{c_X^{3/2}} \mathcal{V}_{\cap,k}^{-1/2} \\
&=: F(x,s). \qquad (5.104)
\end{aligned}$$

Therefore, $\mathcal{F}_k$ is equipped with the measurable envelope F that does not depend on x. The finite size $|\mathcal{F}_k| = N_f(k) \leq 2^{kp}$ is an upper bound for any covering number of $\mathcal{F}_k$ and thus,

$$\sup_{Q \in \mathcal{Q}} N(\mathcal{F}_k, \|\cdot\|_{Q,2}, \kappa \|F\|_{Q,2}) \leq 2^{kp} \leq 2^{kp}/\kappa$$

for all $\kappa \in (0,1]$, implies that $\mathcal{F}_k$ is a VC type class satisfying Definition 2.3 for $A = 2^{kp}$ and $v = 1$.

We proceed by verifying the conditions on the moments. With the definitions of K_k in (4.8) and Ψ_k in (4.9) we obtain

$$\begin{aligned}
\sup_{f \in \mathcal{F}_k} P|f|^2 &= \sup_{x_0 \in [0,1]^p} \sigma^{-2} \Psi_k^{-1}(x_0) \mathbb{E}\left[\varepsilon_1^2 K_k^2(x_0, X_1)\right] \\
&= \sup_{x_0 \in [0,1]^p} \Psi_k^{-1}(x_0) \mathbb{E}\left[K_k^2(x_0, X_1)\right] = 1.
\end{aligned}$$

Thus, $\tilde{\sigma}$ from Theorem 2.4 is equal to 1. For $q \leq \nu$ we denote $\tau_q := \mathbb{E}[|\varepsilon_1|^q]^{1/q}$. The definitions used above, the lower bound for p_{x_0} in (4.7) and $\Psi_k(x_0) \geq c_X C_X^{-2} 2^{2k} \mathcal{V}_{\cap,k}$ from (4.20) further yield

$$\begin{aligned}
\sup_{f \in \mathcal{F}_k} P|f|^3 &= \sup_{x_0 \in [0,1]^p} \sigma^{-3} \Psi_k^{-3/2}(x_0) \mathbb{E}\left[|\varepsilon_1 K_k(x_0, X_1)|^3\right] \\
&= \frac{\mathbb{E}\left[|\varepsilon_1|^3\right]}{\sigma^3} \sup_{x_0 \in [0,1]^p} \Psi_k^{-3/2}(x_0) \mathbb{E}\left[K_k^2(x_0, X_1) \mathbb{E}\left[\mathbb{I}\{X_1 \in A_k(x_0, \omega)\} p_{x_0}(\omega)^{-1} \mid X_1\right]\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{\tau_3^3}{\sigma^3} \sup_{x_0 \in [0,1]^p} \Psi_k^{-3/2}(x_0) \mathbb{E}\left[K_k^2(x_0, X_1)\right] c_X^{-1} 2^k \\
&= \frac{\tau_3^3}{\sigma^3} c_X^{-1} 2^k \sup_{x_0 \in [0,1]^p} \Psi_k^{-1/2}(x_0) \\
&\leq \frac{\tau_3^3}{\sigma^3} c_X^{-3/2} C_X 2^k 2^{-k} \mathcal{V}_{\cap,k}^{-1/2} \\
&= \frac{\tau_3^3}{\sigma^3} \frac{C_X}{c_X^{3/2}} \mathcal{V}_{\cap,k}^{-1/2}.
\end{aligned}
$$

For the envelope F from (5.104) we obtain

$$
\|F\|_{P,\nu} = \mathbb{E}\left[|F(X_1, \varepsilon_1)|^\nu\right]^{1/\nu} = \sigma^{-1} \frac{C_X}{c_X^{3/2}} \mathcal{V}_{\cap,k}^{-1/2} \mathbb{E}\left[|\varepsilon_1|^\nu\right]^{1/\nu} = \frac{C_X}{c_X^{3/2}} \frac{\tau_\nu}{\sigma} \mathcal{V}_{\cap,k}^{-1/2}.
$$

We choose b from Theorem 2.4 as

$$
b := \mathcal{V}_{\cap,k}^{-1/2} C_b := \mathcal{V}_{\cap,k}^{-1/2} \frac{C_X}{c_X^{3/2}} \max\left\{\frac{\tau_\nu}{\sigma}, \frac{\tau_3^3}{\sigma^3}\right\}.
$$

It holds that $b \geq 1$ because $c_X \leq 1 \leq C_X$, $\mathcal{V}_{\cap,k}^{-1/2} \geq 2^{k/2} \geq 1$ and $\tau_q \geq \sigma$ due to Jensen's inequality. Further we have $b = \mathcal{O}(\mathcal{V}_{\cap,k}^{-1/2})$ because C_b is a constant that only depends on the distributions of ε_1 and X_1. We get

$$
\begin{aligned}
\sup_{f \in \mathcal{F}_k \cup -\mathcal{F}_k} \mathbb{G}_n f &= \sup_{f \in \mathcal{F}_k} |\mathbb{G}_n f| \\
&= \sup_{x_0 \in [0,1]^p} |\mathbb{G}_n \sigma^{-1} \Psi_k^{-1/2}(x_0) s K_k(x_0, x)| \\
&= \sup_{x_0 \in [0,1]^p} |\frac{1}{\sqrt{\sigma^2 n}} \Psi_k^{-1/2}(x_0) \sum_{j=1}^n \varepsilon_j K_k(x_0, X_j)| \\
&= \sqrt{\frac{n}{\sigma^2}} \sup_{x_0 \in [0,1]^p} |\Psi_k^{-1/2}(x_0) \hat{U}_{n, r_n, \omega}^{(\varepsilon)}(x_0)|.
\end{aligned}
$$

In summary, the parameters from Theorem 2.4 are $\tilde{\sigma} = 1$, $b = \mathcal{V}_{\cap,k}^{-1/2} C_b$, $A = 2^{kp}$ and $v = 1$. We use $\mathcal{V}_{\cap,k} \geq 2^{-2k}$ from (3.20) and the fact that $2^k = o(n)$ (see (5.12) and (5.3)) to obtain

$$
\begin{aligned}
K_n &= cv(\log n \vee \log(Ab/\tilde{\sigma})) \\
&= c(\log n \vee \log(2^{kp} \mathcal{V}_{\cap,k}^{-1/2} C_b)) \\
&\leq c(\log n \vee \log(2^{k(p+1)} C_b)) \\
&\leq c(\log n \vee (p+1) \log(2^k C_b)) \\
&\leq c(p+1)(\log n \vee \log(2^k C_b)) \\
&= \mathcal{O}(\log n).
\end{aligned}
$$

Let B_k be the centered Gaussian process defined in the claim. For $\gamma = (\log n)^{-1}$ Theorem 2.4 yields that there exists a random variable

$$\mathbf{S}_k \stackrel{d}{=} \sup_{f \in \mathcal{F}_k \cup -\mathcal{F}_k} B_k f = \sup_{f \in \mathcal{F}_k} |B_k f| = \sup_{x_0 \in [0,1]^p} |B_k f_{x_0,k}|$$

such that

$$\begin{aligned}
&\mathbb{P}\left(\Big| \sup_{f \in \mathcal{F}_k} |\mathbb{G}_n f| - \mathbf{S}_k \Big| > \frac{bK_n (\log n)^{1/2}}{n^{1/2-1/\nu}} + \frac{(b \log n)^{1/2} K_n^{3/4}}{n^{1/4}} + \frac{(bK_n^2 \log n)^{1/3}}{n^{1/6}} \right) \\
&\quad \leq C \left((\log n)^{-1} + \frac{\log n}{n} \right)
\end{aligned}$$

since $\tilde{\sigma} = 1$. Hence

$$\begin{aligned}
&\left| \sqrt{\frac{n}{\sigma^2}} \sup_{x_0 \in [0,1]^p} |\Psi_k^{-1/2}(x_0) \hat{U}^{(\varepsilon)}_{n,r_n,\omega}(x_0)| - \mathbf{S}_k \right| \\
&\qquad = \Big| \sup_{f \in \mathcal{F}_k} |\mathbb{G}_n f| - \mathbf{S}_k \Big| \\
&\qquad = \mathcal{O}_{\mathbb{P}} \left(\frac{bK_n (\log n)^{1/2}}{n^{1/2-1/\nu}} + \frac{(b \log n)^{1/2} K_n^{3/4}}{n^{1/4}} + \frac{(bK_n^2 \log n)^{1/3}}{n^{1/6}} \right) \\
&\qquad = \mathcal{O}_{\mathbb{P}} \left(\frac{(\log n)^{3/2}}{\mathcal{V}_{\cap,k}^{1/2} n^{1/2-1/\nu}} + \frac{(\log n)^{5/4}}{\mathcal{V}_{\cap,k}^{1/4} n^{1/4}} + \frac{\log n}{\mathcal{V}_{\cap,k}^{1/6} n^{1/6}} \right).
\end{aligned}$$

since $K_n = \mathcal{O}((\log n))$ and $b = \mathcal{O}(\mathcal{V}_{\cap,k}^{-1/2})$. □

5.6.10.2 Proofs of the approximation error results

Proof of Lemma 5.16. With (5.44) we have

$$\begin{aligned}
&U^{(m)}_{n,r_n,\omega}(x_0) - m(x_0) U^{(1)}_{n,r_n,\omega}(x_0) \\
&\qquad = \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \sum_{j \in I} (m(X_j) - m(x_0)) \frac{\mathbb{I}\{X_j \in A_k(x_0, \omega_I)\}}{\sum_{i \in I} \mathbb{I}\{X_i \in A_k(x_0, \omega_I)\}},
\end{aligned}$$

which implies

$$\begin{aligned}
&\sup_{x_0 \in [0,1]^p} |U^{(m)}_{n,r_n,\omega}(x_0) - m(x_0) U^{(1)}_{n,r_n,\omega}(x_0)| \\
&\qquad \leq \sup_{x_0 \in [0,1]^p} \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \sum_{j \in I} |m(X_j) - m(x_0)| \frac{\mathbb{I}\{X_j \in A_k(x_0, \omega_I)\}}{\sum_{i \in I} \mathbb{I}\{X_i \in A_k(x_0, \omega_I)\}} \\
&\qquad \leq C_H \sup_{x_0 \in [0,1]^p} \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \mathfrak{d}(A_k(x_0, \omega_I))^\alpha \frac{\sum_{j \in I} \mathbb{I}\{X_j \in A_k(x_0, \omega_I)\}}{\sum_{i \in I} \mathbb{I}\{X_i \in A_k(x_0, \omega_I)\}} \\
&\qquad \leq C_H \sup_{x_0 \in [0,1]^p} \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \mathfrak{d}(A_k(x_0, \omega_I))^\alpha.
\end{aligned}$$

For the expectation we get

$$\mathbb{E}\left[\sup_{x_0\in[0,1]^p}\frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}}\mathfrak{d}(A_k(x_0,\omega_I))^\alpha\right]$$
$$=\mathbb{E}\left[\sup_{x_0\in[0,1]^p}\frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}}\mathfrak{d}(A_k(x_0,\omega_I))^\alpha-\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha]\right]+\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha].$$

Using the independence of the ω_I we obtain for the latter expectation that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{x_0\in[0,1]^p}\frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}}\mathfrak{d}(A_k(x_0,\omega_I))^\alpha-\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha]\right]\\
&\le\sum_{x_0\in\mathcal{X}_k}\mathbb{E}\left[\left|\frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}}\mathfrak{d}(A_k(x_0,\omega_I))^\alpha-\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha]\right|\right]\\
&\le N_f(k)\mathrm{Var}\left(\frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}}\mathfrak{d}(A_k(x_0,\omega_I))^\alpha\right)^{1/2}\\
&=N_f(k)\frac{1}{\sqrt{\binom{n}{r_n}}}\mathrm{Var}\big(\mathfrak{d}(A_k(x_0,\omega))^\alpha\big)^{1/2}\\
&\le N_f(k)\frac{1}{\sqrt{\binom{n}{r_n}}}\mathbb{E}\left[\mathfrak{d}(A_k(x_0,\omega))^{2\alpha}\right]^{1/2}\\
&\le N_f(k)\frac{p^{\alpha/2}}{\sqrt{\binom{n}{r_n}}}.
\end{aligned}$$

This leads to

$$\begin{aligned}
&\mathbb{E}\left[\sup_{x_0\in[0,1]^p}|U^{(m)}_{n,r_n,\omega}(x_0)-m(x_0)U^{(1)}_{n,r_n,\omega}(x_0)|\right]\\
&\le C_H\mathbb{E}\left[\sup_{x_0\in[0,1]^p}\frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}}\mathfrak{d}(A_k(x_0,\omega_I))^\alpha\right]\\
&\le C_H\left(\mathbb{E}[\mathfrak{d}(A_k(x,\omega))^\alpha]+N_f(k)p^{\alpha/2}\binom{n}{r_n}^{-1/2}\right).
\end{aligned}$$

□

Proof of Lemma 5.17. For any ω, let $\mathcal{X}_k(\omega)$ denote a set of 2^k points, with exactly one point in each of the partition cells created by omega. That means $|\mathcal{X}_k(\omega)\cap A_k(x_0,\omega)|=1$ for every $x_0\in[0,1]^p$. Using (5.44) for $U^{(1)}_{n,r_n,\omega}(x_0)$ we obtain

$$\mathbb{E}\left[\sup_{x_0\in[0,1]^p}|m(x_0)(U^{(1)}_{n,r_n,\omega}(x_0)-1)|\right]$$

$$
\begin{aligned}
&\leq \|m\|_\infty \mathbb{E}\left[\sup_{x_0\in[0,1]^p}\left|\frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}}\left(\mathbb{I}\{\exists j\in I: X_j\in A_k(x_0,\omega_I)\}-1\right)\right|\right]\\
&\leq \|m\|_\infty \mathbb{E}\left[\frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}}\sup_{x_0\in[0,1]^p}\mathbb{I}\{\nexists j\in I: X_j\in A_k(x_0,\omega_I)\}\right]\\
&= \|m\|_\infty \mathbb{E}\Big[\sup_{x_0\in[0,1]^p}\mathbb{I}\{\nexists j\in [r_n]: X_j\in A_k(x_0,\omega)\}\Big]\\
&= \|m\|_\infty \mathbb{E}\Big[\mathbb{E}\Big[\max_{x_0\in\mathcal{X}_k(\omega)}\mathbb{I}\{\nexists j\in [r_n]: X_j\in A_k(x_0,\omega)\}\mid\omega\Big]\Big]\\
&\leq \|m\|_\infty \mathbb{E}\left[\mathbb{E}\left[\sum_{x_0\in\mathcal{X}_k(\omega)}\mathbb{I}\{\nexists j\in [r_n]: X_j\in A_k(x_0,\omega)\}\mid\omega\right]\right]\\
&= \|m\|_\infty \mathbb{E}\left[\sum_{x_0\in\mathcal{X}_k(\omega)}(1-p_{x_0}(\omega))^{r_n}\right]\\
&\leq \|m\|_\infty 2^k(1-c_X2^{-k})^{r_n}.
\end{aligned}
$$

This completes the proof. □

Remark 5.21. We note that

$$
\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}}\mathbb{I}\{\exists j\in I: X_j\in A_k(x_0,\omega_I)\}-1\right|\right]\\
&\quad=\mathbb{E}\left[\frac{1}{\binom{n}{r_n}}\sum_{I\in B_{r_n,n}}\mathbb{I}\{\nexists j\in I: X_j\in A_k(x_0,\omega_I)\}\right]\\
&\quad=\mathbb{E}\left[\mathbb{I}\{\nexists j\in [r_n]: X_j\in A_k(x_0,\omega)\}\right]\\
&\quad=\mathbb{E}\left[(1-p_{x_0}(\omega))^{r_n}\right]\\
&\quad\geq(1-C_X2^{-k})^{r_n}.
\end{aligned}
$$

Therefore, the bound in Lemma 5.17 is sharp up to the constant and the union bound. The bound

$$
2^k(1-c_X2^{-k})^{r_n}\lesssim\exp\left(k\log 2-c_Xr_n/2^k\right)=\exp\left(-r_n/2^k(c_X-k2^kr_n^{-1}\log 2)\right)
$$

illustrates that the union bound is negligible as long as $k2^kr_n^{-1}\to 0$.

5.6.10.3 Proofs of the remainder results

Proof of Lemma 5.18. We omit x_0 in the notation of

$$
h_{R,n}((X_i,\varepsilon_i)_{i\in I},\omega_I):=\sum_{j\in I}\varepsilon_j\mathbb{I}\{X_j\in A_k(x_0,\omega_I)\}\left(\frac{1}{\sum_{i\in I}\mathbb{I}\{X_i\in A_k(x_0,\omega_I)\}}-\frac{1}{r_np_{x_0}(\omega_I)}\right)
$$

because it is fixed throughout the proof. Then we have

$$R^{(1)}_{n,r_n,\omega}(x_0) = \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} h_{R,n}((X_i,\varepsilon_i)_{i\in I}, \omega_I),$$

which is a generalized U-statistic with centered kernel $h_{R,n}$. For even $q > 0$ Lemma 2.11 yields

$$\mathbb{E}\left[R^{(1)}_{n,r_n,\omega}(x_0)^q\right] \lesssim \left(\frac{r_n}{n}\right)^{q/2} \mathbb{E}\left[h_{R,n}((X_i,\varepsilon_i)_{i=1}^{r_n}, \omega)^q\right]. \tag{5.105}$$

With equation (2.7) and Hölder's inequality we get

$$\begin{aligned}
&\mathbb{E}\left[h_{R,n}((X_i,\varepsilon_i)_{i=1}^{r_n}, \omega)^q\right] \\
&\quad= \mathbb{E}\left[\left(\sum_{j=1}^{r_n} \varepsilon_j \mathbb{I}\{X_j \in A_k(x_0,\omega)\}\left(\frac{1}{\sum_{i=1}^{r_n}\mathbb{I}\{X_i \in A_k(x_0,\omega)\}} - \frac{1}{r_n p_{x_0}(\omega)}\right)\right)^q\right] \\
&\quad\lesssim \mathbb{E}\left[\left(\sum_{j=1}^{r_n} \varepsilon_j^2 \mathbb{I}\{X_j \in A_k(x_0,\omega)\}\left(\frac{1}{\sum_{i=1}^{r_n}\mathbb{I}\{X_i \in A_k(x_0,\omega)\}} - \frac{1}{r_n p_{x_0}(\omega)}\right)^2\right)^{q/2}\right] \\
&\quad= \sum_{q_1+\ldots+q_{r_n}=q/2} \binom{q/2}{q_1,\ldots,q_{r_n}} \\
&\qquad\times \mathbb{E}\left[\prod_{j=1}^{r_n} \varepsilon_j^{2q_j} \mathbb{I}\{X_j \in A_k(x_0,\omega)\}\left(\frac{1}{\sum_{i=1}^{r_n}\mathbb{I}\{X_i \in A_k(x_0,\omega)\}} - \frac{1}{r_n p_{x_0}(\omega)}\right)^{2q_j}\right] \\
&\quad= \sum_{q_1+\ldots+q_{r_n}=q/2} \binom{q/2}{q_1,\ldots,q_{r_n}} \mathbb{E}\left[\prod_{j=1}^{r_n} \varepsilon_j^{2q_j}\right] \\
&\qquad\times \mathbb{E}\left[\prod_{j=1}^{r_n} \mathbb{I}\{X_j \in A_k(x_0,\omega)\}\left(\frac{1}{\sum_{i=1}^{r_n}\mathbb{I}\{X_i \in A_k(x_0,\omega)\}} - \frac{1}{r_n p_{x_0}(\omega)}\right)^{2q_j}\right] \\
&\quad\le \mathbb{E}\left[\varepsilon_1^q\right] \sum_{q_1+\ldots+q_{r_n}=q/2} \binom{q/2}{q_1,\ldots,q_{r_n}} \\
&\qquad\times \mathbb{E}\left[\prod_{j=1}^{r_n} \mathbb{I}\{X_j \in A_k(x_0,\omega)\}\left(\frac{1}{\sum_{i=1}^{r_n}\mathbb{I}\{X_i \in A_k(x_0,\omega)\}} - \frac{1}{r_n p_{x_0}(\omega)}\right)^{2q_j}\right],
\end{aligned} \tag{5.106}$$

with $q_j \in \mathbb{N}_0$. For any $\mathbf{q} = (q_j)_{j\in[r_n]}$ with $\sum_{j=1}^{r_n} q_j = q/2$ let $J_{\mathbf{q}>0} := \{j \in [r_n] : q_j > 0\}$, $J_{\mathbf{q}=0} := \{j \in [r_n] : q_j = 0\}$ and $\tilde{q}(\mathbf{q}) := |J_{\mathbf{q}>0}|$. For the expectation from (5.106) we get with $p_{x_0}(\omega) \le C_X 2^{-k}$ (see (4.6)) and $\tilde{q}(\mathbf{q}) \le q/2$ that

$$\begin{aligned}
&\mathbb{E}\left[\prod_{j=1}^{r_n} \mathbb{I}\{X_j \in A_k(x_0,\omega)\}\left(\frac{1}{\sum_{i=1}^{r_n}\mathbb{I}\{X_i \in A_k(x_0,\omega)\}} - \frac{1}{r_n p_{x_0}(\omega)}\right)^{2q_j}\right] \\
&\quad= \mathbb{E}\left[\prod_{j\in J_{\mathbf{q}>0}} \mathbb{I}\{X_j \in A_k(x_0,\omega)\}\left(\frac{1}{\tilde{q}(\mathbf{q}) + \sum_{i\in J_{\mathbf{q}=0}}\mathbb{I}\{X_i \in A_k(x_0,\omega)\}} - \frac{1}{r_n p_{x_0}(\omega)}\right)^{2q_j}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left(\frac{1}{\tilde{q}(\mathbf{q}) + \sum_{i\in J_{\mathbf{q}=0}} \mathbb{I}\{X_i \in A_k(x_0,\omega)\}} - \frac{1}{r_n p_{x_0}(\omega)}\right)^q \prod_{j\in J_{\mathbf{q}>0}} \mathbb{I}\{X_j \in A_k(x_0,\omega)\}\right] \\
&= \mathbb{E}\Bigg[\left(\frac{1}{\tilde{q}(\mathbf{q}) + \sum_{i\in J_{\mathbf{q}=0}} \mathbb{I}\{X_i \in A_k(x_0,\omega)\}} - \frac{1}{r_n p_{x_0}(\omega)}\right)^q \\
&\qquad \times \mathbb{E}\Bigg[\prod_{j\in J_{\mathbf{q}>0}} \mathbb{I}\{X_j \in A_k(x_0,\omega)\}\Big|\, \omega, (X_i)_{i\in J_{\mathbf{q}=0}}\Bigg]\Bigg] \\
&= \mathbb{E}\left[\left(\frac{1}{\tilde{q}(\mathbf{q}) + \sum_{i\in J_{\mathbf{q}=0}} \mathbb{I}\{X_i \in A_k(x_0,\omega)\}} - \frac{1}{r_n p_{x_0}(\omega)}\right)^q p_{x_0}(\omega)^{\tilde{q}(\mathbf{q})}\right] \\
&\leq C_X^{q/2} 2^{-k\tilde{q}(\mathbf{q})} \mathbb{E}\left[\left(\frac{1}{\tilde{q}(\mathbf{q}) + \sum_{i\in J_{\mathbf{q}=0}} \mathbb{I}\{X_i \in A_k(x_0,\omega)\}} - \frac{1}{r_n p_{x_0}(\omega)}\right)^q\right]. \qquad (5.107)
\end{aligned}
$$

Conditioned on ω, the sum in the expression above is Binomial distributed with parameters $r_n - \tilde{q}(\mathbf{q})$ and $p_{x_0}(\omega)$. We note that $\tilde{q}(\mathbf{q}) \leq q/2$ and hence we get with Lemma 2.19 that

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{\tilde{q}(\mathbf{q}) + \sum_{i=\tilde{q}(\mathbf{q})+1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0,\omega)\}} - \frac{1}{r_n p_{x_0}(\omega)}\right)^q\right] \\
&\quad = \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{\tilde{q}(\mathbf{q}) + \sum_{i=\tilde{q}(\mathbf{q})+1}^{r_n} \mathbb{I}\{X_i \in A_k(x_0,\omega)\}} - \frac{1}{r_n p_{x_0}(\omega)}\right)^q \,\middle|\, \omega\right]\right] \\
&\quad \lesssim \mathbb{E}\left[(r_n p_{x_0}(\omega))^{-q3/2}\right] \\
&\quad \leq (c_X r_n 2^{-k})^{-q3/2}. \qquad (5.108)
\end{aligned}
$$

Combining (5.106), (5.107) and (5.108), we obtain that there exists a constant C, independent of parameters involved, such that

$$
\begin{aligned}
&\mathbb{E}\left[h_{R,n}((X_i,\varepsilon_i)_{i=1}^{r_n},\omega)^q\right] \\
&\quad \leq \frac{C\mathbb{E}\left[\varepsilon_1^q\right] C_X^{q/2}}{c_X^{q3/2}} \left(r_n 2^{-k}\right)^{-q3/2} \sum_{q_1+\ldots+q_{r_n}=q/2} \binom{q/2}{q_1,\ldots,q_{r_n}} 2^{-k\tilde{q}((q_j)_{j\in[r_n]})}. \qquad (5.109)
\end{aligned}
$$

We note that

$$
\sum_{q_1+\ldots+q_{r_n}=q/2} \binom{q/2}{q_1,\ldots,q_{r_n}} = r_n^{q/2}.
$$

For $\check{q} \in \{1,\ldots,q/2\}$, let

$$
N(\check{q}) = \sum_{q_1+\ldots+q_{r_n}=q/2} \binom{q/2}{q_1,\ldots,q_{r_n}} \mathbb{I}\{|\{j\in[r_n], q_j > 0\}| = \check{q}\}.
$$

This implies

$$
\sum_{\check{q}=1}^{q/2} N(\check{q}) = r_n^{q/2}.
$$

Using that $0 \leq \check{q} \leq q/2$, we obtain

$$\begin{aligned}
N(\check{q}) &= \sum_{q_1+\ldots+q_{r_n}=q/2} \binom{q/2}{q_1,\ldots,q_{r_n}} \mathbb{I}\{\check{q} = |\{j \in [r_n], q_j > 0\}|\} \\
&= \sum_{q_1+\ldots+q_{r_n}=q/2} \frac{(q/2)!}{q_1!\ldots q_{r_n}!} \mathbb{I}\{\check{q} = |\{j \in [r_n], q_j > 0\}|\} \\
&\leq \frac{(q/2)!}{(\lfloor \frac{q}{2\check{q}} \rfloor !)^{\check{q}}} \sum_{q_1+\ldots+q_{r_n}=q/2} \mathbb{I}\{\check{q} = |\{j \in [r_n], q_j > 0\}|\} \\
&= \frac{(q/2)!}{(\lfloor \frac{q}{2\check{q}} \rfloor !)^{\check{q}}} \binom{r_n}{\check{q}} \check{q}^{q/2-\check{q}} \\
&= \frac{(q/2)!}{(\lfloor \frac{q}{2\check{q}} \rfloor !)^{\check{q}}} \frac{r_n!}{\check{q}!(r_n-\check{q})!} \check{q}^{q/2-\check{q}} \\
&\leq r_n^{\check{q}} (q/2)! (q/2)^{q/2} \\
&\leq r_n^{\check{q}} (q/2)^q.
\end{aligned}$$

Most importantly this does not depend on the q_j. We note that $\tilde{q}((q_j)_{j\in[r_n]}) = |\{j \in [r_n], q_j > 0\}|$. Using $2^k \leq r_n$, we obtain for the sum in (5.109) that

$$\begin{aligned}
&\sum_{q_1+\ldots+q_{r_n}=q/2} \binom{q/2}{q_1,\ldots,q_{r_n}} 2^{-k\tilde{q}((q_j)_{j\in[r_n]})} \\
&\quad= \sum_{\check{q}=1}^{q/2} 2^{-\check{q}k} \sum_{q_1+\ldots+q_{r_n}=q/2} \binom{q/2}{q_1,\ldots,q_{r_n}} \mathbb{I}\{\check{q} = |\{j \in [r_n], q_j > 0\}|\} \\
&\quad= \sum_{\check{q}=1}^{q/2} 2^{-\check{q}k} N(\check{q}) \\
&\quad\leq \sum_{\check{q}=1}^{q/2} 2^{-\check{q}k} r_n^{\check{q}} (q/2)^q \\
&\quad\leq (q/2)^{q+1} \left(\frac{r_n}{2^k}\right)^{q/2} \\
&\quad= \mathcal{O}\left(\left(r_n 2^{-k}\right)^{q/2}\right). \qquad (5.110)
\end{aligned}$$

Using that $\mathbb{E}[\varepsilon_1^q]$, C_X^q and c_X^{-q} are constants independent of n and k together with (5.110) and (5.109) we obtain

$$\mathbb{E}\left[h_{R,n}((X_i,\varepsilon_i)_{i=1}^{r_n},\omega)^q\right] = \mathcal{O}\left((r_n 2^{-k})^{-q}\right).$$

Thus, with (5.105) we end up with

$$\begin{aligned}
\mathbb{E}\left[\left(R^{(1)}_{n,r_n,\omega}(x_0)\right)^q\right] &\lesssim \left(\frac{r_n}{n}\right)^{q/2} \mathbb{E}\left[(h_{R,n}((X_i,\varepsilon_i)_{i=1}^{r_n},\omega))^q\right] \\
&\lesssim \left(\frac{r_n}{n}\right)^{q/2} \left(\frac{2^k}{r_n}\right)^q = \left(\frac{2^{2k}}{nr_n}\right)^{q/2}. \qquad \square
\end{aligned}$$

Proof of Lemma 5.19. We obtain

$$\begin{aligned}
&\mathbb{E}\left[R^{(2)}_{n,r_n,\omega}(x_0)^2\right] \\
&\quad = \mathrm{Var}\left(R^{(2)}_{n,r_n,\omega}(x_0)\right) \\
&\quad = \frac{1}{n^2}\mathrm{Var}\left(\sum_{j=1}^{n}\varepsilon_j \frac{1}{\binom{n-1}{r_n-1}} \sum_{I\in B_{r_n,n}:j\in I}\left(p_{x_0}(\omega_I)^{-1}\mathbb{I}\{X_j\in A_k(x_0,\omega_I)\}-K_k(x_0,X_j)\right)\right) \\
&\quad = \frac{\sigma^2}{n}\mathrm{Var}\left(\frac{1}{\binom{n-1}{r_n-1}} \sum_{I\in B_{r_n,n}:1\in I}\left(p_{x_0}(\omega_I)^{-1}\mathbb{I}\{X_1\in A_k(x_0,\omega_I)\}-K_k(x_0,X_1)\right)\right) \\
&\quad = \frac{\sigma^2}{n}\frac{1}{\binom{n-1}{r_n-1}}\mathrm{Var}\left(p_{x_0}(\omega)^{-1}\mathbb{I}\{X_1\in A_k(x_0,\omega)\}-K_k(x_0,X_1)\right)
\end{aligned}$$

because

$$\begin{aligned}
&\mathrm{Cov}\left(\frac{\mathbb{I}\{X_1\in A_k(x_0,\omega_1)\}}{p_{x_0}(\omega_1)}-K_k(x_0,X_1), \frac{\mathbb{I}\{X_1\in A_k(x_0,\omega_2)\}}{p_{x_0}(\omega_2)}-K_k(x_0,X_1)\right) \\
&\quad = \mathbb{E}\left[\left(\frac{\mathbb{I}\{X_1\in A_k(x_0,\omega_1)\}}{p_{x_0}(\omega_1)}-K_k(x_0,X_1)\right)\left(\frac{\mathbb{I}\{X_1\in A_k(x_0,\omega_2)\}}{p_{x_0}(\omega_2)}-K_k(x_0,X_1)\right)\right] \\
&\quad = 0.
\end{aligned}$$

The latter equality follows by conditioning on X_1 and using that

$$K_k(x_0,X_1)=\mathbb{E}\left[\frac{\mathbb{I}\{X_1\in A_k(x_0,\omega_1)\}}{p_{x_0}(\omega_1)}\,\middle|\, X_1\right]$$

in conjunction with the independence of ω_1 and ω_2. We get

$$\begin{aligned}
&\mathbb{E}\left[R^{(2)}_{n,r_n,\omega}(x_0)^2\right] \\
&\quad = \frac{\sigma^2}{n}\frac{1}{\binom{n-1}{r_n-1}}\mathrm{Var}\left(p_{x_0}(\omega)^{-1}\mathbb{I}\{X_1\in A_k(x_0,\omega)\}-K_k(x_0,X_1)\right) \\
&\quad = \frac{\sigma^2}{\kappa^2 n}\frac{1}{\binom{n-1}{r_n-1}}\mathbb{E}\left[\left(p_{x_0}(\omega)^{-1}\mathbb{I}\{X_1\in A_k(x_0,\omega)\}-K_k(x_0,X_1)\right)^2\right] \\
&\quad = \frac{\sigma^2}{n}\frac{1}{\binom{n-1}{r_n-1}}\left(\mathbb{E}\left[p_{x_0}(\omega)^{-2}\mathbb{I}\{X_1\in A_k(x_0,\omega)\}\right]-\mathbb{E}\left[K_k^2(x_0,X_1)\right]\right) \\
&\quad \le \frac{\sigma^2}{n}\frac{1}{\binom{n-1}{r_n-1}}\mathbb{E}\left[p_{x_0}(\omega)^{-2}\mathbb{I}\{X_1\in A_k(x_0,\omega)\}\right] \\
&\quad \le \frac{\sigma^2}{n}\frac{1}{\binom{n-1}{r_n-1}}c_X^{-2}2^{2k}\mathbb{E}\left[\mathbb{I}\{X_1\in A_k(x_0,\omega)\}\right] \\
&\quad \le \frac{\sigma^2}{n}\left(\frac{r_n}{n}\right)^{r_n-1}c_X^{-2}2^{2k}p_{x_0} \\
&\quad \le \frac{\sigma^2 C_X}{c_X^2}\frac{2^k}{n}\left(\frac{r_n}{n}\right)^{r_n-1}
\end{aligned}$$

$$= \frac{\sigma^2 C_X}{c_X^2} \frac{2^k}{r_n} \left(\frac{r_n}{n}\right)^{r_n}.$$

This yields the claim. □

Proof of Corollary 5.20. The proof of Corollary 5.20 is analogous to that of Lemma 5.19. In particular, we have

$$\begin{aligned}
&\mathbb{E}\left[R^{(K)}_{n,r_n,\omega}(x_0)^2\right] \\
&= \frac{\sigma^2}{p_{x_0}^2 n} \mathrm{Var}\left(\frac{1}{\binom{n-1}{r_n-1}} \sum_{I \in B_{r_n,n}: j \in I} \mathbb{I}\{X_1 \in A_k(x_0, \omega_I)\} - \mathbb{P}\left(X_1 \in A_k(x_0,\omega) \mid X_1\right) \right) \\
&= \frac{\sigma^2}{p_{x_0}^2 n} \frac{1}{\binom{n-1}{r_n-1}} \mathrm{Var}\left(\mathbb{I}\{X_1 \in A_k(x_0,\omega)\} - \mathbb{P}\left(X_1 \in A_k(x_0,\omega) \mid X_j\right)\right) \\
&\leq \frac{\sigma^2}{p_{x_0}^2 n} \left(\frac{r_n}{n}\right)^{r_n - 1} \mathbb{E}[\mathbb{I}\{X_1 \in A_k(x_0,\omega)\}] \\
&= \frac{\sigma^2}{p_{x_0} r_n} \left(\frac{r_n}{n}\right)^{r_n} \\
&\leq \frac{\sigma^2}{c_X} \frac{2^k}{r_n} \left(\frac{r_n}{n}\right)^{r_n}
\end{aligned}$$

because

$$\begin{aligned}
\mathrm{Cov}\big(&\mathbb{I}\{X_1 \in A_k(x_0,\omega_1)\} - \mathbb{P}(X_1 \in A_k(x_0,\omega) \mid X_1), \\
&\mathbb{I}\{X_1 \in A_k(x_0,\omega_2)\} - \mathbb{P}(X_1 \in A_k(x_0,\omega) \mid X_1) = 0
\end{aligned}$$

with the arguments from the previous proof. □

5.6.11 Proof of Theorem 5.10

The structure of the proof is similar to the proof for the random forest. More precisely, the proof is similar to that of Proposition 5.15 in Section 5.6.2. The claim then follows analogously to Section 5.6.1. We use the decomposition of the estimator and error from (2.3), that is

$$\begin{aligned}
\hat{m}_H(x_0) &= \hat{m}_H^{(m)}(x_0) + \hat{m}_H^{(\varepsilon)}(x_0) \\
&= \sum_{j=1}^n m(X_j) \frac{\mathbb{I}\{X_j \in A_\delta(x_0)\}}{\sum_{i=1}^n \mathbb{I}\{X_i \in A_\delta(x_0)\}} + \sum_{j=1}^n \varepsilon_j \frac{\mathbb{I}\{X_j \in A_\delta(x_0)\}}{\sum_{i=1}^n \mathbb{I}\{X_i \in A_\delta(x_0)\}}.
\end{aligned}$$

The assumption $0 < c_X \leq f_X \leq C_X$ and $\mathbb{V}(A_\delta(x_0)) = \delta^p$ imply

$$c_X \delta^p \leq p_{x_0}(\delta) \leq C_X \delta^p.$$

For

$$\tilde{m}_H^{(\varepsilon)}(x_0) = \frac{1}{n} \sum_{j=1}^n \varepsilon_j p_{x_0}(\delta)^{-1} \mathbb{I}\{X_j \in A_\delta(x_0)\},$$

we have

$$\operatorname{Var}(\tilde{m}_H^{(\varepsilon)}(x_0)) = \sigma^2 p_{x_0}(\delta)^{-1} n^{-1} = \Theta(n^{-1}\delta^{-p}).$$

First we consider the remainder term

$$\hat{m}_H^{(\varepsilon)}(x_0) - \tilde{m}_H^{(\varepsilon)}(x_0) = \sum_{j=1}^{n} \varepsilon_j \mathbb{I}\{X_j \in A_\delta(x_0)\} \left(\frac{1}{\sum_{i=1}^n \mathbb{I}\{X_i \in A_\delta(x_0)\}} - \frac{1}{n p_{x_0}(\delta)} \right).$$

Let $q \in 2\mathbb{Z}$. We note that $\delta^p = \mathcal{O}(n)$ by (5.29). This allows us to apply Lemma 2.19 and similar to the proof of Lemma 5.18, but only considering the part for the moments of the kernel, we get

$$\begin{aligned}
&\mathbb{E}\big[|\hat{m}_H^{(\varepsilon)}(x_0) - \tilde{m}_H^{(\varepsilon)}(x_0)|^q\big] \\
&= \mathbb{E}\left[\Big| \sum_{j=1}^{n} \varepsilon_j \mathbb{I}\{X_j \in A_\delta(x_0)\} \big(\frac{1}{\sum_{i=1}^n \mathbb{I}\{X_i \in A_\delta(x_0)\}} - \frac{1}{n p_{x_0}(\delta)} \big) \Big|^q \right] \\
&\lesssim \mathbb{E}\left[\left(\sum_{j=1}^{n} \varepsilon_j^2 \mathbb{I}\{X_j \in A_\delta(x_0)\} \left(\frac{1}{\sum_{i=1}^n \mathbb{I}\{X_i \in A_\delta(x_0)\}} - \frac{1}{n p_{x_0}(\delta)} \right)^2 \right)^{q/2} \right] \\
&\lesssim n^{q/2} \mathbb{E}\left[\prod_{j=1}^{q/2} \varepsilon_j^2 \mathbb{I}\{X_j \in A_\delta(x_0)\} \left(\frac{1}{\sum_{i=1}^n \mathbb{I}\{X_i \in A_\delta(x_0)\}} - \frac{1}{n p_{x_0}(\delta)} \right)^2 \right] \\
&\lesssim n^{q/2} \mathbb{E}\left[\prod_{j=1}^{q/2} \mathbb{I}\{X_j \in A_\delta(x_0)\} \right] \mathbb{E}\left[\left(\frac{1}{q/2 + \sum_{i=q/2+1}^n \mathbb{I}\{X_i \in A_\delta(x_0)\}} - \frac{1}{n p_{x_0}(\delta)} \right)^q \right] \\
&\lesssim (n\delta^p)^{q/2} \mathcal{O}\big((n\delta^p)^{-3q/2} \big) \\
&= \mathcal{O}\big((n\delta^p)^{-q} \big).
\end{aligned} \tag{5.111}$$

For the approximation error we have

$$\begin{aligned}
\hat{m}_H^{(m)}(x_0) - m(x_0) = \sum_{j=1}^{n} (m(X_j) - m(x_0)) \frac{\mathbb{I}\{X_j \in A_\delta(x_0)\}}{\sum_{i=1}^n \mathbb{I}\{X_i \in A_\delta(x_0)\}} \\
+ m(x_0)(\mathbb{I}\{\exists j \in [n] : X_j \in A_\delta(x_0)\} - 1).
\end{aligned} \tag{5.112}$$

Using the Hölder continuity we get

$$\begin{aligned}
\mathbb{E}\left[\sup_{x_0 \in [0,1]^p} \Big| \sum_{j=1}^{n} (m(X_j) - m(x_0)) \frac{\mathbb{I}\{X_j \in A_\delta(x_0)\}}{\sum_{i=1}^n \mathbb{I}\{X_i \in A_\delta(x_0)\}} \Big| \right] &\leq C_H \mathbb{E}\left[\sup_{x_0 \in [0,1]^p} \mathfrak{d}(A_\delta(x_0))^\alpha \right] \\
&= C_H \delta^\alpha p^{\alpha/2}.
\end{aligned} \tag{5.113}$$

The number of cells is equal to δ^{-p}. Let $\mathcal{X}_\delta \subset [0,1]^p$ with $|\mathcal{X}_\delta| = \delta^{-p}$ denote at set containing exactly one element in each cell. We obtain

$$\mathbb{E}\left[\sup_{x_0 \in [0,1]^p} |m(x_0)(\mathbb{I}\{\exists j \in [n] : X_j \in A_\delta(x_0)\} - 1)| \right]$$

$$
\begin{aligned}
&\leq \|m\|_\infty \mathbb{E}\left[\max_{x_0 \in \mathcal{X}_\delta} \mathbb{I}\{\nexists j \in [n] : X_j \in A_\delta(x_0)\}\right] \\
&\leq \|m\|_\infty \sum_{x_0 \in \mathcal{X}_\delta} \mathbb{E}\big[\mathbb{I}\{\nexists j \in [n] : X_j \in A_\delta(x_0)\}\big] \\
&= \|m\|_\infty \sum_{x_0 \in \mathcal{X}_\delta} (1 - p_{x_0}(\delta))^n \\
&\leq \|m\|_\infty \delta^{-p} (1 - c_X \delta^p)^n.
\end{aligned}
$$

Together with (5.112) and (5.113), this yields

$$
\mathbb{E}\big[\|m - \hat{m}_H^{(m)}\|_\infty\big] \lesssim C_H \delta^\alpha p^{\alpha/2} + \|m\|_\infty \delta^{-p} (1 - c_X \delta^p)^n. \tag{5.114}
$$

We proceed with the leading term. Let

$$
\mathcal{F}_\delta := \{f_{x_0,\delta}(x, s) = \sigma^{-1} s p_{x_0}(\delta)^{-1/2} \mathbb{I}\{x \in A_\delta(x_0)\}\}.
$$

We have

$$
\mathrm{Var}(f_{x_0,\delta}(X_1, \varepsilon_1)) = 1
$$

and

$$
\sup_{x_0 \in [0,1]^p} \mathbb{E}\left[|f_{x_0,\delta}(X_1, \varepsilon_1)|^3\right] \leq \frac{\mathbb{E}\left[|\varepsilon_1|^3\right]}{\sigma^3} (c_X \delta^p)^{-1/2}
$$

as well as

$$
\begin{aligned}
\mathbb{E}\left[\sup_{x_0 \in [0,1]^p} |f_{x_0,\delta}(X_1, \varepsilon_1)|^\nu\right]^{1/\nu} &= \mathbb{E}\left[\sup_{x_0 \in [0,1]^p} |\sigma^{-1} \varepsilon_1 p_{x_0}(\delta)^{-1/2}|^\nu\right]^{1/\nu} \\
&\leq (c_X \delta^p)^{-1/2} \sigma^{-1} \mathbb{E}[|\varepsilon_1|^\nu]^{1/\nu}.
\end{aligned}
$$

For b from Theorem 2.4 this yields $b \lesssim \delta^{-p/2}$. Further we have $|\mathcal{F}_\delta| = \delta^{-p}$ and hence we can choose $A = \delta^{-p}$ and $v = 1$ for A and v from the same theorem. We get

$$
\begin{aligned}
K_n &= cv(\log n \vee \log(Ab/\tilde{\sigma})) \\
&\lesssim c(\log n \vee \log(\delta^{-p} \delta^{-p/2})) \\
&\leq c(\log n \vee \log(\delta^{-p3/2})) \\
&\leq c \log(n^{3/2}) \\
&= \mathcal{O}(\log n).
\end{aligned}
$$

Analogously to Theorem 5.14 we get that

$$
\left|\sqrt{\frac{n}{\sigma^2}} \sup_{x_0 \in [0,1]^p} |p_{x_0}(\delta)^{1/2} \tilde{m}_H^{(\varepsilon)}(x_0)| - \mathbf{S}_\delta\right| = \mathcal{O}_{\mathbb{P}}\left(\frac{(\log n)^{3/2}}{\delta^{p/2} n^{1/2 - 1/q_G}} + \frac{(\log n)^{5/4}}{\delta^{p/4} n^{1/4}} + \frac{\log n}{\delta^{p/6} n^{1/6}}\right). \tag{5.115}
$$

because $q_G \geq 4$. Here $\mathbf{S}_\delta \overset{d}{=} \sup_{x_0 \in [0,1]^p} |B_\delta f_{x_0,\delta}|$ and B_δ is a sequence of centered Gaussian processes indexed by $\mathcal{F}_\delta$ and with covariance function

$$
\mathrm{Cov}(B_\delta(f_{x_1,\delta}), B_\delta(f_{x_2,\delta})) = p_{x_1}(\delta)^{-1/2} p_{x_2}(\delta)^{-1/2} \mathbb{E}\left[\mathbb{I}\{X_1 \in A_\delta(x_1)\} \mathbb{I}\{X_1 \in A_\delta(x_2)\}\right]
$$

$$= \mathbb{I}\{A_\delta(x_1) = A_\delta(x_2)\}.$$

That means

$$\mathbf{S}_\delta \overset{d}{=} \sup_{x_0 \in [0,1]^p} |B_\delta f_{x_0,\delta}| \overset{d}{=} \max_{j=1,\dots,\delta^{-p}} |Z_j|$$

where the $Z_j \sim \mathcal{N}(0,1)$ and i.i.d.. We get confidence bands with radius

$$\sigma c_\delta(\beta) \sqrt{\frac{1}{n p_x(\delta)}}$$

where $c_\delta(\beta)$ satisfies

$$\mathbb{P}\left(\max_{j=1,\dots,\delta^{-p}} |Z_j| \leq c_\delta(\beta) \right) = 1 - \beta.$$

(5.29) implies that

$$\frac{(\log n)^{5/2}}{\delta^{p/2} n^{1/2 - 1/q_G}} = \left(\frac{(\log n)^5}{\delta^p n} n^{2/q_G} \right)^{1/2} \to 0.$$

Similar to the proof of Theorem 5.1, the same argument applied to the other terms in (5.115) implies that

$$\left| \sqrt{\frac{n}{\sigma^2}} \sup_{x_0 \in [0,1]^p} |p_{x_0}(\delta)^{1/2} \tilde{m}_H^{(\varepsilon)}(x_0)| - \mathbf{S}_\delta \right| = o_{\mathbb{P}}((\log n)^{-1}). \tag{5.116}$$

It remains to show that the remainder and the approximation error are negligible. Similar to the main proof we need that they are

$$o_{\mathbb{P}}\left(\frac{1}{n^{1/2} \delta^{p/2} \log n} \right).$$

Using a union bound, (5.111) yields

$$\begin{aligned}
&\mathbb{P}\left(\|\hat{m}_H^{(\varepsilon)} - \tilde{m}_H^{(\varepsilon)}\|_\infty \geq \kappa \frac{1}{n^{1/2} \delta^{p/2} \log n} \right) \\
&\leq \delta^{-p} \mathbb{E}\big[|\hat{m}_H^{(\varepsilon)}(x_0) - \tilde{m}_H^{(\varepsilon)}(x_0)|^{q_R} \big] \kappa^{-q_R} (n \delta^p (\log n))^{q_R/2} \\
&= \mathcal{O}\left(\delta^{-p} (\frac{\log n}{\delta^p n})^{q_R/2} \right) \to 0
\end{aligned} \tag{5.117}$$

due to (5.28). For the approximation error we have with (5.114) that

$$\begin{aligned}
&\mathbb{P}\left(\|m - \hat{m}_H^{(m)}\|_\infty \geq \kappa \frac{1}{n^{1/2} \delta^{p/2} \log n} \right) \\
&\leq \mathbb{E}\big[\|m - \hat{m}_H^{(m)}\|_\infty \big] \kappa^{-1} (n \delta^p (\log n))^{1/2} \\
&= \mathcal{O}\big((\delta^{2\alpha} n \delta^p (\log n))^{1/2} \big) + \mathcal{O}\big(\delta^{-p} (1 - c_X \delta^p)^n (n \delta^p (\log n))^{1/2} \big).
\end{aligned}$$

The first term is $o(1)$ with (5.27) and for the second term we have

$$\delta^{-p} (1 - c_X \delta^p)^n (n \delta^p (\log n))^{1/2} \lesssim \exp(-c_X n \delta^p) \delta^{-p} (n \delta^p (\log n))^{1/2} \to 0$$

because $\delta^{-p} = o_{\mathbb{P}}(n)$ with (5.28) or (5.29). Combining this with (5.116) and (5.117), the claim follows analogously to the proof of Theorem 5.1 based on Proposition 5.15. □

Chapter 6

Simulation study

In this chapter, we present a simulation study to verify our theoretical results in practice. The complete source code related to this chapter is available, see Rabe (2025). Throughout the study we use uniformly distributed X. In Section 5.4 we discussed properties of the Gaussian process that are utilized in this chapter. In particular, we know the covariance of the Gaussian process for uniformly distributed X from (5.31). It is given by

$$\operatorname{Cov}(B_k(f_{x_1,k}), B_k(f_{x_2,k})) = \mathcal{V}_{\cap,k}^{-1}\mathbb{E}\left[\mathbb{V}\left(A_k(x_1,\omega_1)\cap A_k(x_2,\omega_2)\right)\right]. \tag{6.1}$$

In Section 6.1 we illustrate and discuss several aspects of the asymptotic confidence bands. Subsequently, in Section 6.2 we introduce a possible approach to use Bootstrap confidence bands and compare it to the asymptotic bands for a small sample size. At the end of the chapter we discuss some limitations of the simulation study in Section 6.3.

6.1 Asymptotic confidence bands

This section deals with the asymptotic confidence bands given by Theorem 5.1. To construct them, we need to know $\mathcal{V}_{\cap,k}$, σ, and most importantly, the quantiles of $\mathbf{S}_k$ in addition to the estimator. To estimate the distribution of $\mathbf{S}_k$, we can use the observations from Section 5.4. First, we can estimate the distinct entries in the covariance matrix of B_k. Knowing the covariances, we can get the covariance matrix on some grid in $[0,1]^p$. Due to computational costs it is not always feasible to use the 2^k-grid if kp is too large. On this grid we simulate the Gaussian process B_k and compute its supremum. Using the Monte Carlo method, we get an empirical distribution of $\mathbf{S}_k$.

6.1.1 Estimation of the covariance matrix

Lemma 5.11 and Lemma 5.12 provide insights into the estimation of the covariances of the Gaussian process when X is uniformly distributed. It is not necessary to simulate the X because of the identity in (5.31). However, it should be noted that this is equivalent to knowing the density of X.

To estimate the covariance from (6.1), we need to simulate a large number of pairs of independent vectors containing the number of splits per direction for two different cells $A_k(x_1,\omega_1)$ and $A_k(x_2,\omega_2)$. Lemma 5.11 allows us to compute the volume of their

Algorithm 6.1 Covariance Estimation

Input: Number of partitions n_P, number of splits k, split distribution of ω.

$V_{cum} = \{0\}^{\binom{k+p}{p}}$
for $j = 1, \ldots, n_P$ **do**
  Simulate $S(\omega_1) \in \{0, \ldots, k\}^p$ according to the cell split distribution.
  Simulate $S(\omega_2) \in \{0, \ldots, k\}^p$ according to the cell split distribution.
  for $i = 1, \ldots, \binom{k+p}{p}$ **do**
    $V_{cum}(i) = V_{cum}(i) + \mathbb{V}_{\cap}(\tau^{-1}(i), S(\omega_1), S(\omega_2))$
  end for
end for
$\bar{V} = V_{cum}/n_P$
$\hat{\mathcal{V}}_{\cap,k} = \bar{V}(\tau(\{k\}^p))$
$\hat{C} = \bar{V}/\hat{\mathcal{V}}_{\cap,k}$
return $\hat{C}$ and $\hat{\mathcal{V}}_{\cap,k}$

Output: Estimators $\hat{C}(\tau(\mathbf{c}))$ of $\mathrm{Cov}(\mathbf{c})$ for $\mathbf{c} \in \Omega_S$ and $\hat{\mathcal{V}}_{\cap,k}$ of $\mathcal{V}_{\cap,k}$.

intersection from the number of splits $S(x_1, \omega_1)$ and $S(x_2, \omega_2)$. Thus, we can estimate the expected volume, which is equal to the corresponding covariance multiplied by $\mathcal{V}_{\cap,k}$. For $x_1 = x_2$ this is already an estimator for $\mathcal{V}_{\cap,k}$.

Lemma 5.12 implies that there are only $\binom{k+p}{p}$ distinct covariance values for $x_1, x_2 \in [0,1]^p$ and thus further simplifies the estimation. Precisely, it is sufficient to estimate the covariance for all possible values of the ordered version of the closeness vector $\mathfrak{C}_k(x_1, x_2)$ of x_1 and x_2 defined in (5.32). The ordered vectors $\mathfrak{C}_k^{(\uparrow)}(x_1, x_2)$ from Section 5.4.2 are elements of

$$\Omega_S = \{(t_1, \ldots, t_p) \mid t_i \in \{0, \ldots k\} \forall i \in \{1, \ldots, p\}, t_1 \leq t_2 \leq \ldots \leq t_p\}.$$

Instead of calculating the volume of the intersection for all pairs of x_1 and x_2 it is sufficient to do so for all elements of Ω_S. From (5.33) we know that this is

$$\mathbb{V}_{\cap}(\mathbf{c}, S(\omega_1), S(\omega_2)) = 2^{-\left(\sum_{l=1}^{p} \max(S_l(\omega_1), S_l(\omega_2))\right)} \prod_{l=1}^{p} \mathbb{I}\{\min(S_l(\omega_1), S_l(\omega_2)) \leq \mathbf{c}_l\}.$$

Algorithm 6.1 describes the estimation of the covariances

$$\mathrm{Cov}(\mathbf{c}) := \frac{1}{\mathcal{V}_{\cap,k}} \mathbb{E}\left[\mathbb{V}_{\cap}(\mathbf{c}, S(\omega_1), S(\omega_2))\right], \ \mathbf{c} \in \Omega_S.$$

To order the elements of Ω_S, we use some bijection

$$\tau : \Omega_S \to \left\{1, \ldots, \binom{k+p}{p}\right\}.$$

The algorithm gives estimators $\hat{C}(\tau(\mathbf{c}))$ for $\mathrm{Cov}(\mathbf{c})$ and an estimator for $\mathcal{V}_{\cap,k}$. We note that

$$\hat{\mathcal{V}}_{\cap,k} = \bar{V}\left(\tau(\{k\}^p)\right),$$

because we have $\mathfrak{C}_k(x_0, x_0) = \{k\}^p$.

We need to simulate the Gaussian process on an appropriate grid of the feature space. Thus, we need an estimator of its covariance matrix on this grid. When the covariance matrix Σ of a Gaussian process on a grid is known, the process can be simulated on that grid by multiplying the matrix L that satisfies $\Sigma = LL^T$ with a vector that is multivariate normal distributed with identity covariance matrix.

Above we already explained the estimation of all distinct covariance matrix entries. Below we describe how these entries are utilized to form the estimator for the covariance matrix. For some $\tilde{k} \leq k$ we use an equidistant grid with spacing of length $2^{-\tilde{k}}$. The grid is then defined as

$$G_{\tilde{k}} := \{a2^{-\tilde{k}} + 2^{-\tilde{k}+1} \mid a \in \{0, \dots, 2^{\tilde{k}} - 1\}\}^p.$$

The true covariance matrix on this grid is

$$\Sigma_{\tilde{k}} := \frac{1}{\mathcal{V}_{\cap,k}} \left(\mathbb{E}\left[\mathbb{V}\left(A_k(x_1, \omega_1) \cap A_k(x_2, \omega_2)\right)\right]\right)_{x_1, x_2 \in G_{\tilde{k}}}.$$

We recall that $\mathfrak{C}_k^{(\uparrow)}(x_1, x_2)$ is the order statistic of $\mathfrak{C}_k(x_1, x_2)$. The estimator of the covariance matrix on the same grid is

$$\hat{\Sigma}_{\tilde{k}} := \left(\hat{C}\left(\tau\left(\mathfrak{C}_k^{(\uparrow)}(x_1, x_2)\right)\right)\right)_{x_1, x_2 \in G_{\tilde{k}}}.$$

It remains to discuss how we select $\tilde{k}$ and thereby the grid for the simulation of the Gaussian process. The observations about the function class in Section 5.4 imply that the process can be seen as a process in $\mathcal{X}_k$, which is a set that contains one point in every undividable cell. We know that a 2^k-grid would always contain one point in every undividable cell. However, the use of a covariance matrix on a very fine grid can be computationally infeasible. As a way out, Remark 5.13 can be used to determine which grid is fine enough. Since we estimate the covariance for all $\mathbf{c} \in \Omega_S$ we can check which estimated covariance entries are equal to one. If there is a $c^* \in \{0, \dots, k\}$ with

$$\mathbb{P}\left(S_l(\omega_1) \leq c^*\right) = 1 \; \forall l \in \{1, \dots, p\},$$

the grid with $\tilde{k} = c^*$ is fine enough because its grid cells are undividable cells (see Section 3.3.1.3). This implies that all points in these cells correspond to the same function in $\mathcal{F}_k$. Thus, it suffices to simulate the Gaussian process on this grid to estimate the distribution of the supremum. An equivalent perspective is, that the estimator and the stochastic error are constant on these cells and thus it is sufficient to check one point in each cell to calculate the supremum of the stochastic error.

In practice, we can choose $\tilde{k}$ such that the covariance of the Gaussian process at x_1 and x_2 with $\mathfrak{C}_k(x_1, x_2) = \{\tilde{k}\}^p$, which is

$$\frac{1}{\mathcal{V}_{\cap,k}} \mathbb{E}\left[\mathbb{V}_\cap\left(\{\tilde{k}\}^p, S(\omega_1), S(\omega_2)\right)\right],$$

is reasonably close to one. In this case the probability $\mathbb{P}(S_l(\omega_1) \leq \tilde{k})$ should also be close to one for every l. With this method we can determine the finest grid spacing that is

necessary. If this grid is computationally feasible we choose $\tilde{k}$ accordingly. Otherwise, we can use the method to get an idea how good the best computationally feasible grid is.

It may be the case that the estimated covariance $\hat{\Sigma}_{\tilde{k}}$ is not positive semidefinite, which would be necessary for the Cholesky decomposition. Therefore, we slightly adjust the estimated covariance matrix. The first adjustment we try is to add the identity matrix multiplied by a small constant to the covariance matrix before computing the Cholesky decomposition. If this is not sufficient to get a positive semidefinite matrix, we use a reconstruction of the covariance matrix using only the eigenvalues that are greater than a small constant. Let $Q_{\tilde{k}}$ denote a matrix with the eigenvectors of $\hat{\Sigma}_{\tilde{k}}$ as its columns and let $(\lambda_i)_{i=1}^{2^{p\tilde{k}}}$ be the corresponding eigenvalues. Let Λ_ϵ be a diagonal matrix with entries $(\max\{\lambda_i, \epsilon\})_{i=1}^{2^{p\tilde{k}}}$. We use the Cholesky decomposition of the reconstruction

$$\hat{\Sigma}_{\tilde{k},2} := Q_{\tilde{k}} \Lambda_\epsilon Q_{\tilde{k}}^T$$

of $\hat{\Sigma}_{\tilde{k}}$ for the simulations of the suprema.

6.1.2 Estimation of σ

To apply the asymptotic confidence bands in practice we need an estimator for σ. There are different types of estimators in the literature. A standard approach is to use the residuals

$$\hat{\varepsilon}_i := Y_i - \hat{m}(X_i),$$

where $\hat{m}$ is an estimator of the regression function, and estimate σ^2 by

$$\hat{\sigma}^2 := \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

Lemma 5.4 shows that an estimator based on the residuals of $U_{n,r_n,\omega}^{(\mathrm{RF})}$ fulfills the assumptions in Theorem 5.1. It is possible to circumvent the dependence on m by employing an estimator that is not based on residuals. For instance, Müller et al. (2003) have examined estimators of this type. Let K_h be a univariate kernel with bandwidth h. We define the estimator

$$\check{\sigma}_K^2 = \frac{\sum_{j=1}^{n} \sum_{i=1}^{j-1} K_h(\|X_i - X_j\|)(Y_i - Y_j)^2/2}{\sum_{j=1}^{n} \sum_{i=1}^{j-1} K_h(\|X_i - X_j\|)}. \tag{6.2}$$

The kernel gives the pairs of observations weights that are large if the independent variables are close to each other in Euclidean distance. The idea behind the estimator is that

$$Y_1 - Y_2 \approx \varepsilon_1 - \varepsilon_2,$$

if X_1 and X_2 are close. This is the case because

$$Y_1 - Y_2 = m(X_1) + \varepsilon_1 - m(X_2) - \varepsilon_2$$

and

$$|m(X_1) - m(X_2)| \leq C\|X_1 - X_2\|^\alpha.$$

Further we note that

$$\mathbb{E}\left[(\varepsilon_1 - \varepsilon_2)^2\right] = 2\sigma^2.$$

Results for estimators of this type in the univariate case can be found in the work by Müller et al. (2003) or Shen et al. (2020). The latter work also includes results for product kernels in the multivariate case. In our applications product kernels performed worse than the estimator from (6.2). Possible choices for K are the densities of the continuous uniform distribution on $[-1, 1]$ and the standard normal distribution. In practice the choice of the bandwidth depends on the sample size and on σ.

6.1.3 Simulation results

In this section we present and discuss the results for the asymptotic confidence bands and their empirical coverage in a finite sample simulation setting. We want to investigate the effects suggested by the theory in practice. We want to consider the effects of σ, the error distribution, the number of trees N, and the dimension p on the empirical coverage of the confidence bands. We are also interested in whether the asymptotic assumptions can explain the results for different finite sample parameter combinations. When discussing the assumptions that include the approximation error we need to keep in mind that our bound via the expected diameter might not capture the behavior of the approximation error completely, as we discussed in Section 3.3.1.4.

For all simulations, we use 50 000 simulated pairs of splits to estimate the covariance matrix entries. We simulate 100 000 suprema of the Gaussian process with the estimated covariance matrix from Section 6.1.1 to estimate the quantiles of the supremum. Throughout the simulations, the bandwidth $h = n^{-1/2}$ and a Gaussian kernel are employed for the estimation of the standard deviation, as outlined in (6.2), since this choice worked well in our specific application. With the estimated quantiles, the estimation of $\mathcal{V}_{\cap,k}$ and the estimated standard deviation of the errors we are able to compute the confidence band radius.

The grid on which we evaluate the "supremum" of $|U^{(\mathrm{RF})}_{n,r_n,\omega}(x_0) - m(x_0)|$ is related to the grid on which we simulate the Gaussian process. The error is evaluated in the corners of the cells or hyperrectangles, respectively, that are split with probability equal to or close to zero. We use the grid

$$G^{(sup)}_{\tilde{k}} := \left(\{a2^{-\tilde{k}} - \epsilon, a2^{-\tilde{k}} + \epsilon \mid a \in \{1, \ldots, 2^{\tilde{k}} - 1\}\} \cup \{\epsilon, 1 - \epsilon\}\right)^p \tag{6.3}$$

for a small $\epsilon > 0$. The error between a smooth function and a piecewise constant function is the largest at the jump points of the piecewise function, if the smooth function is monotone on the corresponding intervals. Thus, the above grid should capture the almost full uniform error. The fact that at least one grid point is in each almost-undividable cell ensures that the stochastic error, which needs to be the dominating error for confidence bands is fully captured.

We start with simulations in the case $p = 2$. We use the regression function

$$m(x^{(1)}, x^{(2)}) = \frac{1}{10}(\sin(2\pi x^{(1)}) + x^{(2)}). \tag{6.4}$$

Table 6.1: Empirical coverage (top) and average confidence band radius (Rd., bottom) of 1000 RF confidence bands with $k = 5$, $n = 2000$ and $r_n = 1500$ for different error distributions.

σ	RF	N	Normal	Uniform	t-dist. 4 df	t-dist. 6 df
				$1-\beta$		
			.90, .95, .99	.90, .95, .99	.90, .95, .99	.90, .95, .99
0.75	Uni.	50	.656, .782, .941	.679, .816, .954	.617, .755, .911	.606, .752, .924
		100	.682, .812, .946	.697, .828, .957	.636, .772, .923	.621, .760, .943
		Rd.	.228, .243, .272	.228, .243, .273	.228, .242, .272	.228, .243, .272
	Ehr.	50	.648, .809, .945	.681, .825, .958	.646, .778, .918	.620, .767, .933
		100	.675, .810, .953	.702, .826, .953	.646, .786, .925	.620, .767, .931
		Rd.	.236, .252, .283	.236, .252, .283	.235, .251, .282	.236, .252, .282
1	Uni.	50	.734, .839, .961	.750, .884, .970	.699, .821, .935	.689, .815, .952
		100	.748, .870, .965	.773, .873, .975	.722, .816, .941	.707, .830, .953
		Rd.	.304, .324, .363	.304, .324, .363	.304, .323, .363	.304, .323, .363
	Ehr.	50	.745, .866, .970	.760, .876, .967	.710, .825, .944	.711, .822, .954
		100	.754, .866, .966	.770, .885, .971	.724, .824, .942	.701, .831, .952
		Rd.	.314, .336, .377	.314, .336, .377	.314, .335, .376	.314, .335, .376
1.25	Uni.	50	.777, .871, .969	.789, .910, .975	.735, .841, .945	.738, .842, .956
		100	.793, .897, .969	.803, .908, .978	.749, .838, .951	.748, .859, .957
		Rd.	.380, .404, .454	.380, .405, .454	.379, .404, .454	.380, .404, .454
	Ehr.	50	.790, .888, .973	.804, .894, .975	.752, .848, .950	.740, .858, .961
		100	.795, .888, .971	.804, .908, .979	.756, .854, .956	.753, .865, .965
		Rd.	.393, .419, .471	.393, .420, .471	.392, .419, .470	.392, .419, .470

The covariance matrix for both RF types has to be estimated depending on k. The simulations in Table 6.1 are for $k = 5$, $n = 2000$ and $r_n = \frac{3}{4}n = 1500$. We consider a normal, a uniform and two t-distributions for the error distribution. The table shows the empirical coverage and the average radius of the confidence bands. The only variation in the radii comes from the estimation of σ. For a fixed error distribution and σ, we use the same random seed for the training sample generation, to achieve a better comparison. Thus, the uniform and Ehrenfest CPRFs work on the same data for both values of N . The independence of the estimation of σ and m explains why $\hat{\sigma}$ does not vary with N.

The empirical coverage notably differs from the theoretical coverage. The coverage of the uniform RF and the Ehrenfest RF are generally similar. There is no clear trend as to which has better coverage. However, the radius of the confidence bands is larger for the Ehrenfest forest. This is the case because the variance of the estimator scales linearly in $\mathcal{V}_{\cap,k}$, which is larger in this case. As σ increases, we can see that the coverage

increases. The reason for this effect is the increase of the stochastic error in comparison to the approximation error and bias.

The variation of N has more than one effect. On average, a larger N increases the coverage. This effect is larger for the uniform RF than the Ehrenfest RF. This is explained by the characteristics of the two algorithms. The uniform CPRF has more diverse partitions, and thus one would expect it to require a larger N to create this diversity in practice. For $N = 100$, the coverage of the uniform CPRF is close to that of the Ehrenfest CPRF across the entire range of parameters. This is noteworthy because, at first glance, it seems to contradict the larger approximation error bound. Again, a possible reason could be the more diverse partitions. In Section 3.3.1.4 we discussed that this can lead to a smaller approximation error due to the smaller undividable cells, even though the approximation error bound does not capture this.

The sole condition imposed on the error distribution in Theorem 5.1 is the existence of a sufficient number of moments. In practical applications, however, disparities in the empirical coverage of the bands across different error distributions are to be anticipated. In particular, we would expect Gaussian errors to work well because the corresponding empirical process will already be closer to a Gaussian process than the empirical processes for other error distributions.

For the uniform error distribution, the difference in the coverage between $N = 50$ and $N = 100$ is very small, or in some instances even in favor of $N = 50$. This may suggest that a smaller number of trees is sufficient if the error distribution has lighter tails and the outliers are less relevant. The results for the t-distributions show this effect to a lesser extent. Their coverage improves similarly to that of the normal distribution as N increases. Across the board, the coverage for the uniform distribution is the best, explained by its compact support. This suggests that an error distribution with a shape similar to the normal distribution is not necessary due to the asymptotics. The coverage for the t-distribution with six degrees of freedom is worse than that for the normal distribution, but the difference is moderate. This suggests that a larger number of moments is helpful, but a finite number is sufficient, which is in line with the theory.

Figure 6.1 shows the errors of ten uniform CPRF estimators in the case $n = 2000$, $k = 5$ and $\varepsilon \sim \mathcal{N}(0, 1)$, which is also covered in Table 6.1. The vertical placement of the blue dots is equal to $\hat{m}(x_0) - m(x_0)$ where $\hat{m}$ is one of the ten estimators. The horizontal axis corresponds to the test grid of the feature space from (6.3) on which the uniform error is evaluated. It is important to note that this grid is not equidistant and is a subset of the two-dimensional set $[0, 1]^2$. Blocks of 32 values on the horizontal axis correspond to entries in the grid, that have the same first component. The radii of the confidence bands are shown in red. The 0.99 and 0.95 confidence bands cover nine of these ten estimators, the 0.9 confidence band covers eight. The plot provides insight into the dispersion of errors across the width of the confidence bands, demonstrating that errors outside the bands do not constitute extreme outliers relative to the remaining observations.

In Table 6.2 we showcase the empirical coverage for different values of n and k. For all results, the error distribution and N are the same and we have $r_n = 0.75n$. The assumptions in Theorem 5.1 suggest two important effects of n and k. If 2^k is large in relation too n, the distribution of $U^{(\varepsilon)}_{n,r_n,\omega}$ is not well approximated by the asymptotic in the theorem. If 2^k is small relative to n, the approximation error will be larger. Both

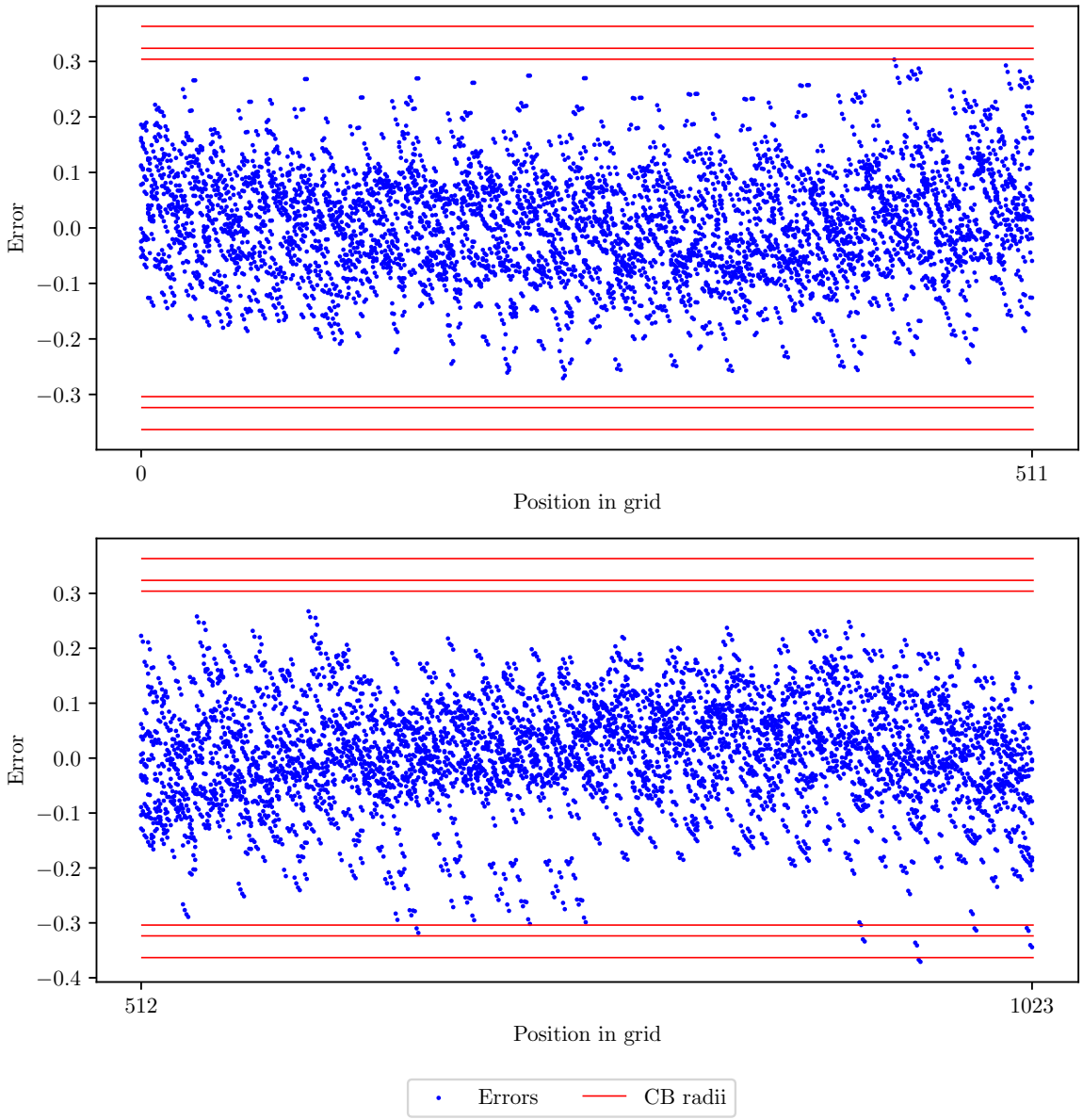

Figure 6.1: Estimation errors of ten uniform CPRF estimators on a non equidistant test grid and confidence band radii.

effects can lead to a worse coverage than the theoretical one. In the table we can see that the best coverage for $k = 5$ is achieved at $n = 1000$ (and $n = 2000$ for the 99% bands), while in the case of $k = 6$ the best coverage is achieved at $n = 4000$. This is in line with the theory. For the smaller values of n we see a suboptimal coverage for both values of k. This suggests that these values of n are too small for a good approximation by the asymptotic distribution. For the largest values of n we get the smallest radii of the confidence bands, and we can see that this affects the coverage negatively because the approximation error is larger in comparison to the stochastic error. This effect is more pronounced for the lower confidence levels. This is the case because the approximation

Table 6.2: Empirical coverage and average confidence band radius of 1000 RF confidence bands with $\varepsilon \sim \mathcal{N}(0,1)$, $N = 100$ and $r_n = \frac{3}{4}n$.

n	RF	$k = 5$		$k = 6$	
		$1-\beta$			
		.90, .95, .99 Coverage	.90, .95, .99 Radius	.90, .95, .99 Coverage	.90, .95, .99 Radius
250	Uni.	.676, .778, .916	.859, .915, 1.027	.599, .712, .877	1.280, 1.351, 1.495
	Ehr.	.668, .781, .909	.889, .949, 1.065	.627, .733, .897	1.333, 1.408, 1.565
500	Uni.	.759, .852, .954	.608, .648, .727	.611, .756, .901	.906, .956, 1.058
	Ehr.	.770, .856, .946	.629, .672, .754	.629, .759, .903	.944, .997, 1.108
1000	Uni.	.802, .878, .963	.430, .458, .514	.778, .867, .962	.641, .676, .748
	Ehr.	.805, .889, .964	.445, .475, .533	.760, .870, .955	.667, .705, .783
2000	Uni.	.748, .870, .965	.304, .324, .363	.786, .881, .967	.453, .478, .529
	Ehr.	.754, .866, .966	.314, .336, .377	.806, .887, .969	.472, .498, .554
4000	Uni.	.658, .797, .947	.215, .229, .257	.814, .898, .967	.320, .338, .374
	Ehr.	.674, .809, .946	.222, .237, .266	.813, .897, .968	.333, .352, .391
8000	Uni.	.450, .640, .873	.152, .162, .182	.763, .856, .967	.226, .239, .264
	Ehr.	.463, .639, .877	.157, .168, .188	.756, .870, .966	.236, .249, .277

error affects the center of the error distribution more than its tails. We can see that the radii of the bands increase when we compare a fixed k and n with $k+1$ and $2n$. It holds that $2^k/n = 2^{k+1}/2n$ and therefore the increase of the radii implies that $c_k(\beta)$ grows faster than $2^k\mathcal{V}_{\cap,k} \leq 1$ decreases. The reason for this is that one has equally many observations per cell on average, but more cells in total. Therefore, the maximum stochastic error will be larger on average.

Table 6.3 shows the results for the histogram confidence bands in the same regression settings as Table 6.2. The histogram confidence bands are evaluated for two choices of δ that lead to similar radii than those of the RF confidence bands in Table 6.2 for the same n. In general, one observes that the histogram confidence bands have a lower coverage. Figure 6.2 adds a graphical comparison to the random forest confidence bands. In the subplots on the left, we observe the empirical coverage and the nominal coverage in red. On the right, the radii are plotted on a logarithmic scale. From top to bottom, the plots are for $1-\beta = 0.9, 0.95, 0.99$. For all β, we observe that the radii of the two random forest bands for $k = 5$ are slightly smaller than the radii of the histogram bands for $\delta^{-1} = 5$. At the same time, the empirical coverage for these random forests is larger than that of the histogram for almost all n. The same observation can be made, when comparing the random forests for $k = 6$ and the histogram for $\delta^{-1} = 7$. This illustrates that the random forest bands work better throughout.

Figure 6.3 shows the histogram and uniform CPRF estimator of m from (6.4) for $n = 8000$, $k = 5$, $\delta^{-1} = 5$ for the same parameters as in Table 6.2 and Table 6.3. The

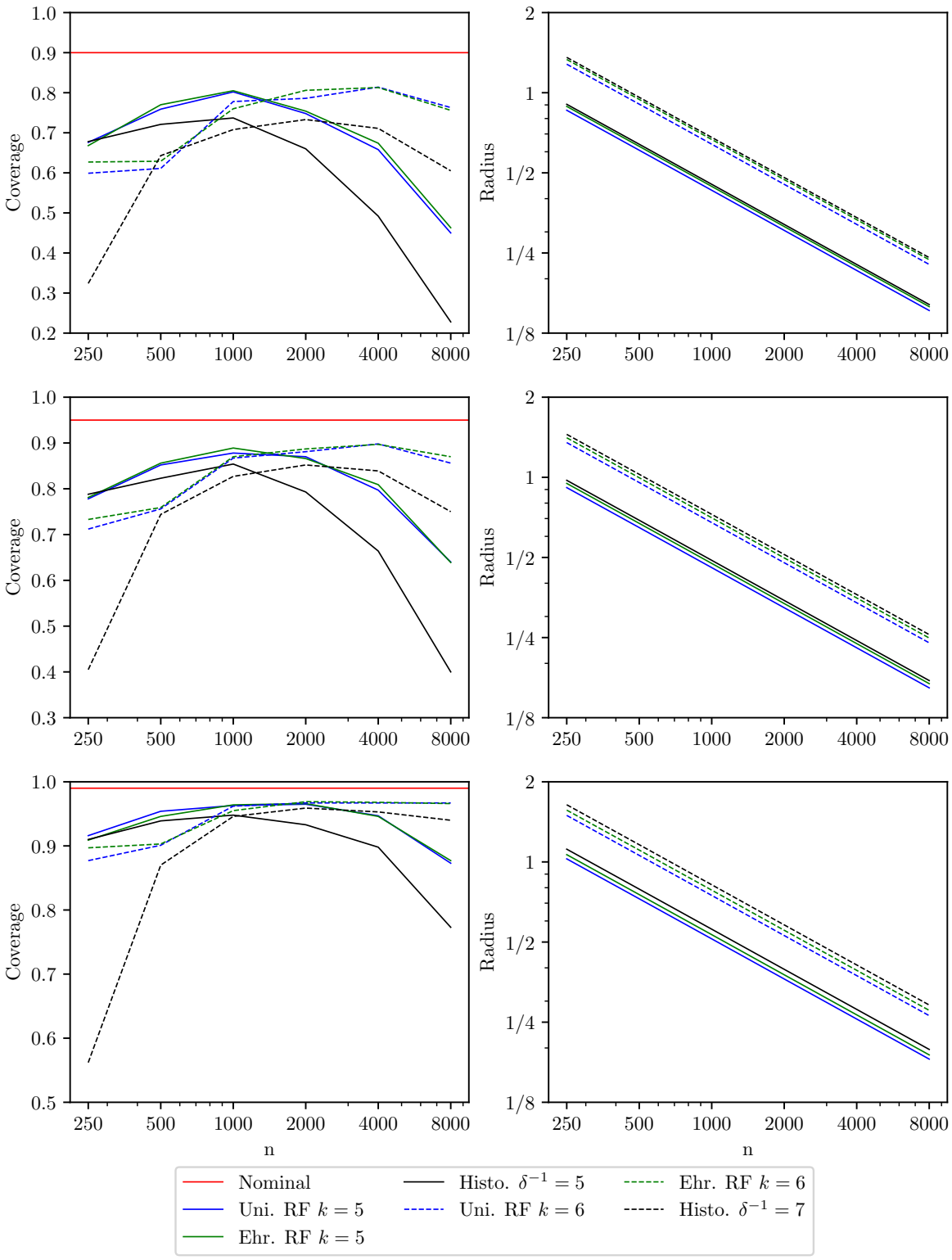

Figure 6.2: Comparison of CB coverage and radii for $1 - \beta \in \{0.9, 0.95, 0.99\}$ depending on n.

Table 6.3: Empirical coverage and average confidence band radius of 1000 Histogram confidence bands with $\varepsilon \sim \mathcal{N}(0,1)$.

n	$\delta^{-1}=5$		$\delta^{-1}=7$	
	$1-\beta$			
	.90, .95, .99 Coverage	.90, .95, .99 Radius	.90, .95, .99 Coverage	.90, .95, .99 Radius
250	.678, .788, .910	.905, .974, 1.117	.324, .405, .562	1.358, 1.450, 1.640
500	.721, .823, .939	.640, .690, .790	.643, .744, .870	.961, 1.026, 1.161
1000	.737, .854, .948	.453, .488, .559	.708, .827, .946	.680, .726, .821
2000	.660, .793, .933	.320, .345, .395	.733, .852, .959	.480, .513, .580
4000	.492, .664, .898	.226, .244, .279	.711, .839, .953	.340, .363, .410
8000	.228, .400, .773	.160, .172, .197	.605, .750, .940	.240, .256, .290

upper right plot shows a histogram estimation on the 2^k-grid, illustrating the data used for both estimators. We note that the scale of this plot is different from the other three.

In Table 6.4 we compare the empirical coverage for two regression models with $p=2$ and $p=4$ for different values of σ. A graphical representation of the results can be found in Figure 6.4. Similar to the previous figure it includes the three values of β from top to bottom. The regression function in the first case is the same as before. In the case of $p=4$ we used

$$m(x^{(1)},x^{(2)},x^{(3)},x^{(4)}) = \frac{1}{10}(\sin(2\pi x^{(1)}) + x^{(2)} + x^{(3)}x^{(4)}).$$

In Theorem 5.1 the dimension p mainly effects the assumption on the diameter which controls the approximation error. Further, p affects the value of $\mathcal{V}_{\cap,k}$ and $N_f(k)$. The approximation error increases with p, which is standard in multivariate regression problems due to the curse of dimensionality. The size of σ directly affects the signal to noise ratio and thus the size of the approximation and the stochastic error. If the noise level is decreased similarly for $p=2$ and $p=4$, the empirical coverage decreases faster in the latter case. This is as expected due to the larger approximation error. We point out that the confidence band radii are smaller for $p=4$. This is due to the effect on $\mathcal{V}_{\cap,k}$. When the empirical coverage is evaluated in a regression model with $m=0$, that means one only considers $U^{(\varepsilon)}_{n,r_n,\omega}$, the results are much better and comparable to the case $p=2$. This demonstrates that the effect of the higher dimension can be attributed almost exclusively to the approximation error.

For all the results where $k=5$ and $p=2$, we used $\tilde{k}=4$. That means, $\tilde{k}=4$ was used to construct the estimated covariance matrix described in Section 6.1.1 and to construct the test grid from (6.3) on which we evaluated the suprema. Remark 5.13 justifies that $\tilde{k}<k$ can be sufficient for the simulation of the Gaussian process and the calculation of the uniform error if the covariance for any x_1, x_2 with $\mathfrak{C}_k(x_1,x_2) = (\tilde{k},\tilde{k})$, see (5.32), is equal or reasonably close to one for both RF types. For the uniform CPRF the covariance

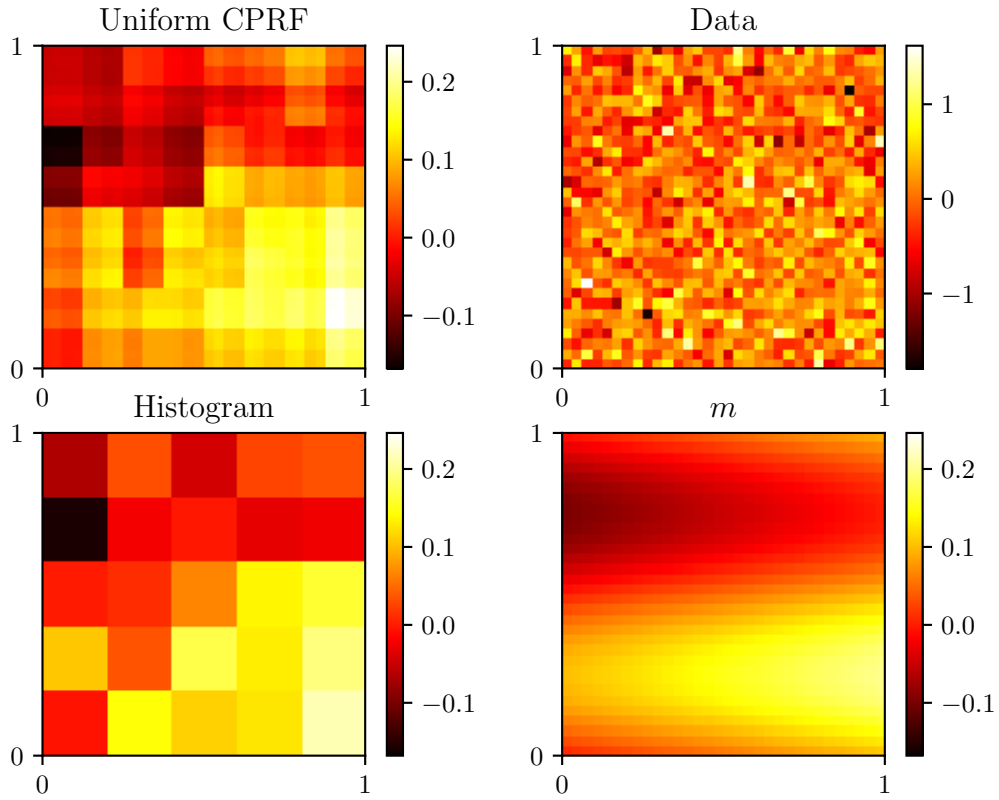

Figure 6.3: Data, true values and estimators of m from (6.4).

Table 6.4: Empirical coverage (left) and average confidence band radius (right) of 1000 confidence bands with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $n = 2000$, $k = 5$, $N = 50$ and $r_n = \frac{3}{4}n$.

σ	RF	$p = 2$		$p = 4$	
		$1 - \beta$			
		.90, .95, .99 Coverage	.90, .95, .99 Radius	.90, .95, .99 Coverage	.90, .95, .99 Radius
0.5	Uni.	.416, .616, .849	.152, .162, .182	.002, .006, .041	.124, .131, .146
	Ehr.	.416, .598, .868	.157, .168, .188	.001, .004, .046	.126, .133, .149
0.75	Uni.	.656, .782, .941	.228, .243, .272	.065, .116, .359	.186, .197, .219
	Ehr.	.648, .809, .945	.236, .252, .283	.056, .132, .382	.189, .200, .223
1	Uni.	.734, .839, .961	.304, .324, .363	.188, .321, .607	.248, .262, .291
	Ehr.	.745, .866, .970	.314, .336, .377	.202, .345, .635	.252, .267, .297
1.25	Uni.	.777, .871, .969	.380, .404, .454	.320, .469, .747	.310, .328, .364
	Ehr.	.790, .888, .973	.393, .419, .471	.338, .502, .742	.315, .334, .371

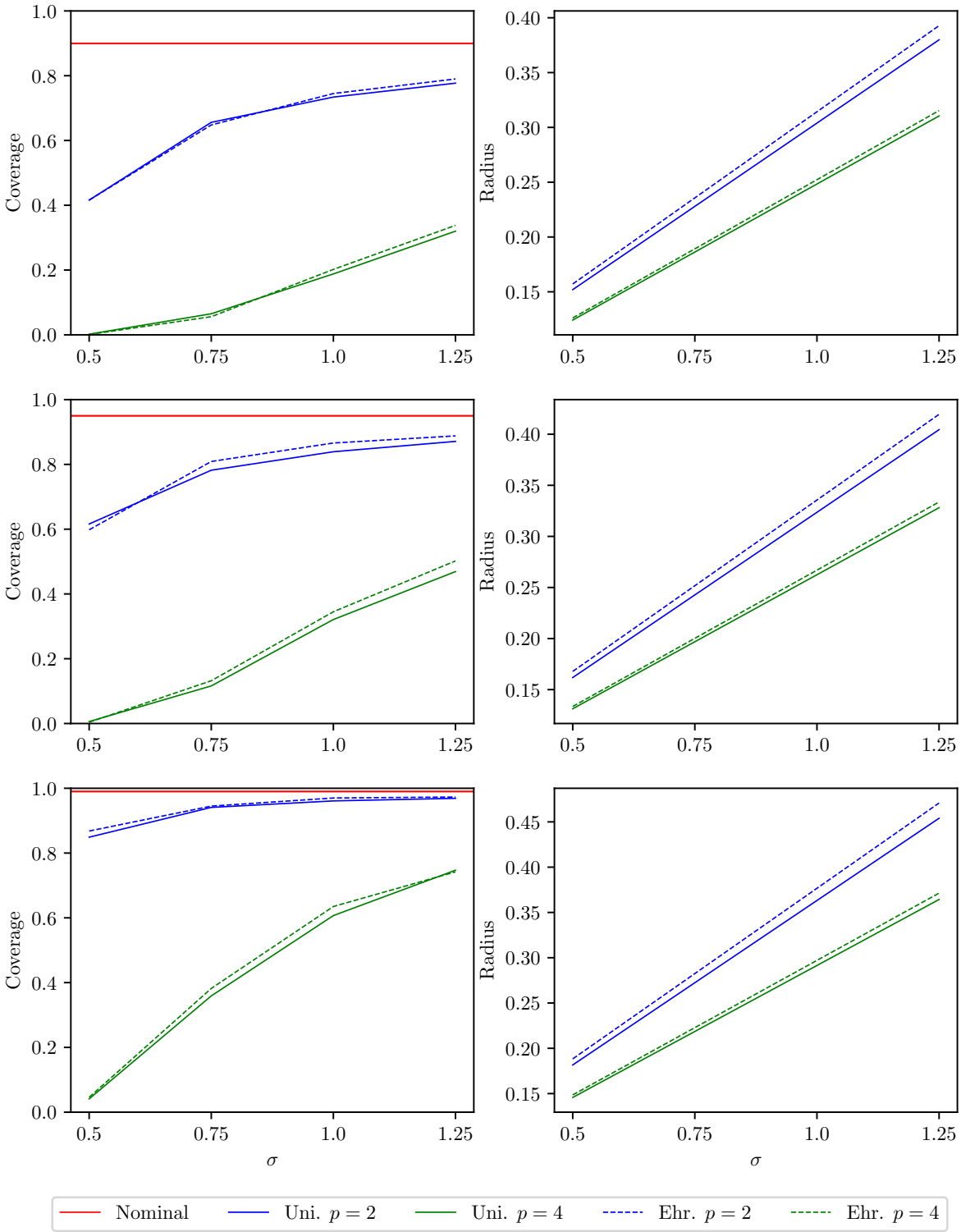

Figure 6.4: Comparison of CB coverage and radii for $1-\beta \in \{0.9, 0.95, 0.99\}$ and $p = 2, 4$ depending on σ.

for any x_1, x_2 with $\mathfrak{C}_k(x_1, x_2) = (4, 4)$ it is equal to 0.9965 and for the Ehrenfest CPRF it is equal to 0.9994. For all the results in the case $k = 6$ and $p = 2$ we used $\tilde{k} = 5$, which was sufficient for the same reason. In the case $k = 5$ and $p = 4$ we used $\tilde{k} = 3$.

6.2 Bootstrap confidence bands

In practice, especially when confronted with relatively small sample sizes, it may be a better option to use bootstrap confidence bands. Even though they are not backed up by our theory we include them in the simulation study. In a bootstrap setting we might be able to include the bias to get closer to the quantiles of the uniform difference of the estimator and the regression function. Suppose we have an estimator for the regression function, possibly for a different, usually smaller, k. Denote this estimator by $\hat{m}$ and the corresponding residuals by

$$\hat{\varepsilon}_i := Y_i - \hat{m}(X_i).$$

One option would be to use the residual bootstrap, where one simulates multiple bootstrap random forest estimators

$$U^{(\mathrm{BRF})}_{n,r_n,\omega}(x_0) := \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \sum_{j \in I} \tilde{Y}_j \frac{\mathbb{I}\{\tilde{X}_j \in A_k(x_0, \omega_I)\}}{\sum_{i \in I} \mathbb{I}\{\tilde{X}_i \in A_k(x_0, \omega_I)\}}$$

based on $\tilde{X}_j$ and

$$\tilde{Y}_j = \hat{m}(\tilde{X}_j) + \tilde{\varepsilon}_j.$$

Here, $\tilde{X}_j \sim U[0,1]^p$ are simulated, the $\tilde{\varepsilon}_j$ are drawn from the residuals with replacement. By simulating enough of these bootstrap random forests we can get an estimator for the quantiles. Like the regular estimator we can decompose the bootstrap estimator and get

$$\begin{aligned}
U^{(\mathrm{BRF})}_{n,r_n,\omega}(x_0) - \hat{m}(x_0) &= \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \sum_{j \in I} \hat{m}(\tilde{X}_j) \frac{\mathbb{I}\{\tilde{X}_j \in A_k(x_0, \omega_I)\}}{\sum_{i \in I} \mathbb{I}\{\tilde{X}_i \in A_k(x_0, \omega_I)\}} - \hat{m}(x_0) \\
&\quad + \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \sum_{j \in I} \tilde{\varepsilon}_j \frac{\mathbb{I}\{\tilde{X}_j \in A_k(x_0, \omega_I)\}}{\sum_{i \in I} \mathbb{I}\{\tilde{X}_i \in A_k(x_0, \omega_I)\}} \\
&\approx \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \sum_{j \in I} \big(\hat{m}(\tilde{X}_j) - \hat{m}(x_0)\big) \frac{\mathbb{I}\{\tilde{X}_j \in A_k(x_0, \omega_I)\}}{\sum_{i \in I} \mathbb{I}\{\tilde{X}_i \in A_k(x_0, \omega_I)\}} \\
&\quad + \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \sum_{j \in I} \tilde{\varepsilon}_j \frac{\mathbb{I}\{\tilde{X}_j \in A_k(x_0, \omega_I)\}}{\sum_{i \in I} \mathbb{I}\{\tilde{X}_i \in A_k(x_0, \omega_I)\}}.
\end{aligned}$$

The first term of this is not exactly zero, but both are piecewise constant and their difference can be rather small. If $\tilde{X}_j$ is in the undividable cell of x_0, it is equal to zero, because $\hat{m}$ is constant on this set. If we only have $\tilde{X}_j \in A_k(x_0, \omega_I)$, the term $\hat{m}(\tilde{X}_j) - \hat{m}(x_0)$ can still be very small. Hence, the residual bootstrap might not help to include the bias term. Mainly because the original bias is partly induced by approximating a smooth function by a piece-wise constant function. This is not the case for the bootstrap. Following this argumentation, it might not be helpful to include $\hat{m}$ in the bootstrap.

Table 6.5: Empirical coverage and average confidence band radius of 1000 bootstrap confidence bands with $k = 5$, $n = 250$, $N = 50$ and $r_n = 187$ for different error distributions.

RF		Normal	Uniform	t-dist. 4 df	t-dist. 6 df
			$1 - \beta$		
		.90, .95, .99	.90, .95, .99	.90, .95, .99	.90, .95, .99
Uni.	Cov.	.932, .976, .996	.948, .981, .998	.933, .973, .996	.928, .966, .994
	Rad.	1.09, 1.21, 1.46	1.00, 1.10, 1.28	1.29, 1.49, 1.93	1.19, 1.36, 1.70
Ehr.	Cov.	.939, .976, .993	.941, .980, .998	.929, .971, .996	.932, .966, .990
	Rad.	1.12, 1.24, 1.50	1.04, 1.14, 1.34	1.32, 1.52, 1.96	1.22, 1.38, 1.75

Instead, we can focus on the term $U^{(\varepsilon)}_{n,r_n,\omega}$ and approximate its finite sample distribution by a bootstrap. We use a multiplier bootstrap with

$$U^{(\varepsilon,B)}_{n,r_n,\omega}(x_0) := \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \sum_{j \in I} \tilde{\varepsilon}_j v_j \frac{\mathbb{I}\{\tilde{X}_j \in A_k(x_0, \omega_I)\}}{\sum_{i \in I} \mathbb{I}\{\tilde{X}_i \in A_k(x_0, \omega_I)\}}.$$

where the v_j are i.i.d. $\mathcal{N}(0,1)$. The bootstrap can be helpful if n is not large enough for the asymptotics to apply, or if the distribution of ε_1 is not normal, such that the Gaussian approximation has a larger error. However, it is not helpful to deal with a large approximation error.

We test the bootstrap in the case $n = 250$, $k = 5$, $N = 50$, $\varepsilon \sim \mathcal{N}(0,1)$ and

$$m(x_1, x_2) = \frac{1}{10}(\sin(2\pi x_1) + x_2)$$

from Table 6.2. We consider the same selection of different error distributions as in Table 6.1, but we do not consider the same sample sizes. In Table 6.5, we observe that the empirical coverage exceeds the theoretical coverage in all cases, which goes hand in hand with larger radii throughout. The larger radii suggest that the sample size is not large enough for the distribution of $\mathbf{S}_k$ to approximate the distribution of $\|U^{(\varepsilon)}_{n,r_n,\omega}\|_\infty$ well. As these distributions differ primarily due to the projection error and the Gaussian approximation, both errors can account for the total deviation. It is not clear which effect is the dominant one.

Nevertheless, the results indicate that the bootstrap confidence bands are robust with respect to different error distributions, even when the sample size is relatively small. The confidence band radii for the t-distributions are larger than the radii for the normal distribution, which meets the expectation. For the uniform distribution, the radii are smaller, which can be attributed to the bounded support of the distribution. As expected, and as in the asymptotic case, the confidence bands are wider for the Ehrenfest CPRF than for the uniform CPRF throughout the error distributions.

6.3 Limitations

The simulation study has a few inherent limitations due to its specific design. Rather than employing the (incomplete) U-statistic version of the random forest, we use a fixed

number of trees, resulting in a different estimator than that utilized in the theory. For this estimator, there is a positive probability that the same subsample is used in two trees. However, this probability is negligible for the considered n and N. The U-statistic structure of the theoretical estimator prevents a subsample from being used twice. Another limitation is the small selection of two different regression functions for two distinct values of p. A wider selection of regression functions for more values of p would provide a better overview of the general performance of the proposed methods.

Since the error can only be evaluated on a finite grid, it will not be captured to its full extend. Nevertheless, the choice of the grid together with continuity of the regression function should ensure that this difference is negligible. Further, the number of simulated confidence bands is of course finite, and a larger number would always lead to a better picture of performance of the methods. The estimation of σ may influence the results independently of the random forest estimator of m. However, in our simulations the estimator performs quite well.

Overall, the results support the theory, but also posses some limitations. We used a rather small signal to noise ratio and observed a coverage clearly different from the theoretical coverage for the smallest values of σ, which is even more pronounced for $p = 4$, see Table 6.4. This indicates that the approximation error is too large which is a major drawback of these random forest variants and a consequence of their inflexible centered structure. Still, our proposed random forest methods outperform the histogram estimator, for which this effect is even more pronounced. The results for $n = 1000$ in Table 6.2 suggest that a k greater than five does not solve the problem. It leads to a smaller approximation error, but the coverage is worse. Nevertheless, the results are valuable, because they support the theory regarding the asymptotic distribution and the goal of the theoretical results for the CPRF is to extend them to more sophisticated random forest methods. If this can be done, these different random forests models can avoid the approximation error drawback due to their different partitioning algorithms.

Chapter 7

Discussion

In this final chapter, we discuss several different aspects that provide different perspectives on the main contributions of the previous chapters. In Section 7.1, we outline the extension of the confidence band results to honest centered random forests, a version that utilizes the data for the partition construction. Subsequently, in Section 7.2, we describe the previously mentioned application of the classical proof structure for confidence bands to CPRFs. We explain the differences from the proof structure used in Chapter 5, and why the limitations of that proof and its result lead us to look for a different proof technique. Finally, in 7.3, we suggest several more possible areas for future research.

7.1 An extension - centered honest random forests

In this section, we discuss how the results from Chapter 5 can be extended to a random forest type that uses data dependent partitions. We consider a centered random forest that is honest according to the definition of the term honesty introduced by Wager and Athey (2018). Their definition of honesty is that in each tree, every observation is used either to create the partition or to estimate m given the partition, but not both. This allows us to use a data dependent partition and keep the independence between partition creation and estimation.

Let $(\tilde{Z}_i)_{i=1}^{\tilde{n}}$ be a sample of i.i.d. copies of (X, Y) that is independent of $(Z_i)_{i=1}^{n}$. This sample will be used for the construction of the tree partition. The original sample will be used for the estimation in this partition. In practice, one would do this by splitting the sample. Intuitively, one would split the sample for each tree. The theory for this is more complex due to the dependencies between the trees. Therefore, we use the same split of the training sample for each tree to avoid technical difficulties. It may still be possible to generalize the results to this version of honest random forests. Honesty is useful because the random variables used for partitioning and those used for the estimation are independent. This makes the mathematical analysis much simpler. We retain the assumption that the trees are built with centered splits.

Similar to honest trees, one can consider a model where the probabilities to split the features depend on the importance of the features, but are fixed in advance. For example, these probabilities can be estimated on a different training sample before actually estimating of the regression function. This means that instead of using the data to choose

the partitions, it is reduced to an assumption of the random forest model. Models with such an assumption have been treated in the literature, for example by Breiman (2004), Biau (2012) and Klusowski (2021).

Let us define

$$U^{(\mathrm{HRF})}_{n,r_n,\omega}(x_0) := \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \sum_{j \in I} Y_j W^H_{j,k}(x_0, \omega_I, I)$$

with

$$W^H_{j,k}(x_0, \omega, I) := \frac{\mathbb{I}\{X_j \in A_k(x_0, \omega, (\tilde{Z}_l)_{l=1}^{\tilde{n}})\}}{\sum_{i \in I} \mathbb{I}\{X_i \in A_k(x_0, \omega, (\tilde{Z}_l)_{l=1}^{\tilde{n}})\}},$$

and $W^H_{j,k}(x_0, I) := W^H_{j,k}(x_0, \omega_I, I)$. Similar to the purely random forest we define

$$U^{(H,\varepsilon)}_{n,r_n,\omega}(x_0) := \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \sum_{j \in I} \varepsilon_j W^H_{j,k}(x_0, I)$$

and

$$U^{(H,m)}_{n,r_n,\omega}(x_0) := \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \sum_{j \in I} m(X_j) W^H_{j,k}(x_0, I).$$

In this section, for simplicity, we will consider the special case where X is uniformly distributed. The case where $0 < c_X \leq f_X \leq C_X$ works similar to Chapter 5. Let $(\check{Z}_l)_{l=1}^{\tilde{n}}$ be an independent copy of the subsample $(\tilde{Z}_l)_{l=1}^{\tilde{n}}$. We define

$$\tilde{\mathcal{V}}_{\cap,k}(x_0) := \mathbb{E}\left[\mathbb{V}(A_k(x_0, \omega_1, (\tilde{Z}_l)_{l=1}^{\tilde{n}}) \cap A_k(x_0, \omega_2, (\check{Z}_l)_{l=1}^{\tilde{n}}))\right].$$

This depends on x_0 because the data dependence of the partition can lead to heterogeneous behavior of the intersections on the feature space. Further we denote $\bar{\mathcal{V}}_{\cap,k} := \inf_{x_0 \in [0,1]^p} \tilde{\mathcal{V}}_{\cap,k}(x_0)$ and define

$$\tilde{f}_{x_0,k}(x, s) := \sigma^{-1} \tilde{\mathcal{V}}^{-1/2}_{\cap,k}(x_0) s \mathbb{P}(x \in A_k(x_0, \omega, (\tilde{Z}_l)_{l=1}^{\tilde{n}})), \tag{7.1}$$

$$\tilde{\mathcal{F}}_k := \{\tilde{f}_{x_0,k} \mid x_0 \in [0,1]^p\}. \tag{7.2}$$

Conjecture 7.1. *Consider an honest centered random forest with at most $N_f(k)$ undividable sets (see Section 3.3.1.3). Let*

$$\mathcal{C}_{k,\tilde{n}}(x_0) := \mathbb{E}\big[\mathbb{V}(A_k(x_0, \omega_1, (\tilde{Z}_l)_{l=1}^{\tilde{n}}) \cap A_k(x_0, \omega_2, (\tilde{Z}_l)_{l=1}^{\tilde{n}}))\big] - \tilde{\mathcal{V}}_{\cap,k}(x_0) \qquad \textit{and} \tag{7.3}$$

$$U^{(H,1)}_{n,r_n,\omega}(x_0) := \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \sum_{j \in I} W^H_{j,k}(x_0, I). \tag{7.4}$$

In addition to several technical assumptions on n, r_n and k assume that $\|\mathcal{C}_{k,\tilde{n}}\|_\infty$ and $\|U^{(H,m)}_{n,r_n,\omega} - mU^{(H,1)}_{n,r_n,\omega}\|_\infty$ converge to zero (in probability) sufficiently fast. For a Gaussian process B_k with covariance

$$\begin{aligned}&\mathrm{Cov}(B_k \tilde{f}_{x_1,k}, B_k \tilde{f}_{x_2,k}) \\ &\quad = \tilde{\mathcal{V}}^{-1/2}_{\cap,k}(x_1) \tilde{\mathcal{V}}^{-1/2}_{\cap,k}(x_2) \mathbb{P}(X_1 \in A_k(x_1, \omega_1, (\tilde{Z}_l)_{l=1}^{\tilde{n}}) \cap A_k(x_2, \omega_2, (\check{Z}_l)_{l=1}^{\tilde{n}}))\end{aligned} \tag{7.5}$$

let $\mathbf{S}_k \overset{d}{=} \sup_{x_0 \in [0,1]^p} |B_k \tilde{f}_{x_0,k}|$ *and for* $c_k(\beta) = F^{-1}_{\mathbf{S}_k}(1-\beta)$ *denote*

$$\mathcal{C}_n(x) = \left[U^{(HRF)}_{n,r_n,\omega}(x) - \sigma c_k(\beta) \sqrt{\frac{2^{2k} \tilde{\mathcal{V}}_{\cap,k}(x)}{n}}, U^{(HRF)}_{n,r_n,\omega}(x) + \sigma c_k(\beta) \sqrt{\frac{2^{2k} \tilde{\mathcal{V}}_{\cap,k}(x)}{n}} \right].$$

For $\alpha \in (0,1]$ *and* $C_H > 0$ *it holds that*

$$\liminf_{n \to \infty} \inf_{m \in \mathcal{H}(\alpha, C_H)} \mathbb{P}\big(m(x) \in \mathcal{C}_n(x),\, \forall x \in [0,1]^p\big) \geq 1 - \beta.$$

The radius of the confidence bands in the result above depends on the function $\tilde{\mathcal{V}}_{\cap,k}$, which is similar to the effect of Ψ_k in Theorem 5.1. In the honest case, the varying radius is due to the data dependence. If the absolute gradient of m is relatively small in a region, we would expect the variance of the estimator to be relatively smaller. If we would extend the above result to the case where f_X is not constant, the radius of the confidence band would depend on the density of X and on the regression function m.

If the dependence of the partitions on the data is suitable, these confidence bands might be able to work in a high-dimensional regression model with an active dimension less than the full dimension of the X. If the cells are never or rarely split orthogonal to the inactive features, the estimator can use most of the splits to generate a useful partition in the active dimensions. This can lead to a sufficiently small approximation error in higher dimensions.

A limitation of this type of random forest is that the splits are still centered and thus the cell size is still constant. This might not be a good property for a data dependent partition, as it limits the ability of the estimator to use the data to generate a good partition.

The term $\tilde{\mathcal{V}}_{\cap,k}(x_0)$ can still be approximated through a Monte Carlo simulation, but it is more complex because we need the $\tilde{Z}_l$ for the partition. Even if the distribution of X is known the partitions also depend on the Y. One can use bootstrap samples of the entire training sample to estimate $\tilde{\mathcal{V}}_{\cap,k}(x_0)$.

In Conjecture 7.1 we need that

$$\begin{aligned} \|\mathcal{C}_{k,\tilde{n}}\|_\infty = \sup_{x_0 \in [0,1]^p} \Big| & \mathbb{E}\big[\mathbb{V}(A_k(x_0, \omega_1, (\tilde{Z}_l)_{l=1}^{\tilde{n}}) \cap A_k(x_0, \omega_2, (\tilde{Z}_l)_{l=1}^{\tilde{n}}))\big] \\ & - \mathbb{E}\big[\mathbb{V}(A_k(x_0, \omega_1, (\tilde{Z}_l)_{l=1}^{\tilde{n}}) \cap A_k(x_0, \omega_2, (\check{Z}_l)_{l=1}^{\tilde{n}}))\big] \Big| \end{aligned}$$

converges to zero fast enough. If $\tilde{n}$ is large enough, $(\tilde{Z}_l)_{l=1}^{\tilde{n}}$ should reflect almost all of the information from the model that is useful for building the partition. If the dependence of the partition construction on the data is well behaved, we would expect $A_k(x_0, \omega, (\tilde{Z}_l)_{l=1}^{\tilde{n}})$ to converge in a useful sense. For instance,

$$\|A_k(x_0, \cdot, (\tilde{Z}_l)_{l=1}^{\tilde{n}}) - A_k(x_0, \cdot, \mathcal{Z})\|_\infty \xrightarrow{\mathbb{P}} 0$$

for $\tilde{n} \to \infty$, where $\mathcal{Z}$ reflects all the information in the model that is useful for partitioning. In this case, the convergence assumption on $\|\mathcal{C}_{k,\tilde{n}}\|_\infty$ is an assumption that $\tilde{n}$ is sufficiently large. Later this assumption will be captured in (7.6).

Similar to the purely random case one would need a sufficiently sharp lower bound for $\tilde{\mathcal{V}}_{\cap,k}$. We note that this needs to be uniform now. It still holds that $2^{-2k} \leq \tilde{\mathcal{V}}_{\cap,k}(x_0) \leq 2^{-k}$. Given a specific data dependent tree construction one can improve this lower bound for $\tilde{\mathcal{V}}_{\cap,k}(x_0)$. From (3.21) we know that

$$\tilde{\mathcal{V}}_{\cap,k}(x_0)2^k = \mathbb{E}\left[2^{-\frac{1}{2}\sum_{l=1}^{p}|S_l(x_0,\omega_1,(\tilde{Z}_l)_{l=1}^{\tilde{n}})-S_l(x_0,\omega_2,(\check{Z}_l)_{l=1}^{\tilde{n}})|}\right].$$

The number of splits behaves differently for all features in an honest regression tree. For a fixed feature, the difference between the number of splits depending on two independent copies of $\omega, (\tilde{Z}_l)_{l=1}^{\tilde{n}}$ should still behave similarly. Hence, we should get a comparable lower bound if $\tilde{n}$ is large enough.

Conjecture 7.1 is a first step towards confidence bands for random forests with data dependent partitions. The centeredness of the splits is a notable limitation. Especially for data dependent partitions, a selective split placement is desirable to create a useful partition given the data.

7.1.1 Proof strategy

We will discuss how to generalize the proof strategy from Chapter 5 to this type of random forest. The goal of data dependent trees is to create partitions that allow for a better approximation of the regression function. Thus, we expect the data dependence to lead to a smaller approximation error. The supposed way to decrease the approximation error is to exploit the structure of the regression function.

The construction of the confidence bands is based on the stochastic error, and the main part of the proof deals with the stochastic error. We do not want to make explicit assumptions about the data dependent construction of the partitions. Therefore, we will only briefly discuss the approximation error. We use the decomposition of the approximation error in (5.45). Since we are still considering centered random forests with depth k, the same bound from Lemma 5.17 holds for

$$m(x_0)(U_{n,r_n,\omega}^{(H,1)}(x_0) - 1).$$

If the regression function has bounded first order derivatives we can use the following bound for the approximation error.

$$|m(x_1) - m(x_2)| \leq \sum_{l=1}^{p} \|\partial_l m\|_\infty |x_1^{(l)} - x_2^{(l)}|$$

Let $a_{l,k}(x_0, \omega, (\tilde{Z}_l)_{l=1}^{\tilde{n}})$ denote the edge lengths of the cell in direction l. We get for the remaining part of the approximation error that

$$\begin{aligned}
&|U_{n,r_n,\omega}^{(H,m)}(x_0) - m(x_0)U_{n,r_n,\omega}^{(H,1)}(x_0)| \\
&\qquad \leq \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \sum_{j \in I} |m(X_j) - m(x_0)| W_{j,k}^H(x_0, I) \\
&\qquad \leq \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \sum_{j \in I} \sum_{l=1}^{p} \|\partial_l m\|_\infty |X_j^{(l)} - x_0^{(l)}| W_{j,k}^H(x_0, I)
\end{aligned}$$

$$\leq \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \sum_{l=1}^{p} \|\partial_l m\|_\infty a_{l,k}(x_0, \omega_I, (\tilde{Z}_l)_{l=1}^{\tilde{n}}) \sum_{j \in I} W_{j,k}^H(x_0, I)$$
$$\leq \frac{1}{\binom{n}{r_n}} \sum_{I \in B_{r_n,n}} \sum_{l=1}^{p} \|\partial_l m\|_\infty 2^{-S_l(x_0,\omega_I,(\tilde{Z}_l)_{l=1}^{\tilde{n}})}.$$

The goal of the splits in the data dependent partition is to minimize the above term. If we have features that carry less information about the regression function they should be split less often. If we have $\|\partial_l m\|_\infty = 0$ for some l, it suffices to have $S_l(x_0, \omega_I, (\tilde{Z}_l)_{l=1}^{\tilde{n}}) = 0$. In general, we assume that the data dependence exploits the structure of the aforementioned upper bound. It should be noted that this can also be achieved by making assumptions about the probabilities of splitting the features, as was mentioned at the beginning of this section.

From here on we will explain the proof strategy for the stochastic error. An important part of the proof is the approximation of the supremum of a suitable empirical process. Analogously to Chapter 5, let us define

$$\hat{U}_{n,r_n,\omega}^{(H,\varepsilon)}(x_0) = \frac{2^k}{n} \sum_{j=1}^{n} \varepsilon_j \mathbb{P}(X_j \in A_k(x_0, \omega, (\tilde{Z}_l)_{l=1}^{\tilde{n}}) \mid X_j).$$

We only condition the probability on X_j to get an empirical process in a function class, where the functions have arguments X and ε. We observe that

$$\begin{aligned}
\operatorname{Var}\big(\hat{U}_{n,r_n,\omega}^{(H,\varepsilon)}(x_0)\big) &= \frac{2^{2k}}{n}\sigma^2 \mathbb{E}\big[\mathbb{P}(X_1 \in A_k(x_0, \omega, (\tilde{Z}_l)_{l=1}^{\tilde{n}}) \mid X_1)^2\big] \\
&= \frac{2^{2k}}{n}\sigma^2 \mathbb{P}(X_1 \in A_k(x_0, \omega_1, (\tilde{Z}_l)_{l=1}^{\tilde{n}}) \cap A_k(x_0, \omega_2, (\check{Z}_l)_{l=1}^{\tilde{n}})) \\
&= \frac{2^{2k}}{n}\sigma^2 \mathbb{E}\big[\mathbb{V}(A_k(x_0, \omega_1, (\tilde{Z}_l)_{l=1}^{\tilde{n}}) \cap A_k(x_0, \omega_2, (\check{Z}_l)_{l=1}^{\tilde{n}}))\big] \\
&= \frac{2^{2k}}{n}\sigma^2 \tilde{\mathcal{V}}_{\cap,k}(x_0).
\end{aligned}$$

Recall that $\tilde{\mathcal{F}}_k$ from (7.2) is the function class of the $\tilde{f}_{x_0,k}$. This implies

$$\begin{aligned}
\sup_{f \in \tilde{\mathcal{F}}_k \cup -\tilde{\mathcal{F}}_k} \mathbb{G}_n f &= \sup_{f \in \tilde{\mathcal{F}}_k} |\mathbb{G}_n f| \\
&= \sup_{x_0 \in [0,1]^p} \big|\mathbb{G}_n \sigma^{-1} \tilde{\mathcal{V}}_{\cap,k}^{-1/2}(x_0) s \mathbb{P}(x \in A_k(x_0, \omega, (\tilde{Z}_l)_{l=1}^{\tilde{n}}))\big| \\
&= \sup_{x_0 \in [0,1]^p} \Big|\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \sigma^{-1} \tilde{\mathcal{V}}_{\cap,k}^{-1/2}(x_0) \varepsilon_j \mathbb{P}(X_j \in A_k(x_0, \omega, (\tilde{Z}_l)_{l=1}^{\tilde{n}}) \mid X_j)\Big| \\
&= \sqrt{\frac{n}{\sigma^2 2^{2k}}} \sup_{x_0 \in [0,1]^p} \big|\tilde{\mathcal{V}}_{\cap,k}^{-1/2}(x_0) \hat{U}_{n,r_n,\omega}^{(H,\varepsilon)}(x_0)\big|.
\end{aligned}$$

Using $\bar{\mathcal{V}}_{\cap,k}$ to lower bound $\bar{\mathcal{V}}_{\cap,k}(x_0)$ we get analogously to Theorem 5.14 that

$$\left|\sqrt{\frac{n}{\sigma^2 2^{2k}}} \sup_{x_0 \in [0,1]^p} \big|\tilde{\mathcal{V}}_{\cap,k}^{-1/2}(x_0) \hat{U}_{n,r_n,\omega}^{(H,\varepsilon)}(x_0)\big| - \mathbf{S}_k\right|$$

$$= \mathcal{O}_{\mathbb{P}}\left(\frac{(\log n)^{3/2}}{\bar{\mathcal{V}}_{\cap,k}^{1/2} n^{1/2-1/\nu}} + \frac{(\log n)^{5/4}}{\bar{\mathcal{V}}_{\cap,k}^{1/4} n^{1/4}} + \frac{\log n}{\bar{\mathcal{V}}_{\cap,k}^{1/6} n^{1/6}}\right)$$

where $\mathbf{S}_k \stackrel{d}{=} \sup_{x_0\in[0,1]^p} |B_k \tilde{f}_{x_0,k}|$ and B_k is a Gaussian process with covariance as in (7.5).

In comparison to Chapter 5 there is a difference in the proof for the remainder terms. Analogous to (5.46) we have

$$\begin{aligned}
&U_{n,r_n,\omega}^{(H,\varepsilon)}(x_0) - \hat{U}_{n,r_n,\omega}^{(H,\varepsilon)}(x_0) \\
&\quad= \frac{1}{\binom{n}{r_n}} \sum_{I\in B_{r_n,n}} \sum_{j\in I} \varepsilon_j \big(W_{j,k}^H(x_0,\omega_I,I) - \frac{2^k}{r_n}\mathbb{I}\{X_j \in A_k(x_0,\omega_I,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\}\big) \\
&\qquad+ \frac{2^k}{n}\sum_{j=1}^{n} \varepsilon_j \frac{1}{\binom{n-1}{r_n-1}} \sum_{I\in B_{r_n,n}: j\in I} \\
&\qquad\qquad \times \big(\mathbb{I}\{X_j \in A_k(x_0,\omega_I,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\} - \mathbb{P}(X_j \in A_k(x_0,\omega,(\tilde{Z}_l)_{l=1}^{\tilde{n}}) \mid X_j)\big) \\
&\quad=: R_{n,r_n,\omega}^{(H,1)}(x_0) + R_{n,r_n,\omega}^{(H,2)}(x_0).
\end{aligned}$$

We consider the first remainder term conditioned on $(\tilde{Z}_l)_{l=1}^{\tilde{n}}$. The remaining randomization of the cells are the ω_I which are still i.i.d. and hence we can apply Lemma 5.18 to $R_{n,r_n,\omega}^{(H,1)}(x_0)$ conditioned on $(\tilde{Z}_l)_{l=1}^{\tilde{n}}$. This yields

$$\mathbb{E}\left[R_{n,r_n,\omega}^{(H,1)}(x_0)^q\right] = \mathbb{E}\left[\mathbb{E}\left[R_{n,r_n,\omega}^{(H,1)}(x_0)^q \mid (\tilde{Z}_l)_{l=1}^{\tilde{n}}\right]\right] \lesssim \left(\frac{2^{2k}}{r_n n}\right)^{q/2}.$$

We need to change how we handle the second remainder term, instead of Lemma 5.19 we need the lemma below.

Lemma 7.2. *For $\mathcal{C}_{k,\tilde{n}}(x_0)$ defined in (7.3) it holds that*

$$\mathrm{Var}(R_{n,r_n,\omega}^{(H,2)}(x_0)) \le \frac{2^k\sigma^2}{r_n}\left(\frac{r_n}{n}\right)^{r_n} + \frac{2^{2k}\sigma^2}{n}\mathcal{C}_{k,\tilde{n}}(x_0).$$

To proceed analogously to the proof of Proposition 5.15 we need that

$$\begin{aligned}
&\mathbb{P}\left(\|\tilde{\mathcal{V}}_{\cap,k}^{-1/2} R_{n,r_n,\omega}^{(H,2)}\|_\infty \ge \kappa(\log n)^{-1}\sqrt{2^{2k}/n}\right) \\
&\le \sum_{x_0\in\mathcal{X}_k} \mathbb{P}\left(|R_{n,r_n,\omega}^{(H,2)}(x_0)| \ge \kappa(\log n)^{-1}\sqrt{2^{2k}\bar{\mathcal{V}}_{\cap,k}/n}\right) \\
&\le \sum_{x_0\in\mathcal{X}_k} \mathrm{Var}(R_{n,r_n,\omega}^{(H,2)}(x_0)) \frac{n(\log n)^2}{\kappa^2 2^{2k}\bar{\mathcal{V}}_{\cap,k}} \\
&\le N_f(k)\frac{\sigma^2}{\kappa^2}\left(\frac{(\log n)^2}{2^k\bar{\mathcal{V}}_{\cap,k}}\left(\frac{r_n}{n}\right)^{r_n-1} + \frac{(\log n)^2}{\bar{\mathcal{V}}_{\cap,k}}\|\mathcal{C}_{k,\tilde{n}}\|_\infty\right) \qquad (7.6)
\end{aligned}$$

converges to zero. The first term is similar to Chapter 5. The second is the new condition already discussed with Conjecture 7.1. It remains to give the proof of Lemma 7.2.

Proof of Lemma 7.2. It holds that

$$
\begin{aligned}
&\mathrm{Var}(R^{(H,2)}_{n,r_n,\omega}(x_0)) \\
&\quad= \frac{2^{2k}}{n^2}\mathrm{Var}\Bigg(\sum_{j=1}^{n}\varepsilon_j \frac{1}{\binom{n-1}{r_n-1}} \sum_{I\in B_{r_n,n}:j\in I} \\
&\qquad\qquad \times\Big(\mathbb{I}\{X_j\in A_k(x_0,\omega_I,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\}-\mathbb{P}(X_j\in A_k(x_0,\omega,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\mid X_j)\Big)\Bigg) \\
&\quad= \frac{2^{2k}\sigma^2}{n}\mathrm{Var}\Bigg(\frac{1}{\binom{n-1}{r_n-1}} \sum_{I\in B_{r_n,n}:j\in I} \\
&\qquad\qquad \times\Big(\mathbb{I}\{X_j\in A_k(x_0,\omega_I,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\}-\mathbb{P}(X_j\in A_k(x_0,\omega,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\mid X_j)\Big)\Bigg) \\
&\quad= \frac{2^{2k}\sigma^2}{n}\frac{1}{\binom{n-1}{r_n-1}}\mathrm{Var}\big(\mathbb{I}\{X_1\in A_k(x_0,\omega_I,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\}-\mathbb{P}(X_1\in A_k(x_0,\omega,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\mid X_1)\big) \\
&\qquad+\frac{2^{2k}\sigma^2}{n}\frac{1}{\binom{n-1}{r_n-1}^2}\sum_{I\in B_{r_n,n}:1\in I}\sum_{J\in B_{r_n,n}:1\in J,J\neq I} \\
&\qquad\qquad\times\mathrm{Cov}\big(\mathbb{I}\{X_1\in A_k(x_0,\omega_I,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\}-\mathbb{P}(X_1\in A_k(x_0,\omega,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\mid X_1), \\
&\qquad\qquad\qquad\mathbb{I}\{X_1\in A_k(x_0,\omega_J,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\}-\mathbb{P}(X_1\in A_k(x_0,\omega,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\mid X_1)\big) \\
&\quad= \frac{2^{2k}\sigma^2}{n}\frac{1}{\binom{n-1}{r_n-1}}\Big(2^{-k}-\tilde{\mathcal{V}}_{A^2}(x_0)\Big)+\frac{2^{2k}\sigma^2}{\kappa^2 n}\mathcal{C}_{k,\tilde{n}}(x_0) \\
&\quad\leq \frac{2^k\sigma^2}{r_n}\left(\frac{r_n}{n}\right)^{r_n}+\frac{2^{2k}\sigma^2}{\kappa^2 n}\mathcal{C}_{k,\tilde{n}}(x_0)
\end{aligned}
$$

because we observe

$$
\begin{aligned}
&\mathrm{Cov}\big(\mathbb{I}\{X_1\in A_k(x_0,\omega_1,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\}-\mathbb{P}(X_1\in A_k(x_0,\omega,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\mid X_1), \\
&\qquad\mathbb{I}\{X_1\in A_k(x_0,\omega_2,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\}-\mathbb{P}(X_1\in A_k(x_0,\omega,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\mid X_1)\big) \\
&\quad= \mathbb{E}\big[(\mathbb{I}\{X_1\in A_k(x_0,\omega_1,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\}-\mathbb{P}(X_1\in A_k(x_0,\omega,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\mid X_1)) \\
&\qquad\times(\mathbb{I}\{X_1\in A_k(x_0,\omega_2,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\}-\mathbb{P}(X_1\in A_k(x_0,\omega,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\mid X_1))\big] \\
&\quad= \mathbb{E}\big[\mathbb{I}\{X_1\in A_k(x_0,\omega_1,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\}\mathbb{I}\{X_1\in A_k(x_0,\omega_2,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\}\big] \\
&\qquad-\mathbb{E}\big[\mathbb{I}\{X_1\in A_k(x_0,\omega_1,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\}\mathbb{I}\{X_1\in A_k(x_0,\omega_2,(\check{Z}_l)_{l=1}^{\tilde{n}})\}\big] \\
&\quad= \mathbb{E}\big[\mathbb{V}(A_k(x_0,\omega_1,(\tilde{Z}_l)_{l=1}^{\tilde{n}})\cap A_k(x_0,\omega_2,(\tilde{Z}_l)_{l=1}^{\tilde{n}}))\big]-\tilde{\mathcal{V}}_{\cap,k}(x_0) \\
&\quad= \mathcal{C}_{k,\tilde{n}}(x_0).
\end{aligned}
$$

□

7.2 The classical proof structure

In this section, we want to discuss a strategy used in the literature to prove confidence bands for different nonparametric regression estimators. Many proofs in the literature do not use a result like the one by Chernozhukov et al. (2014b), instead they use a different

proof structure where one approximates the empirical process instead of its supremum. With minor differences, this proof structure is used for several results, e.g., by Johnston (1982), Härdle (1989), Claeskens and Van Keilegom (2003), and Chao et al. (2017).

Following this proof technique for our method led to unsatisfactory results, which is why we set out to find an improved technique. To illustrate the differences and limitations, we will outline the classical proof structure and compare it to the proof structure used in Chapter 5. In this section, we assume that X is uniformly distributed on $[0,1]^p$. We define the classical empirical process

$$G_n(x,s) := \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \big(\mathbb{I}\{X_j \leq x, \varepsilon_j \leq s\} - \mathbb{P}(X_1 \leq x, \varepsilon_1 \leq s) \big).$$

Due to the uniform distribution of X Remark 4.3 implies

$$\begin{aligned} \hat{U}^{(\varepsilon)}_{n,r_n,\omega}(x_0) &= \frac{2^k}{n} \sum_{j=1}^{n} \varepsilon_j \mathbb{P}(X_j \in A_k(x_0,\omega) \mid X_j) \\ &= n^{-1/2} 2^k \int_{[0,1]^p \times \mathbb{R}} s \mathbb{P}(x \in A_k(x_0,\omega)) dG_n(x,s). \end{aligned}$$

By comparison, in the new proof we use the representation

$$\hat{U}^{(\varepsilon)}_{n,r_n,\omega}(x_0) = \sqrt{\frac{\sigma^2 2^{2k} \mathcal{V}_{\cap,k}}{n}} \mathbb{G}_n f_{x_0,k}.$$

Theorem 3.2 by Dedecker et al. (2014) yields that there exists a sequence of continuous centered Gaussian processes $\check{B}_n(x,s)$ that are uniformly close to $G_n(x,s)$. The idea is that if these processes are uniformly close to each other, then the integrals with respect to the processes should also be close to each other. An older result of this type, used for multiple confidence band results in the univariate regression model, is that by Tusnády (1977). This proof structure is less direct than using the result by Chernozhukov et al. (2014b). The main step is to show that the integrals are uniformly close. This is done using integration by parts, such that the difference of the integrals can be bounded by the integral over the difference of the processes. Results for multivariate integration by parts can be found in the articles by Ansari (2024) and Zaremba (1968). The proof contains several more technical steps and leads to the following result.

Proposition 7.3. *Let W be a p-variate Brownian sheet, i.e. a Gaussian process $W(x)$ for $x \in \mathbb{R}^p$. There exists a random process $V_n(x_0)$ with*

$$V_n(x_0) \stackrel{d}{=} \mathcal{V}_{\cap,k}^{-1/2} \sigma \int_{[0,1]^p} \mathbb{P}(x \in A_k(x_0,\omega)) dW(x)$$

that satisfies

$$\begin{aligned} &\sup_{x_0 \in [0,1]^p} \left| \sqrt{\frac{n}{2^{2k} \mathcal{V}_{\cap,k}}} \hat{U}^{(\varepsilon)}_{n,r_n,\omega}(x_0) - V_n(x_0) \right| \\ &\qquad = \mathcal{O}_{\mathbb{P}}\big(2^{-k} \mathcal{V}_{\cap,k}^{-1/2}\big) + o_{\mathbb{P}}\big(\mathcal{V}_{\cap,k}^{-1/2} n^{-u/6} (\log n)^{r(\kappa+(2p+6)/3)}\big) \end{aligned} \tag{7.7}$$

for some $u < 1$ and every $\kappa > 0$.

The proposition provides a bound for the appropriately standardized error of the Gaussian approximation, which in this case is done uniformly for the entire process. We want to compare this bound with the bound from Theorem 5.14, which is for the Gaussian approximation of the supremum instead of the entire process. The bound from Theorem 5.14 in the case where X is uniformly distributed is

$$\left|\sqrt{\frac{n}{2^{2k}\mathcal{V}_{\cap,k}}}\sup_{x_0\in[0,1]^p}|\hat{U}^{(\varepsilon)}_{n,r_n,\omega}(x_0)| - \sigma\mathbf{S}_k\right| = \mathcal{O}_{\mathbb{P}}\left(\frac{(\log n)^{3/2}}{\mathcal{V}_{\cap,k}^{1/2}n^{1/2-1/\nu}} + \frac{(\log n)^{5/4}}{\mathcal{V}_{\cap,k}^{1/4}n^{1/4}} + \frac{\log n}{\mathcal{V}_{\cap,k}^{1/6}n^{1/6}}\right). \tag{7.8}$$

It is evident that the first bound is asymptotically larger due to the second term on the right-hand side and the fact that $u < 1$. For convergence to zero, we need that $\mathcal{V}_{\cap,k}n^{u/3} \to \infty$ with the first bound from (7.7), but only $\mathcal{V}_{\cap,k}n^{1-2/\nu} \to \infty$ with the bound in (7.8). Further, it is noteworthy that the assumption $\mathcal{V}_{\cap,k}n^{u/3} \to \infty$ contradicts the undersmoothing assumption (5.4) and thus, would prevent the construction of confidence bands unless one were to use a bias correction.

We briefly discuss the reason for the difference between the bounds. It is important to point out that the approximation result by Chernozhukov et al. (2014b) only approximates one random variable. In comparison, the classical result relies on a uniform approximation of an empirical process. This problem is more difficult than approximating a single random variable. Chernozhukov et al. (2014b) discuss the differences between the two problems. The main result that is used for the uniform approximation of the process is Dedecker et al. (2014, Theorem 3.2). For

$$\tilde{G}_n(x,v) = n^{-1/2}\sum_{i=1}^{n}\left(\mathbb{I}\{X_i \le x, F_\varepsilon(\varepsilon_i) \le v\} - \prod_{l=1}^{p} x_l v\right),$$

the result yields that there exists a sequence of Brownian bridges $\tilde{B}_n(x,v)$ with

$$\sup_{(x,v)\in[0,1]^{p+1}}|\tilde{G}_n(x,v) - \tilde{B}_n(x,v)| = \mathcal{O}(n^{-1/6}(\log n)^{\kappa+(2(p+1)+4)/3})$$

almost surely for every $\kappa > 0$. This rate carries over to the approximation of the integral representation of $\hat{U}^{(\varepsilon)}_{n,r_n,\omega}(x_0)$. Hence, the lack of a better uniform approximation of an empirical process is the reason for the weaker result. It is not clear, and beyond the scope of our work, whether a better approximation is possible.

It remains to be discussed why the classical proof structure was sufficient in the literature. Johnston (1982) considers univariate regression and assumes that the bandwidth h_n satisfies $h_n = n^{-\delta}$ for $\delta < \frac{1}{3}$. Translated to our case this means $\mathcal{V}_{\cap,k}n^{u/3} \to \infty$ and it would work in the univariate case. In the multivariate case, it would contradict the approximation error assumptions. Claeskens and Van Keilegom (2003) do not need a similar assumption. The main reason for this is that the uniform approximation of the bivariate empirical process by a Gaussian process has an approximation error of rate $n^{-1/2}$ up to logarithms. This is better than Dedecker et al. (2014, Theorem 3.2), but is only applicable to univariate regression problems. The work done by Chao et al. (2017) is in a multivariate regression model. Their bandwidth needs to satisfy

$$n^{-1/6}h^{-p/2-3p/(b_1-2)} = \mathcal{O}(n^{-\nu})$$

for $\nu > 0$ and some constant b_1. The condition implies that $n^{-1/3}h^{-p} = o(1)$. When translated to our method, h^p is replaced by 2^{-k}, so it is necessary that $n^{-1/3}2^k = o(1)$. However, this would contradict the assumptions required for the bias/approximation error. It is noteworthy that a different regression estimator can have a smaller bias and approximation error. For our estimator, the better approximation with the result by Chernozhukov et al. (2014b) is crucial to prove the confidence band results. A larger error would contradict the assumptions for the approximation error and thereby prevent the construction of the confidence bands.

Another drawback of the classical method is that the distribution of the supremum of the Gaussian approximation is harder to characterize. It is possible that the distribution of

$$\sup_{x_0\in[0,1]^p} \left|V_n(x_0)\right| \stackrel{d}{=} \sup_{x_0\in[0,1]^p} \left|\mathcal{V}_{\cap,k}^{-1/2}\sigma \int_{[0,1]^p} \mathbb{P}(x \in A_k(x_0,\omega))dW(x)\right|$$

is related to the distribution of $\mathbf{S}_k$, but the connection is not obvious. In the literature, convergence to some extreme value distribution is usually used, which is possible with a continuous kernel instead of the discontinuous probability. In contrast, Theorem 2.4 provides a direct description of the required distribution.

7.2.1 Proof sketch

In the first step of the proof we show that for $\Gamma_n = [a_n, b_n]$ with sequences $a_n \to -\infty$ and $b_n \to \infty$ such that $\mathbb{E}[\varepsilon 1_{\Gamma_n}(\varepsilon)] = 0$ and

$$U_{n,r_n,\omega}^{(\varepsilon,\Gamma)}(x_0) = n^{-1/2}2^k \int_{[0,1]^p\times\mathbb{R}} 1_{\Gamma_n}(s)s\mathbb{P}(x \in A_k(x_0,\omega))dG_n(x,s).$$

it holds that

$$\mathbb{P}\big(\|\hat{U}_{n,r_n,\omega}^{(\varepsilon)} - U_{n,r_n,\omega}^{(\varepsilon,\Gamma)}\|_\infty > \kappa\big) \leq \frac{2^k}{\kappa}\mathbb{E}[|\varepsilon_1|1_{\Gamma_n^C}(\varepsilon_1)].$$

Using $U_{n,r_n,\omega}^{(\varepsilon,\Gamma)}(x_0)$ we limit the integral with respect to s to a closed interval which is needed to apply integration by parts. To prove the above equation one needs to apply Markov's inequality to

$$\hat{U}_{n,r_n,\omega}^{(\varepsilon)}(x_0) - U_{n,r_n,\omega}^{(\varepsilon,\Gamma)}(x_0) = \frac{2^k}{n}\sum_{j=1}^n \varepsilon_j 1_{\Gamma_n^C}(\varepsilon_j)\mathbb{P}(X_j \in A_k(x_0,\omega) \mid X_j)$$

and use that the probability therein is bounded by one. In the next step we replace the empirical process in the integrator by a different process. We know that $F_\varepsilon(\varepsilon_j) \sim U[0,1]$ and we denote

$$\tilde{G}_n(x,v) := n^{1/2}\left(n^{-1}\sum_{i=1}^n \mathbb{I}\{X_i \leq x, F_\varepsilon(\varepsilon_i) \leq v\} - \prod_{l=1}^p x_l v\right),$$
$$\tilde{G}_{n,1}(x) := \tilde{G}_n(x,1) = n^{1/2}\left(n^{-1}\sum_{i=1}^n \mathbb{I}\{X_i \leq x\} - \prod_{l=1}^p x_l\right),$$

$$\tilde{G}_{n,2}(v) := \tilde{G}_n(1,v) = n^{1/2}\left(n^{-1}\sum_{i=1}^{n}\mathbb{I}\{F_\varepsilon(\varepsilon_i)\leq v\} - v\right).$$

For $U^{(\varepsilon,\Gamma)}_{n,r_n,\omega}(x_0)$ it holds that

$$\begin{aligned}U^{(\varepsilon,\Gamma)}_{n,r_n,\omega}(x_0) &= n^{-1/2}\int_{[0,1]^p\times\mathbb{R}} 1_{\Gamma_n}(s)s2^k\mathbb{P}(x\in A_k(x_0,\omega))dG_n(x,s)\\ &= n^{-1/2}\int_{[0,1]^{p+1}} 1_{\tilde{\Gamma}_n}(v)F_\varepsilon^{-1}(v)2^k\mathbb{P}(x\in A_k(x_0,\omega))d\tilde{G}_n(x,v).\end{aligned}$$

For

$$\tilde{G}^*_n(x,v) := \tilde{G}_n(x,v) - v\tilde{G}_{n,1}(x) - \prod_{l=1}^{p} x_l\tilde{G}_{n,2}(v). \tag{7.9}$$

we denote

$$U^{(\varepsilon,*)}_{n,r_n,\omega}(x_0) := n^{-1/2}2^k\int_{[0,1]^{p+1}} 1_{\tilde{\Gamma}_n}(v)F_\varepsilon^{-1}(v)\mathbb{P}(x\in A_k(x_0,\omega))d\tilde{G}^*_n(x,v). \tag{7.10}$$

One can prove that

$$\begin{aligned}&n^{1/2}\big(U^{(\varepsilon,\Gamma)}_{n,r_n,\omega}(x_0) - U^{(\varepsilon,*)}_{n,r_n,\omega}(x_0)\big)\\ &\quad= \int_{\mathbb{R}} s1_{\Gamma_n}(s)d\tilde{G}_{n,2}(F_\varepsilon(s))\int_{[0,1]^p} 2^k\mathbb{P}(x\in A_k(x_0,\omega))dF_X(x)\\ &\quad= \int_{\mathbb{R}} s1_{\Gamma_n}(s)d\tilde{G}_{n,2}(F_\varepsilon(s))\\ &\quad= \mathcal{O}_{\mathbb{P}}(1).\end{aligned}$$

which implies

$$\|U^{(\varepsilon,\Gamma)}_{n,r_n,\omega} - U^{(\varepsilon,*)}_{n,r_n,\omega}\|_\infty = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$$

because the term does not depend on x_0. In the next step, we further approximate $U^{(\varepsilon,*)}_{n,r_n,\omega}(x_0)$ by "changing" the integrator from $\tilde{G}^*_n$ to $\tilde{B}^*_n$. Let $\tilde{B}_n(x,v)$ be a sequence of Brownian bridges that fulfills $\sup_{(x,v)\in[0,1]^p}|\tilde{G}_n(x,v) - \tilde{B}_n(x,v)| \xrightarrow{a.s.} 0$. The existence of these Brownian bridges is guaranteed by Dedecker et al. (2014, Theorem 3.2). We define

$$\tilde{B}^*_n(x,v) = \tilde{B}_n(x,v) - v\tilde{B}_n(x,1) - \tilde{B}_n(1,v)\prod_{l=1}^{p} x_l. \tag{7.11}$$

Further let W^*_n be a sequence of Brownian sheets with

$$\tilde{B}_n(x,v) = W^*_n(x,v) - W^*_n(1,1)v\prod_{l=1}^{p} x_l.$$

For

$$U^{(\varepsilon,1)}_{n,r_n,\omega}(x_0) := n^{-1/2}2^k\int_{[0,1]^{p+1}} 1_{\tilde{\Gamma}_n}(v)F_\varepsilon^{-1}(v)\mathbb{P}(x\in A_k(x_0,\omega))d\tilde{B}^*_n(x,v) \tag{7.12}$$

and $U^{(\varepsilon,*)}_{n,r_n,\omega}$ from (7.10) it holds that

$$\|U^{(\varepsilon,*)}_{n,r_n,\omega} - U^{(\varepsilon,1)}_{n,r_n,\omega}\|_\infty = o_{\mathbb{P}}\left(n^{-1/2}2^k n^{-r/6}(\log n)^{r(\kappa+(2p+6)/3)}\right). \tag{7.13}$$

This is the main step of the proof where integration by parts is used. To prove this, we use an auxiliary result. For the process $\tilde{G}^*_n(x,v)$ from equation (7.9) and the Gaussian process $\tilde{B}^*_n(x,v)$ from equation (7.11) it holds that

$$\sup_{(x,v)\in[0,1]^{p+1}} \left| \frac{\tilde{G}^*_n(x,v) - \tilde{B}^*_n(x,v)}{[v(1-v)\prod_{l=1}^p x_l]^a} \right| = o_p(n^{-u/6}(\log n)^{r(\kappa+(2(p+1)+4)/3)}) \tag{7.14}$$

for every $\kappa > 0$ and u, a with $u < 1$ and $a/(1-u) < 1/2$. This can be proved using Theorem 3.2 by Dedecker et al. (2014). In order to prove the validity of (7.13), it is necessary to use two different results for multivariate integration by parts. The first result is by Ansari (2024), while the second is by Zaremba (1968). It is noteworthy that these results hold under different assumptions and consequently yield different original formulas. Nevertheless, it can be demonstrated that these formulas can be written in the same form in our situation. The notation required for the subsequent formulas can be found in Section 7.2.2. Theorem 7.4 (Theorem 1.1 by Ansari (2024)) yields

$$n^{1/2}U^{(\varepsilon,*)}_{n,r_n,\omega}(x_0) = \sum_{m=0}^{p+1}(-1)^m \sum_{I\in B_{m,p+1}:p+1\in I} \int_{[0,1]^m} (\tilde{G}^*_n)^I(\xi)\, dg^I(\xi)$$

for $g(x,v) = 1_{\tilde{\Gamma}_n}(v)F_\varepsilon^{-1}(v)2^k\mathbb{P}(x \in A_k(x_0,\omega))$. It can be applied because $\tilde{G}^*_n$ and g are bounded, right and left continuous, respectively, and of bounded variation. Theorem 7.6, which is by Zaremba (1968), holds for $U^{(\varepsilon,1)}_{n,r_n,\omega}(x_0)$ since $\tilde{B}^*_n$ is continuous and yields

$$n^{1/2}U^{(\varepsilon,1)}_{n,r_n,\omega}(x_0) = \sum_{m=0}^{p+1}(-1)^m \sum_{I\in B_{m,p+1}:p+1\in I} \int_{[0,1]^m} (\tilde{B}^*_n)^I(\xi)\, dg^I(\xi).$$

Thus, the difference of both terms is

$$\begin{aligned} &n^{1/2}(U^{(\varepsilon,*)}_{n,r_n,\omega}(x_0) - U^{(\varepsilon,1)}_{n,r_n,\omega}(x_0)) \\ &\quad = \sum_{m=0}^{p+1}(-1)^m \sum_{I\in B_{m,p+1}:p+1\in I} \int_{[0,1]^m} \big((\tilde{G}^*_n)^I(\xi) - (\tilde{B}^*_n)^I(\xi)\big)\, dg^I(\xi), \end{aligned}$$

and we obtain

$$\begin{aligned} &n^{1/2}|U^{(\varepsilon,*)}_{n,r_n,\omega}(x_0) - U^{(\varepsilon,1)}_{n,r_n,\omega}(x_0)| \\ &\quad \le \sum_{m=0}^{p+1} \sum_{I\in B_{m,p+1}:p+1\in I} \int_{[0,1]^m} \big|(\tilde{G}^*_n)^I(\xi) - (\tilde{B}^*_n)^I(\xi)\big|\, |dg^I(\xi)|. \end{aligned}$$

Using equation (7.14) with $\delta_n = n^{-u/6}(\log n)^{u(\kappa+(2(p+1)+4)/3)}$ we get

$$n^{1/2}|U^{(\varepsilon,*)}_{n,r_n,\omega}(x_0) - U^{(\varepsilon,1)}_{n,r_n,\omega}(x_0)|$$

$$\leq o(\delta_n) \sum_{m=0}^{p+1} \sum_{I \in B_{m,p+1}: p+1 \in I} \int_{[0,1]^m} \left[\xi_{p+1}(1-\xi_{p+1}) \prod_{l=1}^{p} \xi_l \right]^a |dg^I(\xi)|.$$

One can prove that all these integrals are $\mathcal{O}(2^k)$ and hence

$$|U^{(\varepsilon,*)}_{n,r_n,\omega}(x_0) - U^{(\varepsilon,1)}_{n,r_n,\omega}(x_0)| = n^{-1/2} 2^k o(n^{-u/6} (\log n)^{u(\kappa+(2(p+1)+4)/3)}).$$

In the next step we define

$$U^{(\varepsilon,2)}_{n,r_n,\omega}(x_0) := n^{-1/2} \int_{[0,1]^p \times \mathbb{R}} 1_{\Gamma_n}(s) s 2^k \mathbb{P}(x \in A_k(x_0,\omega)) dW^*_n(x, F_\varepsilon(s))$$

Similar to one step for the empirical processes it holds that

$$\begin{aligned}
\tilde{B}^*_n(x,v) - W^*_n(x,v) &= (\tilde{B}^*_n(x,v) - \tilde{B}_n(x,v)) - (W^*_n(x,v) - \tilde{B}_n(x,v)) \\
&= -\left(v\tilde{B}_n(x,1) + \tilde{B}_n(1,v) \prod_{l=1}^{p} x_l \right) - \left(v W^*_n(1,1) \prod_{l=1}^{p} x_l \right).
\end{aligned}$$

and thus, for $U^{(\varepsilon,1)}_{n,r_n,\omega}$ from (7.12) we have

$$\begin{aligned}
&\sqrt{n}(U^{(\varepsilon,1)}_{n,r_n,\omega}(x_0) - U^{(\varepsilon,2)}_{n,r_n,\omega}(x_0)) \\
&\quad = -\int_{[0,1]^p} 2^k \mathbb{P}(x \in A_k(x_0,\omega)) dF_X(x) \int_{\Gamma_n} s d\tilde{B}_n(1, F_\varepsilon(s)) \\
&\quad = -\int_{\Gamma_n} s d\tilde{B}_n(1, F_\varepsilon(s)) \\
&\quad = \mathcal{O}_{\mathbb{P}}(1).
\end{aligned}$$

This yields

$$\|U^{(\varepsilon,1)}_{n,r_n,\omega} - U^{(\varepsilon,2)}_{n,r_n,\omega}\|_\infty = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$$

because the term does not depend on x_0. Let W be a Brownian sheet with p arguments. Define

$$U^{(\varepsilon,3)}_{n,r_n,\omega}(x_0) := n^{-1/2} \sigma_n \int_{[0,1]^p} 2^k \mathbb{P}(x \in A_k(x_0,\omega)) dW(x)$$

with $\sigma_n^2 = \mathbb{E}[\varepsilon^2 1_{\Gamma_n}(\varepsilon)]$. Comparing the covariances one can show that $U^{(\varepsilon,3)}_{n,r_n,\omega} \stackrel{d}{=} U^{(\varepsilon,2)}_{n,r_n,\omega}$. We know that $\sigma_n^2 \to \sigma^2$. If one chooses Γ_n large enough the combination of these steps and Slutsky's theorem yield that there exists a random process

$$V_n(x_0) \stackrel{d}{=} \mathcal{V}_{\cap,k}^{-1/2} \sigma \int_{[0,1]^p} \mathbb{P}(x \in A_k(x_0,\omega)) dW(x)$$

with

$$\begin{aligned}
&\sup_{x_0 \in [0,1]^p} \left| \sqrt{\frac{n}{2^{2k} \mathcal{V}_{\cap,k}}} \hat{U}^{(\varepsilon)}_{n,r_n,\omega}(x_0) - V_n(x_0) \right| \\
&\qquad = \mathcal{O}_{\mathbb{P}}\left(2^{-k} \mathcal{V}_{\cap,k}^{-1/2} \right) + o_{\mathbb{P}}\left(\mathcal{V}_{\cap,k}^{-1/2} n^{-u/6} (\log n)^{u(\kappa+(2p+6)/3)} \right)
\end{aligned}$$

for $u < 1$ as above.

7.2.2 Auxiliary - multivariate integration by parts

In the proof of one of the approximation lemmas we need to use multivariate integration by parts for discontinuous functions. A formula for this is given in Theorem 1.1 by Ansari (2024). Before we state the theorem we introduce the necessary notation and definitions. We consider a hyperrectangle $\Xi = \prod_{j=1}^{d} \Xi_j = \prod_{j=1}^{d} [a_j, b_j]$ of dimension d which in our case is given by $\Xi := [0,1]^{p+1}$ with $d = p+1$. Only in this section we will use $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ instead of $x = (x^{(1)}, \ldots, x^{(d)})$ to denote the components of a vector, to keep the notation clear. For a function $f : \Xi \to \mathbb{R}$ and $I = \{i_1, \ldots, i_k\} \subset [d], I \neq \emptyset$ we define the lower I-marginal f_I of a function as

$$f_I(x_{i_1}, \ldots, x_{i_k}) := \lim_{x_j \downarrow a_j, j \notin I} f(x_1, \ldots, x_d).$$

The upper marginal is given by

$$f^I(x_{i_1}, \ldots, x_{i_k}) := \lim_{x_j \uparrow b_j, j \notin I} f(x_1, \ldots, x_d).$$

We further define $f_\emptyset := \lim_{x_j \downarrow a_j} f(x_1, \ldots, x_d)$ and $f^\emptyset := \lim_{x_j \uparrow b_j} f(x_1, \ldots, x_d)$. A function that is continuous on the boundaries of Ξ is called grounded if

$$\lim_{x_j \to a_j} f(x_1, \ldots, x_d) = 0 \qquad \forall j \in [d] \text{ and } x_i \in \Xi_i, i \in [d] \setminus \{j\}.$$

This is equivalent to $f_I(x) = 0$ for all $x \in \Xi_I := \prod_{j \in I} [a_j, b_j]$ and $I \subsetneq [d]$. Further, for a function $f : \Xi \to \mathbb{R}$ that is continuous on the boundaries of Ξ, define its survival function $\overline{f} : \Xi \to \mathbb{R}$ by

$$\overline{f}(x) := \sum_{k=0}^{d} \sum_{I \subseteq \{1,\ldots,d\}, I = \{i_1,\ldots,i_k\}} (-1)^{|I|} f^I(x_{i_1}, \ldots, x_{i_k}), \qquad x = (x_1, \ldots, x_k) \in \Xi.$$

For the rather complex definition of a measure inducing function we refer to Ansari (2024). If a function h is measure inducing, the measure ν_h denotes the signed measure induced by h. Define $\psi_h(g) := \int_\Xi g(x) d\nu_h$ and

$$\pi_g(f) := \sum_{I \subset \{1,\ldots,d\} : I \neq \emptyset} \int_{\Xi_I} f_I(u) d\nu_{g_I}(u) + f_\emptyset g_\emptyset.$$

The following result is Theorem 1.1 by Ansari (2024).

Theorem 7.4 (Integration by parts)**.** *Let g and h be functions that are continuous on the boundaries of Ξ and satisfy an additional technical directional continuity condition. Let g and all its lower marginals be measure inducing. Let h be measure inducing, bounded and grounded. If g and h have no common discontinuities on the same side of each point, then*

$$\int_\Xi g(x) d\nu_h = \sum_{I \subset \{1,\ldots,p\} : I \neq \emptyset} \int_{\Xi_I} \bar{h}_I(u) d\nu_{g_I}(u) + \bar{h}_\emptyset g_\emptyset.$$

The technical conditions, in particular the abstract directional continuity condition and the assumption that the functions are measure inducing, are satisfied in our application. In particular, both functions must be of bounded variation, which is the case. For more details and the concrete directional continuity condition, we refer to Ansari (2024).

We note that the notation$\bar{h}_I$ means the lower marginal of the survival function, i.e. the bar is "applied" before the index. In general this does not commute. We want to express the formula more explicitly. We have

$$\begin{aligned}
\bar{h}_I((\xi_i)_{i\in I}) &= \lim_{\xi_i \downarrow 0, i\notin I} \bar{h}(\xi) \\
&= \lim_{\xi_i \downarrow 0, i\notin I} \sum_{k=0}^{p+1} \sum_{L\subseteq\{1,\dots,p+1\}, L=\{l_1,\dots,l_k\}} (-1)^{|L|} h^L(\xi_{l_1},\dots,\xi_{l_k}) \\
&= \sum_{k=0}^{p+1} \mathbb{I}\{L\subset I\} \sum_{L\subseteq\{1,\dots,p+1\}, L=\{l_1,\dots,l_k\}} (-1)^{|L|} h^L(\xi_{l_1},\dots,\xi_{l_k}) \\
&= \sum_{k=0}^{|I|} \sum_{L\subseteq I, L=\{l_1,\dots,l_k\}} (-1)^{|L|} h^L(\xi_{l_1},\dots,\xi_{l_k}).
\end{aligned}$$

This implies

$$\begin{aligned}
&\int_\Xi g(x) d\nu_h \\
&\quad= \sum_{I\subset\{1,\dots,p\}: I\neq\emptyset} \int_{\Xi_I} \bar{h}_I(u) d\nu_{g_I}(u) + \bar{h}_\emptyset g_\emptyset \\
&\quad= \sum_{I\subset\{1,\dots,p\}: I\neq\emptyset} \int_{\Xi_I} \sum_{k=0}^{|I|} \sum_{L\subseteq I, L=\{l_1,\dots,l_k\}} (-1)^{|L|} h^L(u_L) d\nu_{g_I}(u) + \bar{h}_\emptyset g_\emptyset \\
&\quad= \sum_{I\subset\{1,\dots,p\}: I\neq\emptyset} \sum_{k=0}^{|I|} \sum_{L\subseteq I, L=\{l_1,\dots,l_k\}} (-1)^{|L|} \int_{\Xi_I} h^L(u_L) d\nu_{g_I}(u) + \bar{h}_\emptyset g_\emptyset \\
&\quad= \sum_{I\subset\{1,\dots,p\}: I\neq\emptyset} \sum_{L\subseteq I} (-1)^{|L|} \int_{\Xi_I} h^L(u_L) d\nu_{g_I}(u) + \bar{h}_\emptyset g_\emptyset.
\end{aligned}$$

Remark 7.5. For $p = 1$ we note that

$$\begin{aligned}
\bar{h}_{\{1\}}(\xi) = \bar{h}(\xi) &= \sum_{k=0}^{1} \sum_{I\subseteq\{1\}, I=\{i_1\}} (-1)^{|I|} h^I(\xi) \\
&= h^\emptyset(\xi) - h^{\{1\}}(\xi) \\
&= h(b) - h(\xi),
\end{aligned}$$

which implies $\bar{h}_\emptyset = h(b) - h(a)$. Thus, the formula simplifies to

$$\int_a^b g(x) d\nu_h = \int_a^b \bar{h}_{\{1\}}(u) d\nu_{g_{\{1\}}}(u) + \bar{h}_\emptyset g_\emptyset.$$

$$
\begin{aligned}
&= \int_a^b (h(b) - h(\xi))dg(\xi) + (h(b) - h(a))g(a) \\
&= -\int_a^b h(\xi)dg(\xi) + h(b)(g(b) - g(a)) + (h(b) - h(a))g(a) \\
&= -\int_a^b h(\xi)dg(\xi) + h(b)g(b) - h(a)g(a),
\end{aligned}
$$

which one would expect. Zaremba (1968, Proposition 2) provides another formula that holds under different assumptions. For any function $\phi : [0,1]^d \to \mathbb{R}$ denote

$$\Delta^j \varphi(x) := \varphi(x_1, \dots x_{j-1}, 1, x_{j+1}, \dots, x_d) - \varphi(x_1, \dots x_{j-1}, 0, x_{j+1}, \dots, x_d).$$

The application of multiple Δ^j commutes and for $J \subset [d]$ we use Δ^J to denote the difference operator in all components $j \in J$.

Theorem 7.6. *If over $[0,1]^d$, one of the functions $g(x)$ and $h(x)$ is of bounded variation in the sense of Hardy and Krause, and the other is continuous, then the Riemann-Stieltjes integral*

$$\int_{[0,1)^d} g(x)dh(x)$$

exists. Let $B_{m,d}$ denote the set of subsets of $\{1, \dots, d\}$ with size m. The integral satisfies

$$\int_{[0,1)^d} g(x)dh(x) = \sum_{m=0}^{d} (-1)^m \sum_{I \in B_{m,d}} \Delta^{[d]\setminus I} \int_{[0,1)^m} h(x)d_I g(x),$$

where d_I denotes integration with respect to the entries of x with index $i \in I$.

7.3 Outlook

In this section, we discuss possible future research and further extensions of our work. These topics will be discussed at a superficial level, as any of them could lead to an extensive discussion. Before beginning this discussion, a brief summary of the results of our work is provided. In Chapter 4 we proved a central limit theorem for CPRFs, which allows the construction of confidence intervals. It is noteworthy that the dependence of the asymptotic variance and the confidence interval radius on the density of X and the error variance σ^2 is known, which to the best of our knowledge is new in the literature. The main contribution of our work is Theorem 5.1, which provides uniform confidence bands. While the uniform confidence bands are a novel contribution to the field of random forests, extending these results to more sophisticated random forest models would be a significant advancement. Alternative approaches include extending the result to a more general regression model, or enhancing the result by incorporating additional terms from the Hoeffding-decomposition.

The proof of Theorem 5.1 relies on the Hájek projection being the dominant term in the Hoeffding-decomposition. The inclusion of higher order terms could have two main beneficial effects. First, the assumptions on the tuning parameters may be weaker, and

second, the finite sample distribution may be better approximated. The quantiles of $\mathbf{S}_k$, which determine the radius of the confidence band, do not belong to an asymptotic distribution but depend on k. Higher order terms could lead to a different distribution whose quantiles might work better in practice. The main difficulty of this extension is that one would have to apply results for U-processes that are similar to those for empirical processes. To the best of our knowledge, there are no such results in the literature.

Generalization of the regression model

A way to extend the results is to relax the assumptions in the regression model. For each part of the regression model, there are ways to relax the assumptions. Starting with the error distribution, the assumption that a finite number of moments exists is already fairly mild. Still, it might be possible to reduce the number of existing moments. Another extension would be to allow heteroscedastic errors. This would change the function class $\mathcal{F}_k$ and the empirical process in our proofs. Including the assumption that the variance is bounded, the result by Chernozhukov et al. (2014b) might still be applicable in a similar way. This would affect the radius of the confidence bands in a heterogeneous way. Furthermore, one would need a consistent estimator of the variance function over the entire feature space.

Another potential generalization is to consider different classes of eligible regression functions, for instance with discontinuities. The piecewise constant structure of the random forest estimator should, in principle, allow it to handle such discontinuities. Under appropriate assumptions on the number, placement, and size of the jumps, it may be possible to obtain results similar to those for continuous functions.

The assumptions on the covariates are already quite general. A density bounded from above and below is usually required. We can consider a higher dimension of these random variables. In a high-dimensional regression model, we would need assumptions on the active dimension or the importance of the different features. If the estimator is also modified and data dependent in some way, it may be able to handle the high dimension by mostly using the important features for the partitions. One can use similar proof techniques from the low-dimensional case on the event that the inactive/unimportant features are seldom split. The honest random forest from Section 7.1 might be able to work in a high-dimensional model.

Extension of the random forest algorithm

There are several ways to modify and relax the assumptions on the estimator. One extension that does not fundamentally change the estimator is to consider a more flexible number of splits. If we are still considering a CPRF, there is no reason to prefer any feature in general when choosing the splits. However, after doing k splits per cell, we can allow a random subset of the cells to be split an additional time. We represent this by choosing $k \in \mathbb{Q}$ with $(k - \lfloor k \rfloor)2^k \in \mathbb{N}$. The latter is the number of cells that will be split an additional time. In theory we consider the asymptotic behavior of 2^k and n, but in practice the discrete assumption on k may be limiting. For a fixed finite sample size n, the extension allows a much more flexible choice of $2^k/n$, which would also be the case in various nonparametric regression estimators that use a bandwidth. Extending the results

to this algorithm should be possible since one can use the bounds $\lfloor k \rfloor \leq k \leq \lceil k \rceil$ without changing the asymptotic behavior.

Staying with purely random forests, one could investigate different splitting rules for centered purely random forests. The choice of the tuning parameters in the Ehrenfest forest already leaves room for further research. Alternatively one could consider algorithms that favor splitting cells with larger volumes relative to the distribution of X, if this distribution is known. This approach might lead to an improved estimation in regions with more observations on average.

Furthermore, the assumption of centered splits can be dropped. The centeredness can hurt the ability of the estimator to approximate a function at the fixed cut points. It also limits the diversity of the trees because the partitions are limited to these fixed cut points. Since random forests benefit from this diversity, this could also limit the performance. By omitting the centeredness of the splits, the constant cell size and the structure of the intersection of two cells are lost. It is not clear whether a similar lower bound holds for $\mathcal{V}_{\cap,k}$ holds in general. If it is smaller, then the variance of the leading term is smaller and thus the remaining terms may no longer be negligible. Still, this would be a desirable modification of the algorithm to be closer to the standard random forest algorithm.

Finally, one would like to extend the results to random forests with data dependent partitions. In the most general case, data-dependence is not achieved by honesty and sample splitting. If the partition is data dependent, the splits should not be centered and the cell sizes should be variable. Otherwise, the ability of the algorithm to construct a partition that is useful for the data is limited. Variable partitions allow the use of small cells when it is useful for approximating the regression function, and large cells for better estimation when the regression function is nearly constant. Examples of technical difficulties in this case are the lower bound on $\mathcal{V}_{\cap,k}$ and the dual use of observations for partition construction and estimation. In particular, the stochastic error would be harder to analyze because the errors and the cells lose their independence.

Confidence bands for such a general type of random forest as the one by Breiman (2001) are very desirable because they are frequently used in applications. A smaller step in the data dependent direction would be to allow the partitions to depend on the independent variables. This can be used to avoid empty cells in the partitions, which are possible for purely random forests and occur more often with finer partitions. We have already made a first step towards data dependent forests in Section 7.1.

Bibliography

Ansari, J. (2024). On a version of a multivariate integration by parts formula for lebesgue integrals. *arXiv preprint: 2203.06772*.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Bhattacharya, R. N. and Waymire, E. C. (2022). *Stationary processes and discrete parameter Markov processes*. Graduate Texts in Mathematics. Springer, Cham, 1. edition.

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

Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research (JMLR)*, 9:2015–2033.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *TEST. An Official Journal of the Spanish Society of Statistics and Operations Research*, 25(2):197–227.

Biau, G., Scornet, E., and Welbl, J. (2019). Neural random forests. *Sankhya A. The Indian Journal of Statistics*, 81(2):347–386.

Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics*, 1:1071–1095.

Birke, M., Bissantz, N., and Holzmann, H. (2010). Confidence bands for inverse regression models. *Inverse Problems. An International Journal on the Theory and Practice of Inverse Problems, Inverse Methods and Computerized Inversion of Data*, 26(11):115020, 18.

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

Breiman, L. (2004). Consistency for a simple model of random forests. Technical report, Statistics department University of California at Berkeley.

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.

Chao, S.-K., Proksch, K., Dette, H., and Härdle, W. K. (2017). Confidence corridors for multivariate generalized quantile regression. *Journal of Business & Economic Statistics*, 35(1):70–85.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014a). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5):1787–1818.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014b). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564–1597.

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

Chow, Y. S. and Teicher, H. (1997). *Probability theory - Independence, interchangeability, martingales*. Springer Texts in Statistics. Springer, New York, third edition.

Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *The Annals of Statistics*, 31(6):1852–1884.

Dedecker, J., Merlevède, F., and Rio, E. (2014). Strong approximation of the empirical distribution function for absolutely regular sequences in $\mathbb{R}^d$. *Electronic Journal of Probability*, 19:no. 9, 56.

Eubank, R. L. and Speckman, P. L. (1993). Confidence bands in nonparametric regression. *Journal of the American Statistical Association*, 88(424):1287–1301.

Frees, E. W. (1989). Infinite order U-statistics. *Scandinavian Journal of Statistics. Theory and Applications*, 16(1):29–45.

Giessing, A. (2023). Gaussian and bootstrap approximations for suprema of empirical processes. *arXiv preprint: 2309.01307*.

Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *The Annals of Statistics*, 38(2):1122–1170.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer, New York.

Hall, P. (1993). On edgeworth expansion and bootstrap confidence bands in nonparametric curve estimation. *Journal of the royal statistical society series b-methodological*, 55:291–304.

Hall, P. and Horowitz, J. (2013). A simple bootstrap method for constructing nonparametric confidence bands for functions. *The Annals of Statistics*, 41(4):1892–1921.

Halmos, P. R. (1946). The theory of unbiased estimation. *Annals of Mathematical Statistics*, 17:34–43.

Härdle, W. (1989). Asymptotic maximal deviation of m-smoothers. *Journal of Multivariate Analysis*, 29(2):163–179.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19:293–325.

Hoeffding, W. (1961). The strong law of large numbers for u-statistics. Technical report, North Carolina State University. Dept. of Statistics.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860.

Johnston, G. J. (1982). Probabilities of maximal deviations for nonparametric regression function estimates. *Journal of Multivariate Analysis*, 12(3):402–414.

Kallenberg, O. (2021). *Foundations of modern probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer, Cham, third edition.

Klusowski, J. (2021). Sharp analysis of a simple model for random forests. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 757–765. PMLR.

Konakov, V. D. and Piterbarg, V. I. (1984). On the convergence rate of maximal deviation distribution for kernel regression estimates. *Journal of Multivariate Analysis*, 15(3):279–294.

Lee, A. J. (1990). *U-statistics - Theory and practice*, volume 110 of *Statistics: Textbooks and Monographs*. Dekker, New York.

Li, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.*, 17(3):1001–1008.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research (JMLR)*, 7:983–999.

Mentch, L. and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research (JMLR)*, 17:Paper No. 26, 41.

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

Müller, U. U., Schick, A., and Wefelmeyer, W. (2003). Estimating the error variance in nonparametric regression by a covariate-matched U-statistic. *Statistics. A Journal of Theoretical and Applied Statistics*, 37(3):179–188.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Neumann, M. H. and Polzehl, J. (1998). Simultaneous bootstrap confidence bands in nonparametric regression. *Journal of Nonparametric Statistics*, 9(4):307–333.

Neumeyer, N., Rabe, J., and Trabs, M. (2025). Asymptotic confidence bands for centered purely random forests. *arXiv preprint: 2511.13199*.

Neumeyer, N., Rabe, J., and Trabs, M. (2026). Asymptotic confidence bands for the histogram regression estimator. *arXiv preprint: 2508.12391*.

Patschkowski, T. and Rohde, A. (2019). Locally adaptive confidence bands. *Ann. Statist.*, 47(1):349–381.

Peng, W., Coleman, T., and Mentch, L. (2022). Rates of convergence for random forests via generalized U-statistics. *Electronic Journal of Statistics*, 16(1):232–292.

Proksch, K. (2016). On confidence bands for multivariate nonparametric regression. *Annals of the Institute of Statistical Mathematics*, 68(1):209–236.

Rabe, J. (2025). Asymptotic-cbs-for-cprf. GitHub repositary: available at https://github.com/Jan-Rabe/Asymptotic-CBs-for-CPRF.

Sabbah, C. (2014). Uniform confidence bands for local polynomial quantile estimators. *ESAIM. Probability and Statistics*, 18:265–276.

Scornet, E. (2016a). On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146:72–83.

Scornet, E. (2016b). Random forests and kernel methods. *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, 62(3):1485–1500.

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

Shen, Y., Gao, C., Witten, D., and Han, F. (2020). Optimal estimation of variance in nonparametric regression with random design. *The Annals of Statistics*, 48(6):3589–3618.

Skorski, M. (2025). Handy formulas for binomial moments. *Modern Stochastics: Theory and Applications*, 12(1):27–41.

Smirnov, N. V. (1950). On the construction of confidence regions for the density of distribution of random variables. *Doklady Akad. Nauk SSSR (N.S.)*, 74:189–191.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

Tusnády, G. (1977). A remark on the approximation of the sample DF in the multidimensional case. *Periodica Mathematica Hungarica. Journal of the János Bolyai Mathematical Society*, 8(1):53–55.

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer, New York. With applications to statistics.

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.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A*, 26:359–372.

Xia, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 60(4):797–811.

Xu, T., Zhu, R., and Shao, X. (2024). On variance estimation of random forests with infinite-order U-statistics. *Electronic Journal of Statistics*, 18(1):2135–2207.

Zaremba, S. C. (1968). Some applications of multidimensional integration by parts. *Polska Akademia Nauk. Annales Polonici Mathematici*, 21:85–96.

Zhang, Y., Ji, W., and Bradic, J. (2024). Adaptive split balancing for optimal random forest. *arXiv preprint: 2402.11228*.

Appendix A

Formalities

A.1 Abstract

One of the most prominent estimation problems in mathematical statistics, essential for a wide range of applications, is described by regression models. A popular class of estimators in regression models are random forests, which are often referred to as machine learning methods. Random forests are nonparametric estimators that can be applied to regression and classification problems. In our work, we focus on their application in a nonparametric multivariate regression model

$$Y = m(X) + \varepsilon,$$

with regression function m, response $Y \in \mathbb{R}$, covariates $X \in [0,1]^p$ and errors $\varepsilon \in \mathbb{R}$ that are independent of X. A random forest averages the estimates of several randomized regression trees that are based on different subsamples of the training sample. Regression trees construct a hierarchical partition of the feature space by performing iterative, axis-aligned splits and use a piecewise constant estimation of the regression function on the cells in this partition. In practice, random forests appear to be successful, even when applied to high-dimensional data. The statistical analysis of random forests is challenging, especially due to the dependence of the tree partitions on the data. Even for the simpler purely random forests, which omit this dependence, the majority of the literature focuses on consistency results.

In this thesis, we analyze the asymptotic behavior of centered purely random forests with a focus on the construction of confidence intervals and bands, which provide quantitative information about the quality of an estimator and are thus a desirable tool in applications. In centered purely random forests, the regression trees perform the splitting during the partition construction in the center of each cell. We consider two specific types of centered purely random forests. One of them is the classical type, which selects the splitting direction uniformly distributed on all p axes, and the other variant, which is the Ehrenfest centered purely random forest, is newly introduced and its partitioning algorithm is influenced by the Ehrenfest model for diffusion. We prove that this new variant achieves the minimax optimal convergence rate for Hölder continuous regression functions m, which is not the case for the classical centered purely random forest variant.

Similar to several articles in the literature, we exploit an interpretation of random forests as generalized incomplete U-statistics to analyze their asymptotic distribution.

Important concepts for U-statistics employed in the asymptotic analysis are the Hoeffding-decomposition and the Hájek projection. First, we prove pointwise asymptotic normality of centered purely random forests based on results for generalized U-statistics by Peng et al. (2022), which allows for the construction of asymptotic confidence intervals. These results are explicit in the sense, that the only possibly unknown objects affecting the asymptotic variance and the radius of the confidence intervals are the variance of the errors and the density of the covariates X. To the best of our knowledge, the explicit form of our results is new among the asymptotic results for random forests in the literature.

The main theorem we prove provides uniform asymptotic confidence bands based on centered purely random forests. As far as we are aware, this is the first confidence band result for any random forest type. Since confidence bands are a uniform generalization of confidence intervals, the proof of this result is substantially different from the pointwise case and requires more sophisticated techniques for the handling both the dominating and the remainder terms from the Hoeffding-decomposition. An important component of the proof is the application of a result by Chernozhukov et al. (2014b), which approximates the supremum of an empirical process by the supremum of a Gaussian process. This allows us to approximate and thereby characterize the distribution of the supremum of the Hájek projection, which is the dominating term from the Hoeffding-decomposition and also an empirical process. The same proof technique allows us to construct asymptotic confidence bands for the histogram regression estimator. Further, we obtain a uniform consistency result for centered purely random forests. A possible direction of future research is to generalize the results to more sophisticated and data dependent random forests. We illustrate the theoretical results with a simulation study.

A.2 Zusammenfassung

Eines der bekanntesten Schätzprobleme in der mathematischen Statistik, welches für eine Vielzahl von Anwendungen essenziell ist, wird durch Regressionsmodelle beschrieben. Eine verbreitete Klasse von Schätzern in Regressionsmodellen sind Random Forests, die häufig als Methoden des maschinellen Lernens angesehen werden. Random Forests sind nichtparametrische Schätzer, die in Regressions- und Klassifikationsproblemen angewendet werden können. In dieser Arbeit liegt der Fokus auf ihrer Anwendung in einem nichtparametrischen multivariaten Regressionsmodell

$$Y = m(X) + \varepsilon,$$

mit Regressionsfunktion m, abhängiger Variable $Y \in \mathbb{R}$, unabhängiger Variable $X \in [0,1]^p$ und Fehlern $\varepsilon \in \mathbb{R}$, die stochastisch unabhängig von X sind. Ein Random Forest mittelt die Schätzungen von vielen randomisierten Regressionsbäumen, die auf verschiedenen Teilstichproben der Trainingsdaten beruhen. Regressionsbäume konstruieren eine hierarchische Partition von $[0,1]^p$, indem sie iterativ Teilungen (Splits) parallel zu den Achsen durchführen und eine stückweise konstante Schätzung der Regressionsfunktion auf den Zellen in dieser Partition verwenden. In der Praxis liefert die Anwendung von Random Forests oft gute Ergebnisse, selbst bei der Anwendung auf hochdimensionale Daten. Die statistische Analyse von Random Forests ist eine Herausforderung, insbesondere wenn

die Partitionen der Regressionsbäume von den Daten abhängen. Selbst für die weniger komplexen Purely Random Forests, bei denen diese Abhängigkeit wegfällt, konzentriert sich der Großteil der Literatur auf Konsistenzresultate.

In dieser Arbeit wird das asymptotische Verhalten von Centered Purely Random Forests analysiert, wobei der Schwerpunkt auf der Konstruktion von Konfidenzintervallen und -bändern liegt, die quantitative Informationen über die Qualität eines Schätzers liefern und somit ein nützliches Werkzeug in Anwendungen sind. In Centered Purely Random Forests teilen die Regressionsbäume die Zellen während der Partitionskonstruktion immer zentriert. Wir betrachten zwei bestimmte Typen von Centered Purely Random Forests. Einer davon ist der klassische Typ, der die Richtung der Splits uniform verteilt auf allen p Achsen wählt, und die andere Variante, der sogenannte Ehrenfest Centered Purely Random Forest, wird neu eingeführt und sein Partitionierungsalgorithmus ist angelehnt an das Ehrenfest Modell für Diffusion. Wir beweisen, dass diese neue Variante die minimax optimale Konvergenzrate für hölderstetige Regressionsfunktionen m erreicht, was für den klassischen Centered Purely Random Forest nicht der Fall ist.

Angelehnt an mehrere Artikel aus der Literatur nutzen wir eine Interpretation von Random Forests als verallgemeinerte unvollständige U-Statistiken, um ihre asymptotische Verteilung zu analysieren. Wichtige Konzepte für U-Statistiken, die in der asymptotischen Analyse verwendet werden, sind die Hoeffding-Zerlegung und die Hájek Projektion. Zunächst wird die punktweise asymptotische Normalverteilung von Centered Purely Random Forests auf der Grundlage der Resultate für verallgemeinerte U-Statistiken von Peng et al. (2022) bewiesen, was die Konstruktion von asymptotischen Konfidenzintervallen ermöglicht. Diese Resultate sind explizit in dem Sinne, dass die einzigen möglicherweise unbekannten Objekte, die einen Einfluss auf die asymptotische Varianz und den Radius der Konfidenzintervalle haben, die Varianz der Fehler und die Dichte der Zufallsvariablen X sind. Unseres Wissens nach ist die explizite Form unserer Ergebnisse eine Neuheit in der Literatur zum asymptotischen Verhalten von Random Forests.

Das zentrale Theorem, das wir beweisen, liefert gleichmäßige asymptotische Konfidenzbänder basierend auf Centered Purely Random Forests. Soweit es uns bekannt ist, ist dies das erste Konfidenzbandresultat überhaupt für eine Random Forest Variante. Da Konfidenzbänder eine gleichmäßige Verallgemeinerung von Konfidenzintervallen sind, unterscheidet sich der Beweis dieses Theorems wesentlich vom punktweisen Fall und erfordert komplexere Beweistechniken sowohl für die dominierenden als auch für die vernachlässigbaren Terme aus der Hoeffding-Zerlegung. Ein wichtiger Bestandteil des Beweises ist die Anwendung eines Resultats von Chernozhukov et al. (2014b), das erlaubt, das Supremum eines empirischen Prozesses durch das Supremum eines Gauß-Prozesses zu approximieren. Auf diese Weise können wir die Verteilung des Supremums der Hájek Projektion, die der dominierende Teil der Hoeffding-Zerlegung und zugleich ein empirischer Prozesses ist, approximieren und damit charakterisieren. Dieselbe Beweistechnik ermöglicht es uns, asymptotische Konfidenzbänder für den Histogramm Regressionsschätzer zu konstruieren. Außerdem erhalten wir ein gleichmäßiges Konsistenz Resultat für Centered Purely Random Forests. Eine mögliche Richtung zukünftiger Forschung ist die Verallgemeinerung der Ergebnisse auf komplexere und datenabhängige Random Forests. Wir veranschaulichen die theoretischen Ergebnisse mit einer Simulationsstudie.

A.3 Publications related to this dissertation

Extracts of the results of this dissertation have been published in preprints on arXiv, in collaboration with my supervisors Natalie Neumeyer and Mathias Trabs.

- Neumeyer et al. (2025) covers the confidence band results for centered purely random forests.
- Neumeyer et al. (2026) covers the confidence band results for the histogram regression estimator.

Declaration of personal contribution

Chapters 1 to 3 of this dissertation are introductory material and contain few original results. The novel results in Chapters 3 to 5 are are jointly due to Natalie Neumeyer, Mathias Trabs and the candidate, where all detailed proofs have been carried out by the candidate. The simulation study in Chapter 6 has been carried out by the candidate. Chapter 7, which contains the discussion of extensions and a comparison with a classical proof method, is due to the candidate. It is planned that the results of the dissertation, especially those in Chapter 4 and Chapter 5, will lead to a joint publication, co-authored by Natalie Neumeyer, Mathias Trabs and the candidate.

A.4 Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Jan Rabe
Hamburg, den 21.03.2025